FIGfams: Yet Another Set of Protein Families

by
Folker Meyer[1,2], Ross Overbeek[3], Alex Rodriguez[2]

[1] Argonne National Laboratory, Argonne, IL
[2] University of Chicago, Computation Institute, Chicago, IL
[3] Fellowship for the Interpretation of genomes, Burr Ridge, IL

Journal: NAR or Bioinformatics or BMC Bioinformatics
Reviewers: - O. White, UMaryland
- J. Selengut, TIGR
- M DeJongh, Hope College
- C. Wu or Peter from PIR

Abstract:
We present FIGfams, a new collection of over 100,000 protein families that are the product of manual curation and close strain comparison. The manual curation is carried out by using the Subsystem approach, ensuring a previously unattained degree of throughput and consistency. FIGfams are based on over 950,000 manually annotated proteins. Associated with each FIGfam is a two-tiered rapid, accurate decision procedure to determine family membership for new proteins.

License: FIGfams are freely available under an open source license.
Download: ftp://ftp.theseed.org/FIGfams/
Website: http://www.theseed.org/wiki/FIGfams//

1. Introduction

Progress in DNA sequencing technology has led to an abundance of nucleotide sequences in our databases (1). As the pace of sequencing increases (see, e.g., (2)) so does the importance of creating tools to accurately describe the protein functions encoded in the DNA sequences. These descriptions, or "annotations," are created using a variety of bioinformatics tools and databases. Our most valuable clues to decipher functions of unknown proteins is their comparison with existing proteins in some form (3).

A number of groups are curating large sets of existing genomes using a variety of approaches (4-6), and even more groups are focusing their curation efforts on sets of proteins (6-12). The one common denominator of all these approaches is that they need to rely on automatic propagation of "correct" annotations using bioinformatics techniques because the number of proteins clearly exceeds the available manpower when following the established "one-protein-at-a-time" annotation approach.

The same issues explain why few authors of genome data sets are able to spend time curating their genome data sets in Genbank. For the majority of genomes, new discoveries are not included if made after the time of initial submission. As a result, even as our knowledge of protein function is developing and the number of "hypothetical proteins" is decreasing, the existing genome data sets are most often not updated with new information. Since comparison with existing data is the source for annotations of new data, this situation presents a serious dilemma for the analysis of new sequences.

Tools for the analysis of new sequence data sets are used by many groups worldwide, and the resulting data is again submitted to a number of repositories (3). Most of the tools for analyzing new sequence data rely on comparison with existing sequence data, either directly using, for example, BLAST (13) or using more sophisticated bioinformatics tools (HMMs, (7, 14-16); PSSMs, (17); or integrations of multiple tools, (18)). Both with BLAST-type searches and with more complex representations, the construction and use of protein families are central to most accurate annotation efforts (6-8, 19); see (3) for a discussion.

The common requirement for these approaches is that curation of initial protein sets (in the case of TIGRFAMS (7)) or assignment of protein family functions (in the case of PIRfams or OrthoMCL(20)) needs to be performed by an expert in the protein being analyzed. As with manual curation of complete genomes, however, manpower for the creation of these core data sets has been the limiting factor so far. Curation of data has restricted the number of protein families; for example, in the manually curated TIGRFAM core set, only 1972 *TIGR equivalogs* exist in Release 8.0 of the TIGRFAMs (21).

This bottleneck for manually curating the protein families can be overcome by using the ***Subsystem*** approach (22) for the construction and maintenance of protein families. Subsystem-based curation provides a scalable, alternative to the traditional manual curation efforts for protein families.

1.1. *Subsystems and FIGfams*

Basically, a *Subsystem* is a collection of abstract functional roles and a spreadsheet mapping those functional roles to genes across multiple genomes. The spreadsheet has functional roles as columns, and each row corresponds to a single genome. Each cell contains the genes in the corresponding genome that implement the functional row given by the column. Together, the *Subsystem* and the *Subsystem spreadsheet* are referred to as a *populated Subsystem*. The current collection of manually curated Subsystems includes over 800 subsystems containing over 6,400 functional roles, to which over 950,000 genes are connected; see (5) for details.

The FIGfam effort may be thought of as constructing the infrastructure needed to automatically project the manual annotations maintained within the Subsystem collection.

1.2. *Defining FIGfams*

FIGfams are sets of isofunctional homologues. In other words each FIGfam is supposed to contain a set of proteins that are end-to-end homologous and share a common function. The current release (8.0) contains roughly 113,000 families, from careful manual curation using Subsystems (22) and automatic annotation of closely related strains. The families from closely related strains are based on sequence similarity and conserved genomic context. Figure 1 gives an overview of FIGfam creation and the use of FIGfams for automatic annotation.

More formally, each FIGfam can be defined as a 4-tuple: (*ID*, *protein-set, decision-procedure, family-function),* where

1. the *ID* is a stable, unique identifier that describes the family and allows linking to a web site describing the protein family;
2. the *protein-set* is a set of protein sequences that are similar over essentially their entire length (i.e., they share a common domain structure; we allow for slight differences in the C-terminal because the correct determination of start codons is still somewhat imperfect and would artificially split protein-sets otherwise belonging to the same family) and are believed to implement a common function;
3. the *decision-procedure* is a decision procedure that, given a new protein sequence as input, will decide whether the new sequence should be considered as “part of the same family”; and
4. the *family-function* is the function believed to be implemented by all members of the protein-set.

