Validating Annotations for Uncharacterized Proteins in *Shewanella oneidensis*

Brenton Louie,¹ Peter Tarczy-Hornoch,¹ Roger Higdon,² and Eugene Kolker^{1,2}

Abstract

Proteins of unknown function are a barrier to our understanding of molecular biology. Assigning function to these “uncharacterized” proteins is imperative, but challenging. The usual approach is similarity searches using annotation databases, which are useful for predicting function. However, since the performance of these databases on uncharacterized proteins is basically unknown, the accuracy of their predictions is suspect, making annotation difficult. To address this challenge, we developed a benchmark annotation dataset of 30 proteins in *Shewanella oneidensis*. The proteins in the dataset were originally uncharacterized after the initial annotation of the *S. oneidensis* proteome in 2002. In the intervening 5 years, the accumulation of new experimental evidence has enabled specific functions to be predicted. We utilized this benchmark dataset to evaluate several commonly utilized annotation databases. According to our criteria, six annotation databases accurately predicted functions for at least 60% of proteins in our dataset. Two of these six even had a “conditional accuracy” of 90%. Conditional accuracy is another evaluation metric we developed which excludes results from databases where no function was predicted. Also, 27 of the 30 proteins’ functions were correctly predicted by at least one database. These represent one of the first performance evaluations of annotation databases on uncharacterized proteins. Our evaluation indicates that these databases readily incorporate new information and are accurate in predicting functions for uncharacterized proteins, provided that experimental function evidence exists.

Introduction

PROTEINS OF UNKNOWN FUNCTION comprise 30–40% of proteins predicted from virtually any genome (Bork, 2000). This impedes our understanding of organisms at the molecular level, an important focus of biological research. In this work we begin to address the difficult challenge of assigning function to these proteins, which is known as *annotation*. For this we used *Shewanella oneidensis*, a bacterium studied for its bioremediation potential (Heidelberg et al., 2002), as an example organism. After the *S. oneidensis* genome was initially sequenced, about 40% of its proteins were of unknown function (uncharacterized). To gain insight into organisms such as *S. oneidensis*, biological researchers struggle to annotate uncharacterized proteins (Kolker et al., 2005). This can be a daunting task given the challenges in the annotation process and the scale of the problem.

Biological researchers commonly search either protein or “pattern” databases (Attwood, 2000) to annotate proteins. We refer to both as “annotation” databases. These annotation databases operate by “extrapolating” function (Valencia, 2005) based on sequence similarity. Their predictions can be unreliable (Galperin and Koonin, 1998). Error estimates in protein databases are as high as 30% (Devos and Valencia, 2001), for instance. Validation is also difficult as predic-

tions from various annotation databases may not agree, and supporting evidence can be hard to track (Karp et al., 2001). Because it is unclear which databases are most accurate, annotation is difficult. While the gold standard for annotation is biological experimentation, predictions based on sequence similarity remain necessary. Experimental function characterizations happen slowly even for *Escherichia coli* (Thomas, 1999), an aggressively studied organism. In light of this, the accuracy of annotation databases should be assessed, given the potential negative consequences of inaccurate annotations (Koonin and Galperin, 2003). However, these evaluations are difficult to undertake; the overall dearth of gold standard reference sets is cited as a primary barrier (Ouzounis and Karp, 2002).

To address the challenge of annotating uncharacterized proteins reliably, we developed a benchmark annotation dataset of 30 proteins to serve as a gold standard. All proteins in the dataset were formerly uncharacterized proteins in *S. oneidensis* from the initial annotation of its proteome (2002). The 5 years that have elapsed have allowed for these new function characterizations. We chose this organism given our interest and expertise in its biology and potential scientific application. Our annotations are reliable in that they are based on high sequence similarity to experimentally characterized proteins. We used this dataset to evaluate predictions from sev-

¹Division of Biomedical and Health Informatics, Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, Washington.

