**Visualizing the Protein Sequence Universe**

Larissa Stanberry 1-3, Roger Higdon 1-4, Winston Haynes 1-3, Natali Kolker 2-4, William Broomall 3-4, and Eugene Kolker1-5

**Geoffrey Fox, Judy Qiu, Saliya Ekanayake, Yang Ruan 6**

1 Bioinformatics & High-Throughput Analysis Laboratory, Seattle Children’s Research Institute, Seattle, Washington.

2 Predictive Analytics, Seattle Children’s Hospital, Seattle, Washington.

3 Data-Enabled Life Sciences International (DELSA), Seattle, Washington.

4 High-Throughput Analysis Core, Seattle Children’s Research Institute, Seattle, Washington.

5 Departments of Medical Education & Biomedical Informatics and Pediatrics, School of Medicine, University of Washington, Seattle, Washington.

6 School of Informatics and Computing and Pervasive Technology Institute, Indiana University Bloomington. {gcf, xqiu, sekanaya, yangruan}@indiana.edu

**Abstract**

Modern biology has experienced a rapid growth of data that challenges our analytical skills and existing cyberinfrastructure. Functional annotation of newly sequenced genomes requires a fusion of standard and novel analytical approaches and computational means. Visualization of the protein sequence universe can be instrumental in meeting the need for annotation. Current existing resources lack scalable visualization tools to study the structure of the protein sequence universe. Here, we discuss a visualization approach based on multi-dimensional scaling that allows creation of a 3D embedding of the protein sequence space. We demonstrate the method by creating a 3D representation of prokaryotic clusters of orthologous proteins. The merits of the method are examined by comparing the resultant embedding with functional grouping of the original data.

**Introduction**

Functional annotation of newly sequenced genomes and meta-genomes is one of the principal challenges of modern biology. Rapidly advancing sequencing technologies generate peta- and even exa-scale data, exponentially expanding the protein sequence universe (REF: Microbiome Project, i5K Insect, Arthropod Genome Sequencing Initiative). Assigning functions to this glut of newly sequenced proteins is an immense computational challenge that requires advanced analytical tools and scaling capabilities [Louie2009, Louie2010, Kolker2011, Kolker2003, Ragh..2004, Kolker2004, Higdon2009].

Surmounting the annotation challenge requires the ability to coherently display a vast amount of protein information. Currently, none of the existing resources provide interactive tools to visualize and analyze data across large sets of proteins. The analysis is typically done on the experiment level and in the context of known relationships, e.g. pathways, complexes. Tools for pathway and network visualization (e.g. Ingenuity or Biobase) do not relate to sequence similarity or extend to the entire protein sequence universe. To map the proteins beyond their network or pathway, to estimate the mutual proximity of proteins identified in the experiment (with respect to sequence similarity, expression levels, structure, etc.) or to project the identified proteins into the subspace of known proteins requires ample computational effort. Clearly, such a resource-intensive approach is a major undertaking for individual laboratories.

In order to infer protein function, sequence data analysis relies on sophisticated statistical and machine-learning methods including pairwise and multiple sequence alignment algorithms [SC3-6], structure prediction models [SC7-8], motif and domain [SC9-12] finding algorithms and clustering methods [SC1-2, SC13-16]. Numerous databases provide information about functions of proteins, protein domains, and protein families including general resources (UniProt, GenBank), pathways (KEGG, PANTHER, MetaCyc, Reactome), protein structure (PDB), protein domains (Pfam, CDD), protein families (ProtClustDB, COG, eggNog, CluSTr, Protomap, SYSTERS).

Currently, functional annotation and analysis is done on a protein-by-protein basis which is tedious, time-consuming, and relies on a multitude of resources. While the 'manual' approach is feasible for a small group of proteins, it quickly becomes unsustainable for a large number of proteins [SC31-32]. Given the scale of modern research studies, the inability to quickly and efficiently analyze protein sequence data creates an ever expanding backlog of un-annotated proteins [Bork2000, Kolker2003, Kolker2004].

A viable computational approach to functional annotation uses clustering to identify functionally similar groups of proteins. The cluster annotation can then be propagated to newly assigned uncharacterized proteins. In view of the exponential growth of data, this approach is computationally advantageous as it facilitates the annotation of large numbers of proteins.