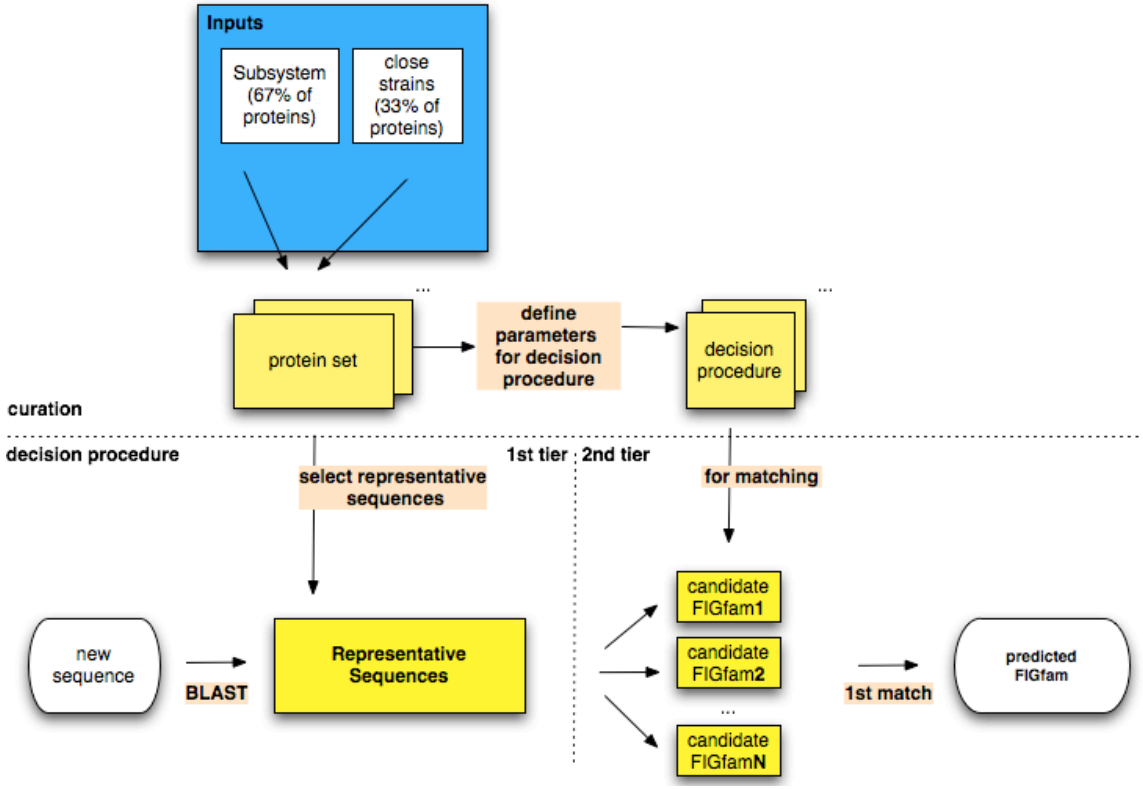

Figure 1: Overview of FIGfam creation and use. From two different input sets, FIGfam protein sets are defined. Subsequently, parameters for decision procedures and representative proteins are computed. On the decision procedure side, a new sequence is classified via BLAST searching the representative database (1st tier); then the decision procedure associated with each candidate FIGfam is executed (2nd tier); and finally the highest score for any candidate FIGfam is reported back.

1.3. *Creation and Maintenance of FIGfams*

The construction of FIGfams is based on forming *protein-sets* in cases in which it can more or less reliably be asserted that sequences implement identical functions. Currently, there are two scenarios for creating a FIGfam: based on Subsystems and based on closely related strains.

The FIGfams are constructed by inferring which pairs of genes must be placed in the same FIGfam (see below for detailed discussion in each of the scenarios) and then forming the set of FIGfams as the maximum set of protein-sets consistent with the pairwise constraints.

1.3.1. Families Constructed from Subsystems

Two proteins will be placed in the same FIGfam if they are similar over their entire length and they occur within the same column of a Subsystem (see Figure 2). Further, if two sequences occur in Subsystems, but do not share a common function (i.e., are not connected to the same column in any Subsystem and have distinct assigned functions), then they cannot be part of the same FIGfam.

A postprocessing step checks for the rare case in which a resulting FIGfam contains two protein sequences, each of them contained in Subsystems, having differing functions. Such a case is evidence of an error in the curation of the relevant Subsystems and is corrected manually.

	Organism	Domain	Variant	argA/J	argB	argC	argD	argE	argF/I	argG	argH	ArgJ	ArgR
		Bacteria		All	All	All	All	All	All	All	All	All	All
☐	Nocardia farcinica IFM 10152 (247156.1)	Bacteria	1	1942	1943	1941	1944	4560	1945	1967	1968	1942	1946
☐	Rubrobacter xylanophilus DSM 9941 (266117.6)	Bacteria	1	2525	2526	2524	2527	1493 2176	2528	2521	897	2525	2523
☐	Bacillus anthracis str. 'Ames Ancestor' (261594.1)	Bacteria	1	4320	4319	4321	4318	2417 766	4317	4815	3608 4814	4320	4361
☐	Bacillus cereus ATCC 14579 (226900.1)	Bacteria	1	3958	3957	5486	3956	2123	358 3955	4448	3395 4447	3958	356
☐	Bacillus halodurans C-125 (272558.1)	Bacteria	1	2899	2898	2900	2897	1059 2678	2894	3187	3186	2899	
☐	Listeria monocytogenes str. 1/2a F6854 (267409.1)	Bacteria	1	2095	2094	2096	2093	1809	2092	1114	1115	2095	1263
☐	Lactobacillus plantarum WCFS1 (220668.1)	Bacteria	1	441	442	440	443	1597 2351	444	657	658	441	1166 1331
☐	Rhodopseudomonas palustris CGA009 (258594.1)	Bacteria	1	3604 589	626	2476	3712 4750	2313 3015	4749	389	4720	589	
☐	Pelagibacter ubique HTCC1062 (335992.3)	Bacteria	1	878	721	1238	812	466	813	57	522	878	
☐	Geobacter metallireducens GS-15 (269799.3)	Bacteria	1	1027	234	636	235		236	237	240	1027	

Figure 2: FIGfams from a Subsystem. The manual curation of the Arginine Biosynthesis Subsystem led to the creation of multiple FIGfams. The colored background indicates FIGfam membership. A single column can contain multiple FIGfams.