²Seattle Children’s Hospital Research Institute, Seattle, Washington.

eral commonly utilized annotation databases. The accuracy benchmarks in this study provide some of the first insights into how well these databases predict function for uncharacterized proteins. Our benchmark dataset, our approach for reliably predicting function, and performance assessments of several annotation databases are described herein.

Materials and Methods

Benchmark annotation dataset

The annotations in the benchmark dataset are based on high sequence similarity to experimentally characterized proteins. This alleviates known sources of error such as incorrect database records (Koonin and Galperin, 2003), which can be difficult to detect. The intervening 5 years since the initial annotation of *S. oneidensis* proteins has allowed for new function characterizations to be published and made the creation of this dataset possible. These publications describe direct experimental function characterization for a protein, such as an enzyme assay. The following series of steps are then applied to assign function to uncharacterized proteins:

1. An experimentally characterized protein is compared to all proteins in an organism using BLAST or PSI-BLAST until convergence, with low-complexity filtering on. If the top-hit is an uncharacterized protein with an e-value $\leq e-04$, a threshold which indicates homology (Brenner et al., 1998), it is a candidate for function assignment.
2. The length of the experimentally characterized protein should be at least 90% of the length of the uncharacterized protein, or vice versa. The exception is if they both contain at least 90% of a commonly shared conserved domain. For example, SO_0887 from *S. oneidensis* and experimentally characterized protein NP_905579 have greater than 10% length difference, but both contain 99% of pfam04371, a 329 amino acid peptidyl-arginine deiminase domain.
3. If both criteria are met, then the function from the experimentally characterized function is assigned to the uncharacterized protein.

Annotation databases

We refer to both protein and pattern databases as “annotation” databases. The primary distinction between protein and pattern databases is that protein databases are based on pairwise sequence similarity determined by BLAST or PSI-BLAST searches, whereas pattern databases are based on similarity to multiple sequences. Two protein databases were evaluated in this study: UniProt (The UniProt Consortium, 2007), and SwissProt (Barrioch and Apweiler, 1996). Ten pattern databases were also selected and searched using their respective search tools: BLOCKS (Henikoff et al., 1999), Clusters of Orthologous Groups (COG) (Atwood, 2000), Protein Clusters (PRK) (O'Neill et al., 2007; Tatusov et al., 2000), The Conserved Domain Database (CDD) (Marchler-Bauer et al., 2002), InterPro (Apweiler et al., 2001), Pfam (Sonnhammer et al., 1997), PIRSFScan (Wu et al., 2003), SuperFamily (Gough et al., 2001), SMART (Schultz et al., 1998), and TIGRFAM (Haft et al., 2003).

Annotation database evaluation

The highest scoring result (top-hit) by e-value from each annotation database was compared to the benchmark anno-

tation. Each was scored as “Agree,” “Disagree,” or “Indeterminate.”

- An “Agree” was scored when the top-hit agreed with the benchmark (annotation database top-hit = benchmark).
- An “Indeterminate” was scored when the annotation database returned no hits or no function was specified in the top-hit (annotation database top-hit = “hypothetical” or no hits returned).
- A “Disagree” was scored when the annotation database top-hit was not equal to the benchmark and not Indeterminate (annotation database top-hit $\neq$ benchmark and annotation database $\neq$ Indeterminate).

These scores are summarized as “Accuracy” and “Conditional Accuracy” metrics for each annotation database. Accuracy is the proportion of correct predictions made on the whole dataset. It is defined as the number of “Agrees” divided by size of the dataset (30). Conditional Accuracy is the proportion of correct predictions made on a benchmark protein where a prediction was returned by the database, that is, it excludes “Indeterminate” predictions. It is defined as the number of Agrees divided by the sum of (“Agrees” + “Disagrees”).