To demonstrate the complexity of the protein functional annotation task, we completed a first of a kind all-versus-all sequence alignments for 9.9 million proteins in the UniRef100 database. The alignment was done on the Microsoft Windows Azure cloud system with 475 eight-core virtual machines that produced over 3 billion filtered records in six days. Protein classification into functional groups was then performed using an innovative implementation of a single-linkage algorithm on a Hadoop compute cluster using Hive and the MapReduce paradigm (Kolker2011). Using the normalized alignment score, we have assigned 68% of 5.1 million bacterial proteins into clusters in the Clusters of Orthologous Genes (COG) database.

The remaining proteins were classified into functional groups using Hive and custom jars implemented on top of Apache Hadoop utilizing the MapReduce paradigm. This implementation significantly reduced the run time for non-indexed queries and optimized clustering performance. Consequently, nearly 2 million proteins were agglomerated into half a million functional groups. A similar approach was applied to 2.8 million eukaryotic sequences from UniRef100, thereby expanding the eukaryotic database KOG by over 1 million proteins and producing 100,000 new functional groups.

The UniRef100 clustering project showcased both the promise and the challenges of the protein annotation. It took the considerable efforts of a diverse group of researchers along with multiple cloud systems to successfully complete the task. Established open-source resources are struggling to cope with the influx of data as well and are either no longer supported (SYSTERS, CluStr, COG) or have limited capabilities (ProtClustDB, eggNog). All this highlights a pressing need of the biological community for a scalable and efficient computational approach to visualize, explore and assign functional annotations to new proteins.

Protein sequence annotation is an example of one of the grand challenges of modern biology that requires a focused, concentrated effort of experts from multiple scientific fields. In addition, the current funding climate severely limits the capabilities of any single research laboratory [Hather2010] thus compelling scientists to forge alliances and leverage skills across different discipline [Ozdemir2011]. The drive for collective innovation in data-enabled sciences translates into community efforts such as DELSA, the Data-enabled Life Sciences Alliance, whose goal is to create a synergy between the computer science and life sciences to tackle modern biological challenges through best computational practices and most advanced cyberinfrastructure [Ozdemir2011, Kolker2011].

In this paper, we propose a novel visualization tool to explore and visualize the protein sequence universe. The proposed method uses a multidimensional scaling (MDS) to create a 3D embedding of the protein sequence space. The parallel implementation on a multiple grid platform utilizes Iterative MapReduce, the usual Message Passing Interface MPI and threading. The proposed method provides an advanced visualization tool that allows thorough exploration of the protein sequence data. Furthermore, the approach could be extended to allow mapping and annotation of newly sequenced proteins.

In what follows, we describe the methods to create a 3D rendering of 100,000 prokaryotic protein sequences from the COG database. We briefly describe the data and outline the implementation of multi-dimensional scaling approach. The 3D representation of the protein subspace is given in the Results section. The application and relevance of the proposed approach to the functional annotation task adjourns the paper.

**Materials and Methods**

**COG Database**

A major principle of molecular evolution is that functionally important proteins tend to be conserved across species. Clusters of Orthologous Groups of proteins (COGs) is a project by the National Center for Biotechnology Information (NCBI). The project constructed clusters of proteins from 66 prokaryotic and seven eukaryotic genomes. For each protein, the best aligned protein in every other genome was determined using a sequence similarity search. If three proteins from three organisms were mutual best hits, they created a triple. COGs are the result of exhaustive, successive merging of triples with two common members. Manual curation of the clusters was done by experts to ensure correct grouping and functional annotations The COG database is separated into COGs for prokaryotic genomes and KOGs for eukaryotic genomes (Tatusov et al., 1997, Tatusov et al., 2003). In this paper, we are using the COG database of prokaryotic genomes that we will refer to as COGs.