These rules firmly ground the FIGfams in the manual curation effort maintaining the Subsystems. If at any point it appears that (because of either of the non-Subsystem grouping strategies discussed below) two proteins are part of a single protein-set but are believed to implement distinct functions, the solution is to make sure that Subsystems exist to which the proteins are attached. That is, if there is a solid reason to believe that the proteins implement different functions, the way to force this to occur within FIGfams is to make sure that the manual effort reflects the reasoning that the functions are distinct.

1.3.2. Families Constructed from Closely Related Strains

If two or more sequenced genomes are from closely related strains, it is usually possible to trivially establish a reliable correspondence between 90% and 95% of the genes within the genomes. This can is illustrated by the display of corresponding regions from the genomes shown in Figure 3. Of course, one needs an automated tool that uses specific rules to detect reliable correspondences, and many have been constructed.

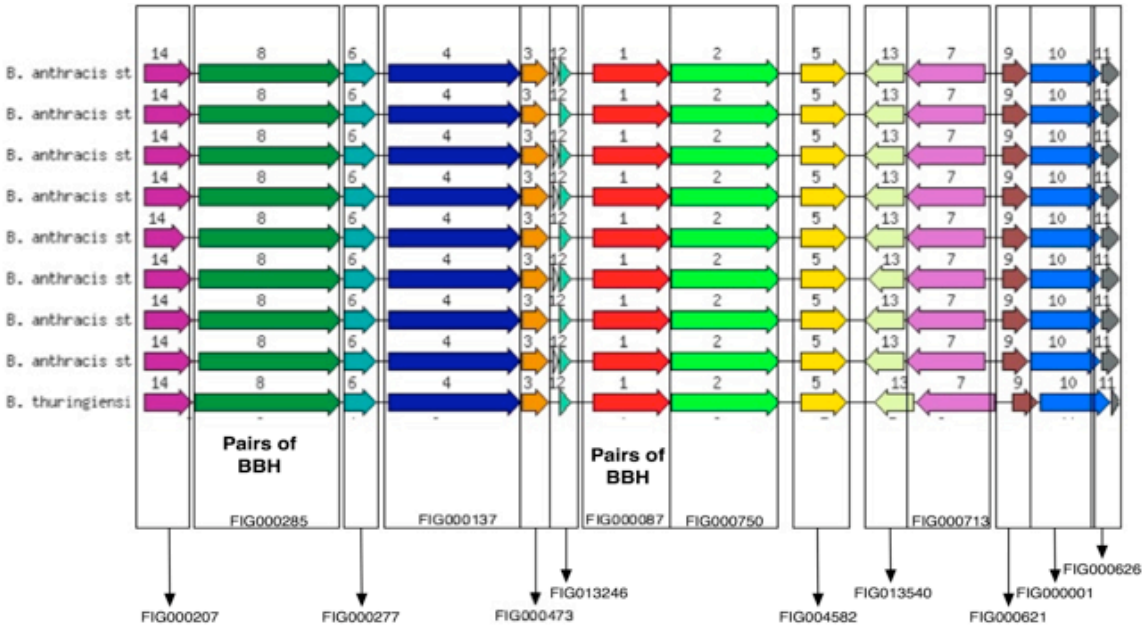

Figure 3: FIGfams constructed from closely related strains – The graphic depicts the chromosomal neighborhood of histidyl-tRNA synthetase in closely related bacillus. Same color indicates a set of similar proteins (bi-directional best hits) that form a FIGfam. Each of the FIGfams has a different functional role; see Table 1 in the Supplemental Material.

We use a simple tool to implement this process; a description is provided in the appendix. We note that as more genomes are sequenced from closely related strains, the number of correspondences due to this type of rule will rapidly grow.

1.4. *Curating FIGfams over Time — Connecting Changes in Subsystems to Changes in FIGfams*

The current set of FIGfams will rapidly become outdated as the characterization of specific proteins continues to improve. New experimental results reported in the literature and careful manual annotation of the current Subsystems naturally force changes and additions to the FIGfam collection. A central feature of the existing collection is that families will automatically be split, merged, and added in response to the addition of new Subsystems or corrections of errors in the existing collection. In a field experiencing such rapid advance, this automated coupling of changes in Subsystems to derived changes in FIGfams is vital.

Once each month, the existing FIGfams are scanned looking for cases in which a family contains two proteins such that both proteins occur in Subsystems and the functions of the proteins are not identical. Such a situation forces a split of the FIGfam, which can be achieved automatically. Similarly, if two families are found to each contain proteins that occur in Subsystems and if these proteins are globally similar and implement identical functions, then the families are automatically merged. As new Subsystems are implemented, we find cases in which globally similar proteins that all connect to Subsystems and implement the same function are not yet members of any FIGfam. If they can be added to existing FIGfams, they are; if not, new FIGfams are automatically created.

2. Comparison of FIGfams, PIRSFs, and TIGRFAMs

Many groups have attempted the curation of protein families over time; here we discuss the differences and similarities among three prominent efforts. All three efforts curate families of homeomorphic proteins, requiring full-length sequence similarity and common domain structure within each family. These requirements make them different from efforts such as the PFAM database (14) that provide protein domains.

The technologies used for curation are very different, resulting in vastly different throughput of the various protein family curation strategies. In the case of FIGfams roughly 23,000 protein families are the results of manual curation; that number is in stark contrast to 1,900 TIGRFAM equivalogs with manually curated SEED alignments.

It is our understanding that curation of TIGRFAMs starts with manual creation of a SEED alignment, the creation of an HMM from said alignment. The subsequent curation effort is the definition of thresholds that allow the HMM-based decision procedure to reliably detect new members of the protein family.

The PIRSF discusses the formation of a shallow hierarchy (superfamilies, containing families, containing subfamilies). The goal is somewhat different from, and perhaps more ambitious than, that present in FIGfams. The PIRSF hierarchy attempts to group things into a hierarchy based on physical properties, realizing that significant shifts in physical properties usually correlate closely with functional properties. The FIGfams are based on the *Subsystems* view (22) in which the cell is composed of a set of functional Subsystems, and each active variant of a Subsystem is thought of as a set of functional roles. Proteins implement one or more functional roles. Grouping sets of functional roles induces the shallow hierarchy imposed by Subsystems.