Some annotation databases are composed of multiple “component” annotation databases, which can complicate evaluation. Pfam is available individually, and is also incorporated in CDD, InterPro, SMART, and TIGRFAM, for instance. Determining the top-hit for these “integrated” databases can be difficult, as some do not provide results sorted by e-value. Evaluating the component databases individually helps in this regard, and some interfaces enable this flexibility, such as InterPro. Evaluating all component databases and combinations thereof is onerous, however. InterPro alone has 13 component databases. Because Pfam is widely incorporated, we assessed the impact of Pfam on the databases which incorporate it, when the interface allowed. This was possible in TIGRFAM and InterPro. We evaluated TIGRFAM and InterPro in aggregate and with Pfam removed. We were able to evaluate CDD in aggregate and COG and PRK individually.

Heterogeneous function descriptions can also make comparing and scoring annotations difficult. Function synonyms were checked for using the BRENDA enzyme database (Schomburg et al., 2002), and we attempted to normalize descriptions using a Gene Ontology database (Ashburner et al., 2000). Still, it is challenging to determine agreement between structural and biochemical descriptions, or when comparing enzyme functions that are hierarchical. For this study, we maintained strict standards for scoring “Agrees” to remain true to the benchmark. Enzyme functions at different hierarchical levels or structural characterizations sharing only a common fold were scored as “Disagree.” Although these are not necessarily incorrect, specific function annotations, such as those in the gold standard, are more preferred by biological researchers than high-level enzyme family or structural descriptions.

Results and Discussion

Benchmark annotation dataset

All annotations were made to proteins that were originally uncharacterized in 2002. Our benchmark dataset is provided in Table 1. The database identifiers are *S. oneidensis* locus

TABLE 1. BENCHMARK ANNOTATION DATASET OF 30 PROTEINS FROM *S. ONEIDENSIS*

<i>Locus Id</i>	<i>PubMed Ids</i>	<i>Function</i>
SO_0342	14702315	prpF protein required for repair/synthesis of Fe-S center of AcnD
SO_0506	11029449, 782527, 12799002	3-octaprenyl-4-hydroxybenzoate Decarboxylase UbiD
SO_0887	10377098	Peptidylarginine deiminase
SO_1313	15901686, 16452451	Anhydro-N-acetylmuramic acid kinase
SO_1523	11488932	NAD kinase
SO_1597	12055309	omega-3 polyunsaturated fatty acid synthase PfaD subunit
SO_1608	15767583	7-cyano-7-deazaguanine reductase
SO_1789	12000770, 12000771	UDP-2,3-diacylglycerolamine hydrolase
SO_1851	17010378	Methyltransferase
SO_1963	10876237	Homogenetisate 1,2-dioxygenase
SO_2042	15355966, 16042411	Sulfite oxidase subunit YedY
SO_2043	15355966, 16042411	Sulfite oxidase subunit yedZ
SO_2593	10924516	NAD-specific glutamate dehydrogenase
SO_2603	17535911	Methionine- α -sulfoxide reductase
SO_2614	11011151	aminodeoxychorismate lyase
SO_2627	11931773, 12426582	ATP-dependent Clp protease adaptor protein ClpS
SO_3014	12100548	chromosome segregation and condensation protein B
SO_3015	12100548	chromosome segregation and condensation protein A
SO_3367	12730187	tRNA guanine-N(7)-methyltransferase
SO_3436	12756329	tRNA pseudouridine synthase TruD
SO_3542	16086247	Phosphoketolase
SO_3578	16740638	Multicopper polyphenol oxidase (laccase)
SO_3667	16376031	Heme iron utilization protein HugZ
SO_3668	16376031	Heme iron utilization protein HugX
SO_3957	12639950	3-deoxy-D-manno-octulosonate 8-phosphate phosphatase
SO_4227	10572301	S-adenosylmethionine-dependent methyltransferase MraW
SO_4398	10383414	D-tyrosyl-tRNA deacylase
SO_4413	9264543, 9477966	Kynureninase
SO_4677	10531340, 10952982	3-deoxy-D-manno-octulosonic-acid kinase
SO_4680	10648531	CDP-glycerol:poly(glycerophosphate) glycerophosphotransferase

identifiers defined by Genbank. The associated publications (PubMed Ids) provide the supporting evidence for the given function for each protein. Additional detailed information about each protein, such as BLAST e-values, protein lengths, and database identifiers, is in Supplementary Table 1.