**UniRef**

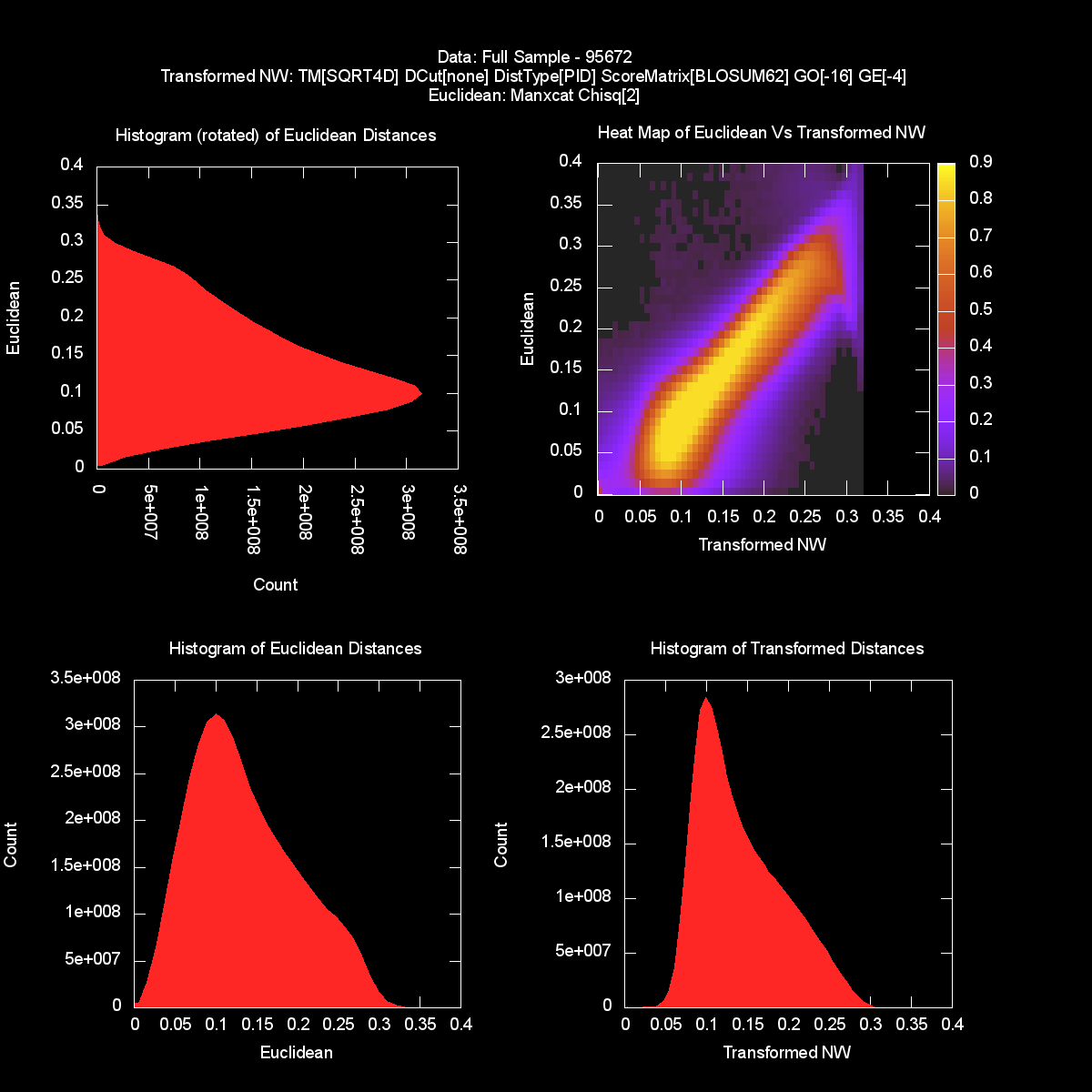
UniRef is composed of the distinct databases UniRef100, UniRef90, and UniRef50, which have 100%, 90%, and 50% sequence similarity, respectively, within protein clusters and reduce the UniProt database size by approximately 10%, 40%, and 70%, respectively. Each cluster contains one reference sequence and all proteins within the similarity threshold to the reference. UniRef retains annotation from all members of the protein cluster to prevent information loss [Suzek et al., 2007].

**Multi-Dimensional Scaling**

Multi-Dimensional Scaling (MDS) algorithm was used to project the protein sequence similarity data into a low-dimensional space (REF). The method has an *O(n2)* computational complexity to find *n* unknown 3D vectors X(*i*) from *n* sequences *i*. It can be heuristically solved in several ways including Expectation Maximization (EM) and *χ2* minimization [GCF REFS]. In this paper, we used the Sammon *χ2* optimization approach where terms of the cost function are inversely weighted by the distance (REF). Experience has found that the increased importance of smaller dissimilarities in Sammon helps studies of clustered sequences. Unfortunately the elegant EM SMACOF methods for MDS and our deterministic annealing improvements [GCF REFS] are not directly applicable to Sammon’s objective function given in (1) below. We used a highly robust implementation of nonlinear *χ2* minimization using the well-known Marquardt Levenberg method to regularize Newton’s equations.