Both notions involve protein families made up of proteins that are globally (i.e., full-length) similar. In most cases, the lowest-level PIRSF family (either a family or a subfamily) is composed of proteins that are believed to implement a common function. Hence, we believe there exists a close correspondence between the families produced by the two efforts, and the correspondence will improve as uncertainties are gradually eliminated. At this point the differences in perspective become most apparent in the way

families are constructed. In the FIGfam effort, the major concern is to avoid placing two proteins with different functions into the same set. This leads to many small protein families (and many distinct families that contain closely similar sequences). In a somewhat oversimplified view, the families built by PIR are large groupings of homologous proteins in which the precise, distinct functions of subfamilies are gradually worked out, whereas the FIGfams start with no groups and conservatively gather proteins of identical function. To provide perspective on what this means, we note that the FIGfams collection now includes over 100,000 families, over half of which contain three or fewer members, whereas the PIRSF contains 32,000 families.

All three groups maintain sets of proteins and suggest a function for all members of that family. For TIGRfams, the set of proteins is used for the SEED alignment subsequent used to create an HMM. For FIGfams and PIRSF, complete sets of proteins are maintained for each family. Table 1 lists further differences and similarities.

Table 1: Comparison of protein family creation and maintenance.

	FIGfams	**PIRSFs**	**TIGRFAM**
Family creation	Via Subsystem curation and close strains	Via automatically generated sets of homeomorphic proteins incorporating protein domain knowledge	Via manually curated SEED alignment
Extending an existing family	Include new genomes in Subsystem	Automatic placement in homeomorphic family[3]	Adjust threshold for trusted HMM score
Creating new families	Via new Subsystem creation		New SEED and HMM
Curation of function (for all proteins in set)	Via Subsystem inclusion	Define family function for set	
Number of families	107,233	33,599	3603
Families with proteins with manually curated function	20,699	327[2]	1920
Number of proteins in manually curated families	970,682	6040[2]	–[1]

1) TIGRFAM protein sets are not curated; only SEED sets and HMM thresholds are curated.
2) info from ftp://ftp.pir.georgetown.edu/pir_databases/pirsf/data/pirsf_full_validated_oo.readme
3) info from http://pir.georgetown.edu/pirwww/about/doc/tutorials/pirsftutorial.ppt
*) homeomorphic = full length homologuous with common domain architecture

3. FIGfam Decision Procedures

The second component of each FIGfam, the *decision-procedure*, is used to answer the question "For a new sequence X, should X be considered part of the FIGfam?"

Various technologies exist to implement this decision procedure, ranging from a simple "take the best BLAST hit" approach to more sophisticated approaches using machine learning technology such as hidden Markov models and position-specific scoring matrices.

The current implementation of the FIGfam decision is two-tiered. A global fast screening procedure will create a set of candidate FIGfams for a target sequence *X*. A slower, more accurate decision procedure is associated with the individual family.

In order to provide maximum throughput, the initial screening is implemented by using a database of representative sequences for all FIGfams (see the Supplemental Material for details on the construction of this database). Each candidate FIGfam has its own decision procedure; we currently implement two distinct procedures.

Similarity bounds decision procedure – A bounds list is generated for each member of the protein family by using the learning data. The bounds list is essentially a threshold for trusted BLAST scores, in which the user can safely assign a functional role if the blast BLAST score falls below a designated threshold. The decision procedure goes through the closest BLAST hits in the family (from the sequence being considered for membership), and the individual members are examined to see whether the hits fall within the "safe" threshold. If there are ever more "safe" hits than those that are not, the process ends successfully. Otherwise, the sequence cannot reliably be assigned to the set.

BLAST voting decision procedure – The top 10 and top 20 BLAST results are voted on to select the functional role with the most hits. When two or more functional roles have an equal number of votes as the top choice, no assignment is given.

HOW DO WE CHOOSE THE DECISION PROCEDURE FOR EACH FIGFAM

We have evaluated a series of decisions procedures when designing and implementing FIGfams. In the remainder of this section, we revisit some of the issues that led to the existing implementation. The first test targets pure runtime performance, the second tests for robust classification performance in the face of noisy data, and the third compares performance with two related protein family efforts.

3.1. *Test 1: A Simple Case — Finding Ribosomal Protein L33p*

FIGfam FIG000053 has a family function of LSU ribosomal protein L33p. The decision procedure for this family is straightforward. The central issue in choosing a decision procedure is just performance.

Specifically, we evaluated the use of an HMM as opposed to the use of BLAST using the set of sequences (1,313 sequences) from the FIGfam FIG000053. The BLAST test was performed by first BLASTing the sequences against a set of representative sequences of the FIGfams. Subsequently, the resulting FIGfams in a threshold were further evaluated by BLASTing against the specific FIGfams. The HMM test was performed by doing an HMM search against the set of HMMs for each FIGfam. Currently, there are over 100,000 FIGfams, which hampers the ability to perform an HMM search on the FIGfams.

Both methods generated the same prediction for the test sequences. Runtime requirements were dramatically different, however, with the HMM procedure taking significantly more resources. As shown in Figure 4 the time required for the HMM case depends on the number of families searched, whereas the two-tiered FIGfam strategy requires linear time. Note that the initial rapid screening stage does not show significant variations with growing numbers of families.