Accuracy of annotation databases

Annotation databases were queried in October, 2007. The top six databases in our study have overall accuracy rates of at least 60%: TIGRFAM, CDD, SwissProt, InterPro, PRK, and

TABLE 2. SUMMARY OF ANNOTATION DATABASE PERFORMANCE

<i>Annotation database</i>	<i>Agree</i>	<i>Disagree</i>	<i>Indeter.</i>	<i>Accuracy</i>	<i>Cond_Acc</i>
TIGRFAM (7.0)	20	7	3	66.7%	74.1%
CDD (2.12)	19	5	6	63.3%	79.2%
SwissProt (54.4) (B)	19	2	9	63.3%	90.5%
InterPro (16.1)	18	9	3	60.0%	66.7%
PRK (1.0)	18	1	11	60.0%	94.7%
UniProt (37.4) (PB)	18	9	3	60.0%	66.7%
Pfam (22.0)	15	10	5	50.0%	60.0%
COG (1.0)	14	9	7	46.7%	60.9%
SMART (5.1)	14	11	5	46.7%	56.0%
InterPro (16.1) (-Pfam)	13	6	11	43.3%	68.4%
TIGRFAM (7.0) (-Pfam)	12	0	18	40.0%	100.0%
SuperFamily (1.69)	10	15	5	33.3%	40.0%
BLOCKS (14.3)	7	10	13	23.3%	41.2%
PIRSF	5	0	25	16.7%	100.0%

"InterPro(-Pfam)" and "TIGRFAM (-Pfam)" are results from InterPro and TIGRFAM without Pfam results. "(B)" and "(PB)" stand for BLAST and PSI-BLAST searches of the associated database, respectively. "Cond_Acc" is "Conditional Accuracy." Database versions, in parenthesis, are supplied as well.

UniProt (Table 2). Two of these six databases have conditional accuracy rates above 90%: SwissProt and PRK (Table 2). The annotations produced by these databases were very likely to agree with the benchmark. Overall, 27 out of 30 proteins were correctly predicted by at least one database. This level of accuracy can only be achieved by using top-hits from multiple databases. For instance, the top hits from TIGRFAM, Pfam, PRK, and COG can achieve this level of accuracy. This combination of databases is not found in any integrated annotation database such as CDD or InterPro.

Some annotation databases, such as Pfam and SuperFamily, search for very distant relationships between proteins. SuperFamily uses structure relationships that allows for function transfer at very low sequence similarity. Consequently, their annotations may not be specific enough to be scored as “Agree” according our strict evaluation criteria, but may still be biologically correct. For instance, SuperFamily predicts “Thiamine diphosphate binding fold” for SO_3542, a common structure in an enzyme family related to phosphoketolases, the benchmark function.

No annotation database could correctly predict functions for three proteins in the dataset (SO_2043, SO_2603, and SO_3578). Their database identifiers are highlighted in Supplementary Table 2. Some databases gave hints at the function but not at the level of specificity required in our evaluation. For example, SO_2043 is a “sulfite oxidase subunit YedZ” in our benchmark. Pfam, however, suggests a “Fer-ric reductase-like transmembrane component,” which appears to refer to a general family of proteins distantly related to the benchmark annotation.

Annotation database versions, the top-hit for each protein database, and evaluation scores are included in Supplementary Table 2.

Limitations

Our process for creating benchmark datasets is manual and not scalable. The result is a small benchmark dataset in a single organism. The accuracy assessments described here are therefore preliminary given the limited sample size and diversity of protein functions. *S. oneidensis* is also quite similar to *E. coli*, an organism where function characterizations are commonly performed. It may not be possible to create benchmark sets in other organisms that are less comparable to *E. coli*.