HSammon = Σ*i< j=*1*n* (f(δ(*i*, *j*)) - d(X(*i*) *,* X(*j*)))2 / f(δ(*i*, *j*) ) (1)

In equation (1), d is the three dimensional Euclidean metric and δ(*i*, *j*) any measure of the pairwise dissimilarity between sequences *i* and *j*. In MDS one does not need the formal properties of Euclidean distance measures and one can replace the initial measure of distance by any monotonic transformation of distances [Ref] which is represented by the function f in equation (1). We chose heuristically f to reduce the formal dimension of distance data (in this case, from 244 with original δ(*i*, *j*) to 14 for f(δ(*i*, *j*) ) after mapping), which allows for a more uniform coverage of the target Euclidean space by the MDS projections. In future papers, we will explore different choices for f; for example if f is taken as identity, the high initial dimension implies that essentially all sequences are projected into surface of 3D structure, which lowers the utility of the mapping. One choses f to increase the ratio of standard deviation to mean for f(δ(*i*, *j*)) and spread out range of pairwise distance values.



*Figure 2: Projected Histograms and Density Map comparing the n2 f(d(i,j)) (abscissa) with mapped distances calculated from Euclidean embedding (ordinate). The two histograms on left are identical apart from rotation.*

Figure 2 shows a logarithmic density map of the scatterplot of the original versus the dimension reduced pairwise distances between all points in the CoG sample. You see clearly the peak near the diagonal line. The excess of points with mapped distances less than original values can be traced to equation (1) dividing by original rather than mapped distance.

**Implementation**

We used a scaling parallel traditional MPI with threading intranode for MDS implementation. We also used Twister which is a MapReduce extension to support more efficient and broader range of communication collectives (including reduce, gather and broadcast in an MPI language) in the Reduce phase of MapReduce [13]. In Twister, all communication avoids using intermediate disk and is built around ActiveMQ, an Apache publish-subscribe environment, in Java Twister and around Azure primitives in the Microsoft cloud.

The method was applied to obtain a three dimensional projection of 100,000 protein sequences from well-characterized COG clusters. Pairwise distances were calculated using an MPI implementation of the Needleman-Wunsch (NW) alignment algorithm. In later papers, we will compare NW with alternative dissimilarity measures from Blast and Smith Waterman Gotoh. Note that generally used software for NWare rather inefficient. Hence, we implemented a parallel distance computation on the 24-core node system. The efficiency of the parallel distance computation was less than that of MDS due to saturation of memory bandwidth.

Further, we applied a monotone transformation to the pairwise NW distances. To map the data into a 3D Euclidean space, we fed the transformed distances into an MPI implementation of the χ2 MDS. The resulting 3D projection were visualized in PlotViz software [REF].

Each calculation was performed on 768 cores (32 nodes with 24 cores each) of a Microsoft HPC Cluster.