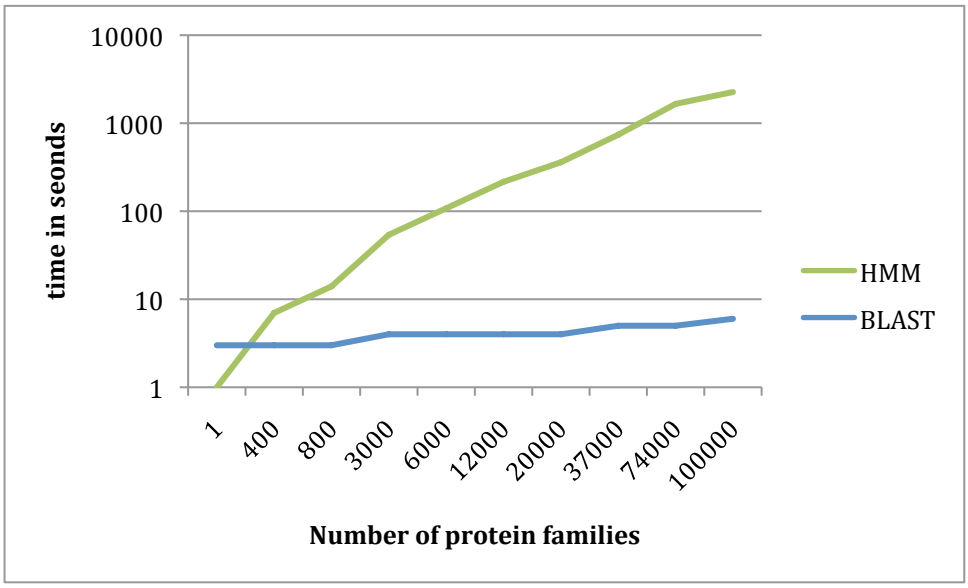

Figure 4: Comparison of the time (in seconds) spent searching growing numbers of FIGfams via HMMs and BLAST. Time for searching via HMMs increases with the number of families, for the two-tiered BLAST based decision procedure, the time is constant. The time required to perform the search with BLAST remains at less than 10 seconds. These computations were performed on a current desktop machine running Linux (3Ghz Intel CPU, 4GB RAM, with faster CPUs the ratios will remain stable).

3.2. *Test 2: HisAb — Using Different Decision Procedures to Distinguish between Two Similar Proteins*

Curation of data is an error-prone process, and any decision procedure employed to recognize new family members (or predict functions for novel genes) is likely to include erroneously annotated false positive members. We therefore have devised a test to "poison" carefully verified protein families with errors.

Several decision procedures were tested using protein sequences from a protein family implementing the function *Histidyl-tRNA synthetase (EC 6.1.1.21)* and a second family implementing *ATP phosphoribosyltransferase regulatory subunit (EC 2.4.2.17)*. These two families contain protein sequences that are similar to one another (see (23) for details), which has led to numerous errors in annotation over the past few years.

We call the union of these two families the HisAb set, and we consider in detail the issue of how well different decision procedures separate the entries of *HisAb* into the two protein families. The *HisAb* set offers a framework for comparison in that the sequences are closely homologous, and we believe that we have accurately annotated (manually) the entire set of sequences. The FIGfam FIG000087 implements *Histidyl-tRNA synthetase (EC 6.1.1.21),* and FIG000865 implements *ATP phosphoribosyltransferase regulatory subunit (EC 2.4.2.17)*.

One obvious way to evaluate each decision procedure would be to take each sequence from the *HisAb* set, delete the sequence from the family containing it, and then examine the results of asking, for each of the two families, "Does the sequence belong in this family?" For each of these single-sequence experiments there are four possible outcomes: the decision procedure can place the sequence into FIG000087, FIG000865, both, or neither. If we perform this experiment for each sequence in each of the two families, for each of the decision procedures we wish to evaluate, we gain some insight into the relative merits of the set of decision procedures (we display the results of this experiment below). However, we can also investigate the situation in which some percentage of the sequences in *HisAb* has been assigned to the wrong protein family. This more closely resembles the real situation for most paralogous families, and we believe that it offers a more comprehensive way to evaluate the relative merits of the decision procedures.

Overall, the decision procedures that performed the best in the presence of misannotated sequences were the BLAST voting algorithms (top 1, top 20 BLAST results). The number of BLAST hits to vote on was directly proportional to the size of the protein family being tested. The HMM decision procedure was outperformed by all other decision procedures, and it was also more time consuming.

3.2.1. Test Methodology

Each decision procedure was tested by using a jack-knife approach, where a sequence was used for testing the decision procedure, while the rest of the sequences were used as the learning data to create the model. This process was iterated several times over the number of total sequences in the learning data. In addition to experimenting with each decision procedure using the gold standard, errors were introduced to the gold standard

assignments by switching a sequence's assigned functional role in the learning data. The goal is to view how each decision procedure behaves in the presence of errors in the annotations using a controlled environment. Each decision procedure was tested with 0%, 10%, 20%, 30%, and 40% annotation errors in the learning data. The accuracy, sensitivity, and specificity measurements were calculated in order to compare the results of the different decision procedures for each of the protein families tested. The sensitivity measures how well a binary classification test correctly identifies a condition. The specificity measures how well a binary classification test correctly identifies the negative cases, or those cases that do not meet the condition under study. The specificity, sensitivity, and accuracy measures were calculated by counting the number of true positives (tp), true negatives, (tn), false positives (fp), and false negatives (fn).

For FIGfams, we used the built-in method, in this case the BLAST voting procedure; for TIGRfams, we used the HMMs provided with the cut-off values; and for PIRSF, we used the decision procedure provided by PIR.

3.2.2. Test Results

For FIG000087, the sensitivity of all three procedures is identical without errors present. As errors are introduced into the data set, the BLAST voting procedure clearly outperforms the other procedure with almost no loss of specificity at 20% errors and a .96 specificity rate at 30% errors (see Figure 5).

The Figure should be 5

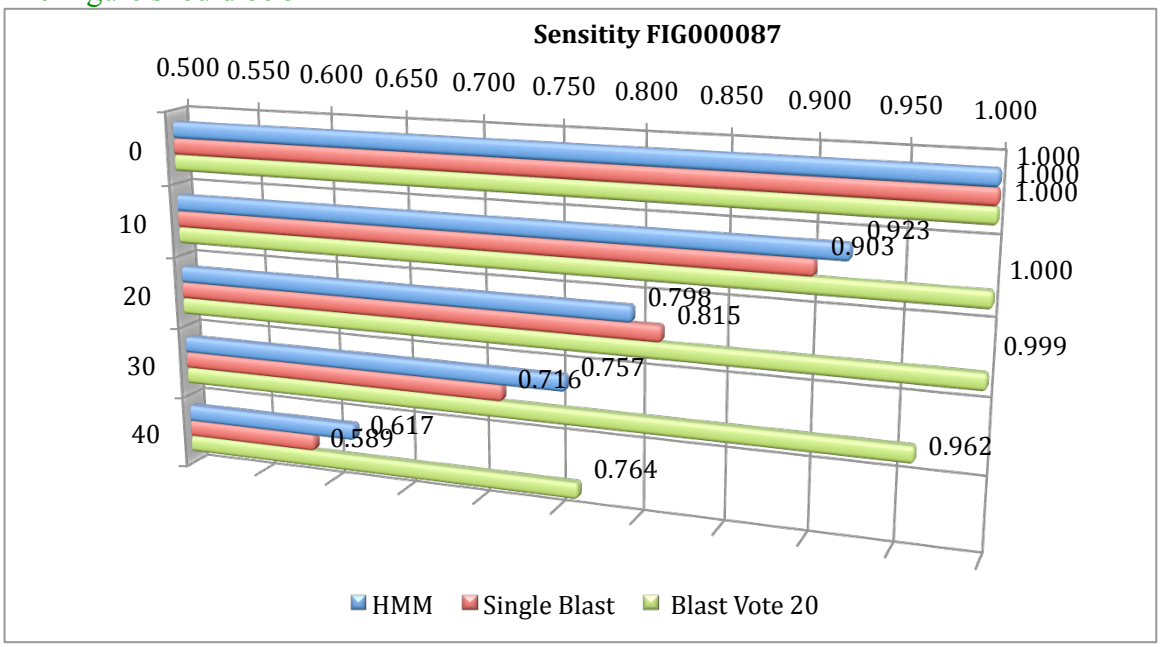

Figure 5: Sensitivity results in the presence of varying error rates for FIG000087. The BLAST voting procedure clearly outperforms the alternatives, losing virtually no sensitivity at 20% error.

The specificity, or rate of false positive predictions, is another important performance measure for a classification tool. Again the BLAST voting procedure clearly outperforms the simple BLAST and also the HMM (see Figure 6).

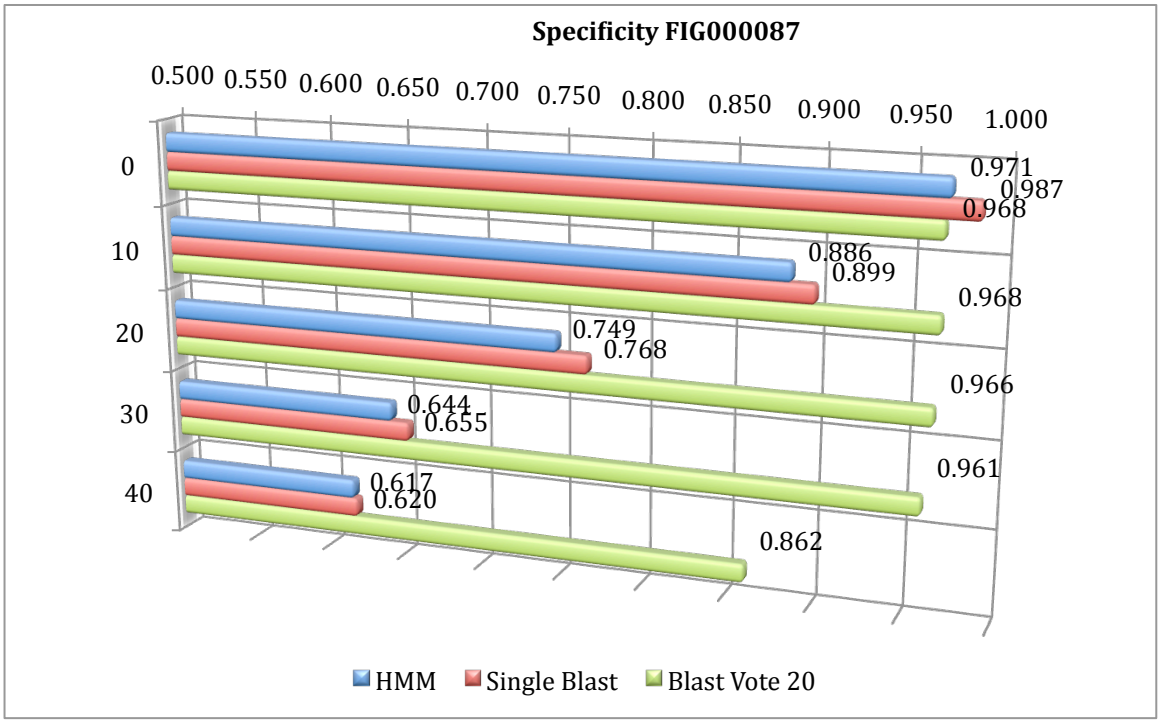

Figure 6**:** The Specificity for protein family FIG000087 in the presence of varying error rates. All three procedures perform equally well in the presence of no errors, but performance drastically drops once errors are introduced into the protein family data. The BLAST voting procedure clearly outperforms the other procedures with >0.96 specificity in the presence of 30% errors.

In some cases the decision procedure associated with a FIGfam will not be the BLAST voting procedure. Instead we use the similarity bounds procedure described above. Figure 7 shows the performance characteristics of this procedure. Similarity bounds provide very good specificity (>.914 for 50% errors in the data), but the sensitivity degrades badly with increasing error rates. The complete results are available in the appendix.