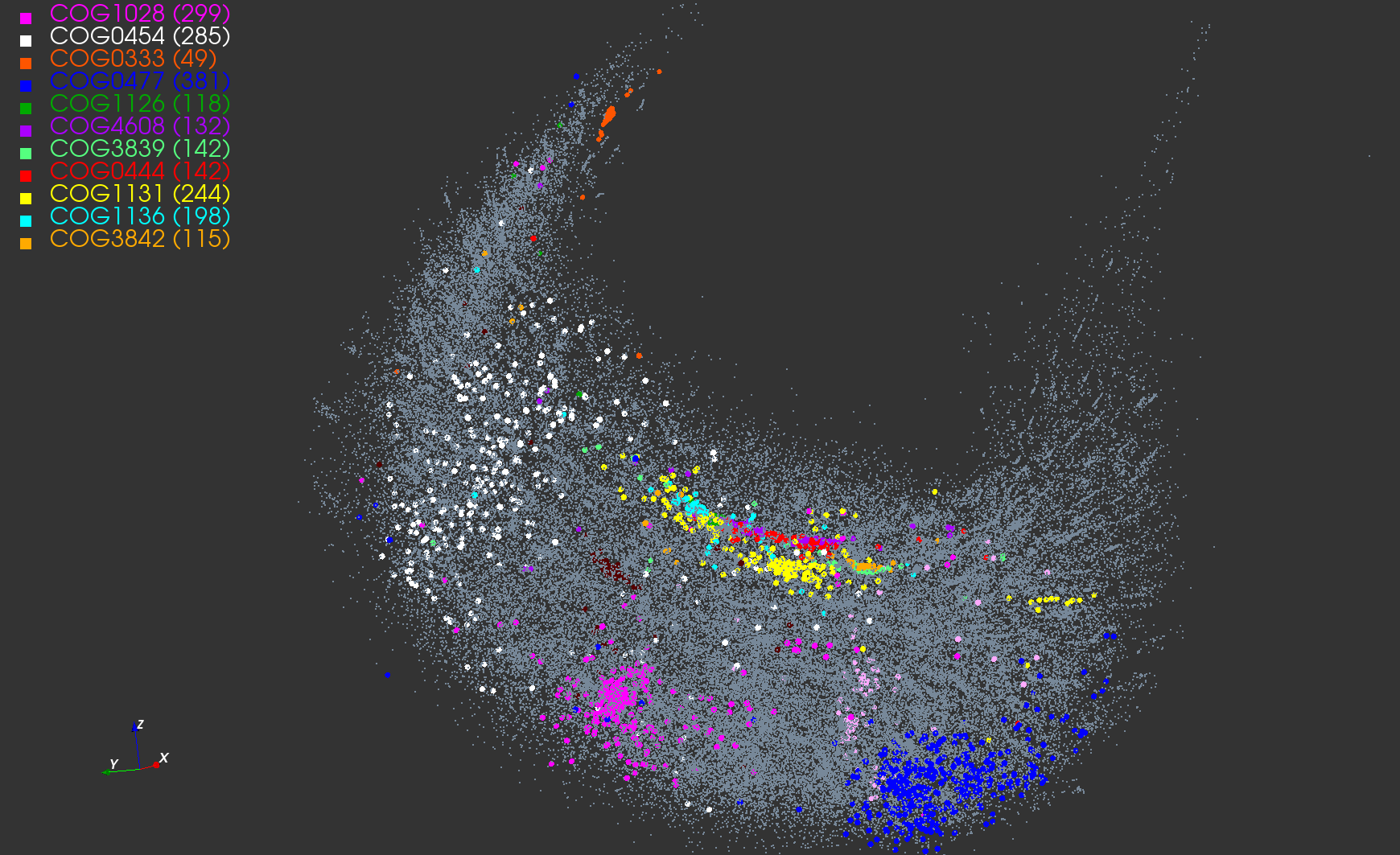
**GEOFFREY: Need more info on processors and cluster.**

The NW distance calculation required one day to complete and the MDS job ran for three days. The parallel efficiency of the code was approximately 60% based on earlier studies that discuss both the inter-node and intra-node cases and find that it is essential to adopt a hybrid model with intra-node threading and MPI between nodes.

**Results**

Figure 1 shows the 3D embedding of the 100,000 prokaryotic protein sequences. Each point represents a particular sequence. To illustrate the results, we overlaid eleven color-coded COG clusters on the map.

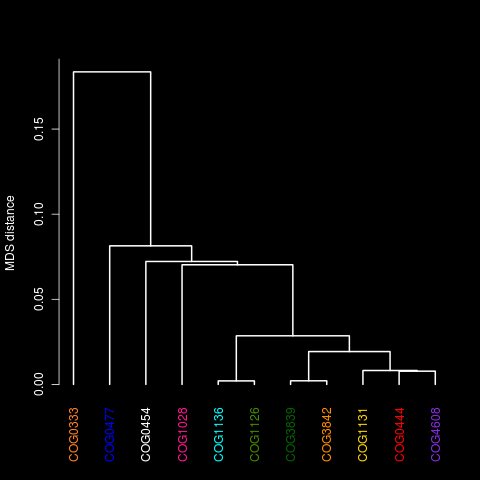
**Figure 1. MDS representation of the 100,000 well-characterized prokaryotic sequences from the COG database. Each point represents a protein sequence. Seven COG clusters were color-coded as marked in the legend. The number of proteins in each cluster is given in parentheses.**