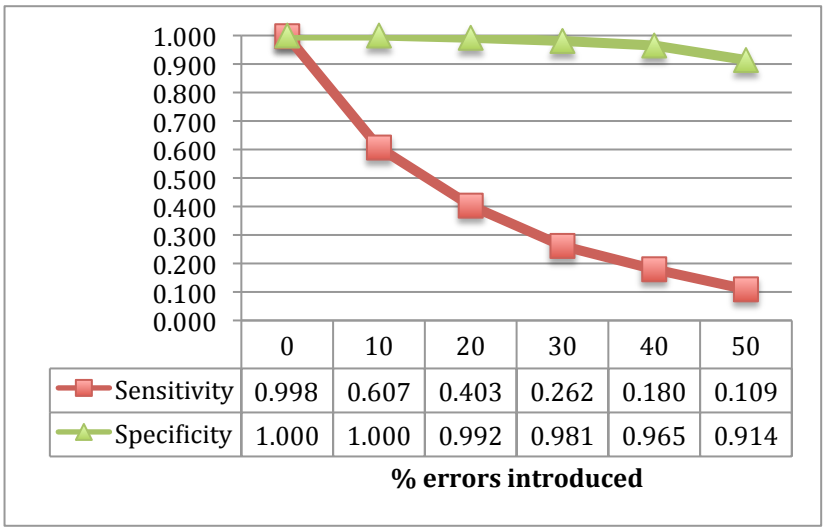

Figure 7: The similarity bounds procedure has very good specificity in the face of errors, but the specificity degrades rapidly in the presence of errors.

3.2.3. Discussion of Test 2

The BLAST voting procedure clearly outperforms HMMs in the chosen example. Since we do not include this procedure with all FIGfams, however, we also show data for the similarity bounds procedure. This is a very conservative procedure with very poor sensitivity. The decision to use such a procedure in complicated cases was taken to minimize the noise introduced into the predictions.

3.3. *Test 3: Comparing FIGfams, TIGRFAMs, and PIR HMMs*

NOTE - the reference should be to Table 2.

The efficiency of the decision procedures from different groups such as PIRSF [1], and TIGRFAM [4] was tested to evaluatte the accuracy of the three databases. Table 1 provides the number of protein families intersecting the contents of the *HisAb* set that is associated with each database. TIGRFAM release 7.0 provides the public with a set of protein families covering a range of functional roles. TIGRFAM's preferred decision procedure is a set of HMMs that provide a trusted and noise cutoff score indicating the ranges for which the results can either be trusted or not as the said functional role. The HMMER package was used to make an assignment. PIRSF (July 2007 release) also provides a set of HMMs along with a decision procedure for its protein families. The decision procedure uses the available HMMs together with BLAST results to assign a sequence to a PIRSF group.

Table 2: HisAb protein families from TIGRFAM, PIRSF and FIGFAM.

Protein Family Group	Name	Description
PIRSF	PIRSF001549	histidyl-tRNA synthetase [Validated]
TIGRFAM	hisS hisS_second	histidyl-tRNA synthetase histidyl-tRNA synthetase
FIGFAM	FIG000087 FIG000865	Histidyl-tRNA synthetase (EC 6.1.1.21) ATP phosphoribosyltransferase regulatory subunit (EC 2.4.2.17)

The TIGRFAMs, PIRSF protein families, and FIGFAMs provide a number of related families that intersect with the contents of *HisAb*. The same *HisAb* sequences used in the previous section were used to compare the accuracy, sensitivity, and specificity against the three different protein family groups' decision procedures. The main interest is to test whether the manually curated *HisAb* sequences were correctly characterized as any of the respective *HisAb* families in TIGRFAM, PIRSF, and FIGfam. A similar procedure was used to count the *tp*, *fp*, *tn,* and *fn* as before. The same equations as in the above section were used to calculate the accuracy, sensitivity and specificity of the *HisAb* families. The specificity comparisons between the *HisAb* protein families from FIGFAMS, PIRSF, and TIGRFAMS show that none tries to overpredict the sequence's functional role (keep false positives to a minimum). However, using the FIGFAM decision procedures resulted in more functional roles assigned correctly (sensitivity). A summary of the comparisons is shown in Figure 8.

Note – this should be Figure 8. But the figure seems a bit weird – why is the top half of the figure blank? Why not choose a better unit? Also, the axis labels get smothered in the rest.

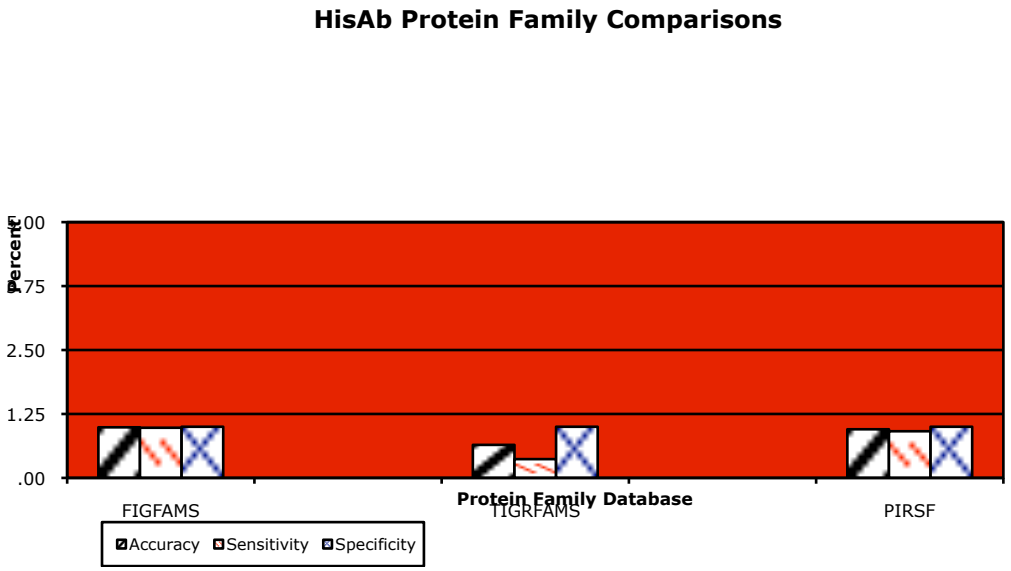

Figure 8: Comparisons of FIGfams, TIGRFAMS, and PIRSF hisAb protein families. The high specificity levels on each protein group indicate that no group tries to overpredict the functional roles of the sequences. Both PIRSF and FIGfams performed well in annotating the correct function. The TIGRFAM

predictions used the HMM trusted and noise cutoff scores to assign the functions. The top hits of most TIGRFAM's nonassigned sequences was one of the hisAb families; however, it failed to annotate it as a member of the HisAb family because it missed the noise-cutoff score.

3.4. *Test 4: Another Comparison of FIGfams and PIRfams*

Alex will provide another set of data that shows FIGfams outperforming PIRSF for some stuff.

4. Summary and Discussion

The propagation of errors from the sequence databases has been a significant problem in genome annotation and other areas. Several techniques have been used to handle the results of the noise in the databases. We present a novel solution to the problem by providing a set of protein families that can be used for automatic annotation and other purposes based on a set of consistent manual high quality annotation. The fact that Subsystems cover approximately 50% of the known bacterial and archaeal protein space make FIGfams a very useful resource. By allowing for variable decision procedures on a per family basis, we have ensured rapid processing at a rate that enables the annotation of several genomes per day on a current desktop machine.

The primary benefits of our approach are as follows:

- FIGfams are fast, reliable, and robust to noise in the data. Moreover, as more diverse genomes are sequenced and annotated, the speed and accuracy of FIGfam based annotation will increase.

- The time to classify a single protein averages around 10 seconds on a modest desktop machine, allowing processing of ~8,640 proteins per day on a single machine.

- The BLAST voting procedure, the most frequently used decision procedure for FIGfams, performs at least as well as simple BLAST and HMM-based procedures. If errors are present in the data set, it outperforms all other procedures.

- FIGfam performance is optimized to minimize false positive assignments.

- As Subsystems cover more and more of the known protein space, the FIGfams will increase in value over time. In addition, new results from the literature are incorporated into the FIGfams via Subsystem curation, guaranteeing that the FIGfams remain up to date.

As the number of proteins in FIGfam increases, automatic annotation pipelines such as RAST (24) will be able to reduce the number of genes subjected to costly in-depth database searches. Thus, **by linking accurate, mass creation of protein annotations and protein family construction using Subsystems we have achieved a previously unseen productivity and accuracy in protein family creation.**

5. Availability

The FIGfams have been used as a central component in the RAST server (25) (http://RAST.nmpdr.org), a system that provides rapid, accurate annotation of prokaryotic genomes. They are also used in MG-RAST(26), a public server focusing on the annotation of metagenomic data (http://metagenomics.nmpdr.org).

Release 9 of FIGfams is made freely available to anyone for any use. It contains 1,389,417 proteins grouped into 107,233 families. All families can be downloaded from ftp.theseed.org/FIGfams/.

Running the FIGfam decision procedure locally requires a LINUX/UNIX/OS-X operating system and Perl 5.6 or greater.

Acknowledgments

Part of this project has been funded with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400042C. Argonne National Laboratory's work was supported under U.S. Department of Energy contract DE-AC02-06CH11357. We thank Veronika Vonstein and her team of expert Subsystem annotators.

References

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2007;35(Database issue):D21-5.
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.
3. Overbeek R, Bartels D, Vonstein V, Meyer F. Annotation of Bacterial and Archaeal Genomes: Improving Accuracy and Consistency. Chemical Reviews. 2007.
4. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35(Database issue):D61-5.
5. McNeil LK, Reich C, Aziz RK, Bartels D, Cohoon M, Disz T, et al. The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. Nucleic Acids Res. 2007;35(Database issue):D347-53.

6. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008;36(Database issue):D480-4.
7. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, et al. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res. 2007;35(Database issue):D260-4.
8. Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, et al. PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res. 2004;32(Database issue):D112-4.
9. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC bioinformatics [electronic resource]. 2003;4:41.
10. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research. 2003;13(9):2178-89. PMCID: 403725.
11. Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, et al. The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Res. 2003;31(1):452-5. PMCID: 165509.
12. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, et al. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008;36(Database issue):D419-25. PMCID: 2238974.
13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389-402.
14. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic Acids Res. 2004;32(Database issue):D138-41.
15. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, et al. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 2005;33(Database issue):D284-8. PMCID: 540032.
16. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 2004;32(Database issue):D226-9. PMCID: 308773.
17. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, et al. CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res. 2005;33(Database issue):D192-6. PMCID: 540023.
18. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. New developments in the InterPro database. Nucleic Acids Res. 2007;35(Database issue):D224-8.
19. Schneider M, Tognolli M, Bairoch A. The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools. Plant Physiol Biochem. 2004;42(12):1013-21.
20. Chen F, Mackey AJ, Stoeckert CJ, Jr., Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 2006;34(Database issue):D363-8. PMCID: 1347485.
21. TIGRFAM 8.0 release notes. 2009 [updated 2009; cited]; Available from: ftp://ftp.jcvi.org/pub/data/TIGRFAMs/RELEASE_NOTE_8.0.

22. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005;33(17):5691-702. PMCID: 1251668.
23. Sissler M, Delorme C, Bond J, Ehrlich SD, Renault P, Francklyn C. An aminoacyl-tRNA synthetase paralog with a catalytic role in histidine biosynthesis. Proc Natl Acad Sci U S A. 1999;96(16):8985-90.
24. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008.
25. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75. PMCID: 2265698.
26. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC bioinformatics [electronic resource]. 2008;9:386. PMCID: 2563014.

The submitted manuscript has been created by UChicago Argonne, LLC, operator of Argonne National Laboratory (“Argonne”). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

PLEASE Give Judy B. a copy to clear.