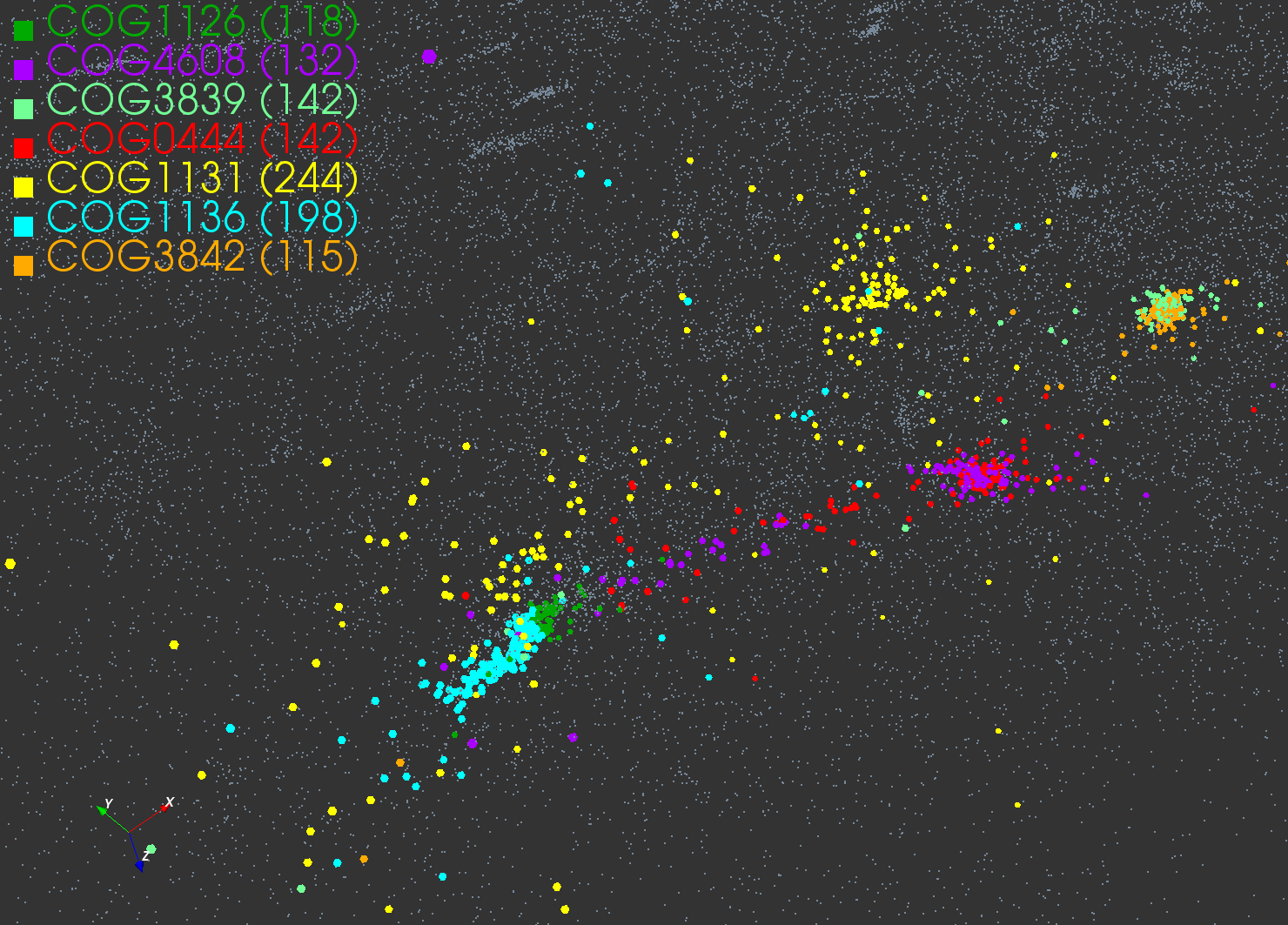
Using the 3D coordinates, we computed the centroids of the eleven clusters and the corresponding distances. The dendrogram tree in Figure 2 shows the proximity of the cluster centroids. Clusters COG1131 (yellow) and COG1136 (cyan) are the tightest with respect to the mean intra-cluster distance. These two clusters are a part of a group that includes seven COGs (right branch of the dendrogram). The other four COGs 1028, 0333, 0477, 0454 appear to be less similar to these group of seven or to each other.

The magnified view in Figure 3 details the neighborhood structure of the COG1131 and COG1136 showing five more COGs lying in close proximity. Remarkably, all seven clusters are functionally similar and correspond to ABC-type transport system, ATPase component (Table 1).

Figure 1 also shows the diversity of protein clusters with respect to their location, shape, dispersion and size. While some clusters are rather tight, others are scattered throughout a sizeable domain. For example, compare the tight COG0333 cluster of ribosomal protein L32 with the diffuse COG0454 (HPA2) and COG0454 (Permeases of the major facilitator superfamily); see also Table 1.

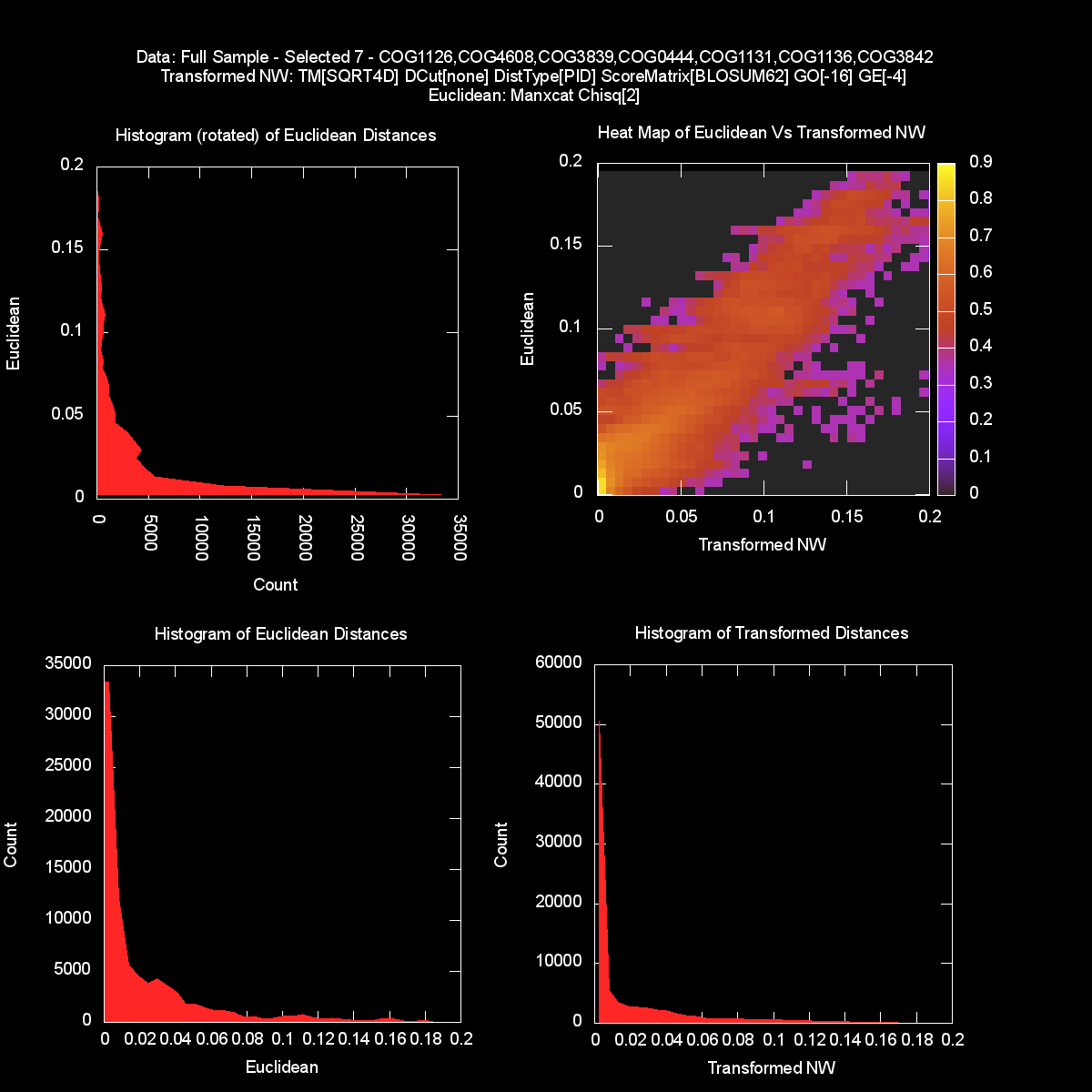


**Figure 2. The dendrogram tree of the cluster centroids. The cluster labels are color-coded as in Figure 1.**



**Figure 3. Magnified version of the prokaryotic database showing the seven functionally similar COG clusters**

From the biological standpoint, this spatial distinction conforms well with clusters' functionality. For example, a tight COG3839 cluster contains 142 proteins sequences of the sugar transport systems that are similar both in function and composition. Similarly, COG1126 of the polar amino acid transport system proteins with very specific functions appears as a very tight cluster. In turn, COG1131 is fairly diffuse as it combines 244 multidrug transport system proteins that differ in the amino acid composition and functional mechanisms. The similarity between different clusters is reflected by the distance that separates them. For example, the two oligopeptide transport systems, COG4608 and COG0444, have similar shape and are located in close proximity to one another.



*Figure 6: Projected Histograms and Density Map comparing the pairwise f(d(i,j)) (abscissa) with mapped distances calculated from Euclidean embedding (ordinate) for points internal to 7 clusters discussed here.*

Figure 6 repeats the diagrams of figure 2 but for the 7 clusters considered here. The distances are much smaller but again good tracking between original and 3D embedded distances.

As mentioned, in our previous work, we used all-versus-all alignment of 10 million UniRef100 proteins to populate the existing COG clusters [Kolker2011]. The last column in Table 1 shows the number of UniRef100 proteins added to each of the eleven clusters from Figure 1. Notably the most diffuse clusters were expanded most.

**Table 1: Annotations for COG clusters in Figures 1-3**

|  |  |  |
| --- | --- | --- |
| **COG** | **Annotation** | **UniRef100 Proteins** |
| COG1131 | ABC-type multidrug transport system, ATPase component | 14,406 |
| COG1136 | ABC-type antimicrobial peptide transport system, ATPase component | 7,306 |
| COG1126 | ABC-type polar amino acid transport system, ATPase component | 4,061 |
| COG3839 | ABC-type sugar transport systems, ATPase component | 4,121 |
| COG0444 | ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component | 3,520 |
| COG4608 | ABC-type oligopeptide transport system, ATPase component | 3,074 |
| COG3842 | ABC-type spermidine/putrescine transport systems, ATPase component | 3,665 |
| COG0333 | Ribosomal protein L32 | 1,148 |
| COG0454 | Histone acetyltransferase HPA2 and Related acetyltransferases | 14,085 |
| COG0477 | Permeases of the major facilitator superfamily | 48,590 |
| COG1028 | Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) | 37,461 |

**Discussion**

Functional protein annotation is one of the most,important and most resource-intensive challenges in biology. The number of sequenced genomes is poised to drastically increase in the next five years. The Earth Microbiome Project alone is expected to sequence 500,000 microbial genomes, which will contain on the order of 1.5 billion protein sequences and half a trillion amino acids. This is well over a 100 fold increase in the number of sequenced microbial genomes and proteins currently contained in GenBank. The i5K Insect and other Arthropod Genome Sequencing Initiative plans to sequence the genomes of 5,000 insects and related species over the next five years, yielding nearly 100 million new protein sequences. Assigning functions to this glut of newly sequenced proteins is an immense computational challenge that requires advanced computational approach with analytic and scaling capabilities.

The proposed visualization method could be instrumental in enabling prompt and reliable annotation and characterization of newly sequenced proteins. The 3D mapping of the protein space allows interactive exploratory data analysis that is a mandatory precursor to statistical modeling and comparisons. The projection provides a unique perspective on the organization of protein space that so far had been largely described by volumes of summary statistics. The agreement between the 3D projection and the COG clusters demonstrate the utility of the proposed methods to provide efficient representation of the protein sequence universe that can be readily visualized and analyzed across domains of interest. Note that the method is not restricted to protein sequences and can be readily adapted to other types of biological data.

For newly sequenced proteins, the neighborhood information could be effectively used to infer their function and provide accurate annotations. For example, UniRef mapping resulted in 70-fold increase in the size of clusters. Clearly, the manual curation of new clusters is a daunting task. By inspecting the expanded universe, one could identify protein features within each cluster and use them to verify the annotations.

The challenges associated with the functional annotation of newly sequenced genomes cannot be solved by the life sciences community alone. A successful and sustainable solution requires a new trans-disciplinary collaborative model that would leverage and adopt most prominent advances of modern sciences. This turn to collective innovation in data-enabled sciences is essential for truly ground-breaking medical discoveries and advances that may benefit public health. Scientific alliances like HPV, RaDiCAL and DELSA [Ozdemir2011] stand to harness the diversity of skills and expertise, quickly and efficiently translating the influx of new data into tangible innovations and long-awaited treatments.

**Acknowledgements**

We would like to thank Elizabeth Stewart, Christopher Moss, Courtney MacNealy-Koch, Maggie Lackey and Randy Salamon for their comments and help with this manuscript. This work was made possible by the support from NSF (under DBI grants 0969929 and 1045040 to E.K.) and NIH (NIGMS grant 5R01 GM076680-02).

**Andrew: Are these the right grants?**

**GEOFFREY: Please add your acknowledgements and funding sources.**

**References**

**I am still formatting the references**

**GEOFFREY: Can you please email me the final version of the PSU main text and the reference file.**