Our evaluation has some caveats. First, it is based on the assumption that functional evidence is available and published, which is not the case for the majority of uncharacterized proteins. Also, proteins can have more than one function, a possibility our process does not yet consider. Additionally, annotation databases can provide more than one prediction. We selected the top-hit by score, which is not always the correct choice. In our study we observed only a few cases where the correct answer was not the top hit. For example, when querying TIGRFAM with SO_3014, the second hit “TIGR:00281: segregation and condensation protein (3e-27),” rather than the first “DUF387: putative transcriptional regulators (4e-69),” agrees with the benchmark annotation.

Conclusions

Benchmark datasets of proteins, such as the one we have created here, allow for accuracy estimation of annotation

databases in regard to predicting protein function. In this study, we found that several commonly utilized annotation databases can produce highly accurate function predictions for uncharacterized proteins in *S. oneidensis*, according to our strict evaluation criteria. In addition, we observed that combining databases provides further annotation improvements. These accuracy benchmarks are among the first for assigning function to uncharacterized proteins using protein databases. They are also an indicator that protein databases are incorporating new experimental function evidence appropriately. In future work we plan on more definitively validating the accuracy of protein databases for predicting function through the use of larger benchmark datasets and extended evaluation criteria.

Acknowledgments

We thank Andrew Bauman, Caroline Dombrowski, Michael Galperin, Kira Makarova, Jared Roach, and Gerald van Belle for helpful discussions and numerous suggestions. We also appreciate support from: NIH NHGRI/NLM R01 HG02288, NIH T15 LM07442, NSF IIS-0513877, and NSF EF-0748739.

Author Disclosure Statement

The authors declare that no competing financial interests exist.

References

- Apweiler, R., Attwood, T., Barioch, A., Bateman, A., and Birney, E. (2001). The InterPro database, and integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* 29, 37–40.
- Attwood, T.K. (2000). The quest to deduce protein function from sequence: the role of pattern databases. *Int J Biochem Cell Biol* 32, 139–155.
- Barioch, A., and Apweiler, R. (1996) The Swiss-Prot protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* 24, 21–25.
- Bork, P. (2000). Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res* 10, 398–400.
- Brenner, S., Chothia, C., and Hubbard, T. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci USA* 95, 6073–6078.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25–29.
- The UniProt Consortium (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Res* 35, D193–D197.
- Devos, D., and Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends Genet* 17, 429–431.
- Galperin, M.Y., and Koonin, E.V. (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol*.
- Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 4, 903–919.
- Haft, D.H., Selengut, J.D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res* 31, 371–373.

- Heidelberg, J., Paulsen, I., Nelson, K., Gaidos, E., Venter, J.C., Nealson, K., et al. (2002). Genome sequence of the dissimilarity metal ion-reducing bacterium *Shewanella oneidensis*. *Nat Biotechnol* 20, 1118–1123.
- Henikoff, S., Henikoff, J., and Pietrokovski, S. (1999). Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15, 471–479.
- Karp, P., Paley, S., and Zhu, J. (2001). Database verification of studies of SWISS-PROT and GenBank. *Bioinformatics* 17, 526–532.
- Kolker, E., Piccone, A., Galperin, M., Smith, R., Giometti, C., Nealson, K., et al. (2005). Global profiling of *Shewanella oneidensis* MR-1: expression of hypothetical genes and improved functional annotations. *Proc Natl Acad Sci USA* 102, 2099–2104.
- Koonin, E., and Galperin, M. (2003). *Sequence–Evolution–Function*. (Kluwer Academic Publishers, Boston).
- Marchler-Bauer, A., Panchenko, A., Shoemaker, B., Thiessen, P., Geer, L., and Bryant, S. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* 30, 281–283.
- O'Neill, K., Klimke, W., and Tatusova, T. (2007). *Protein Clusters: A Collection of Proteins Grouped by Sequence Similarity and Function*. (NCBI, Bethesda, MD).
- Ouzounis, C., and Karp, P. (2002). The past, present and future of genome-wide re-annotation. *Genome Biol* 3, 2001.2001–2001.2006.
- Schomburg, I., Chang, A., and Schomburg, D. (2002). BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* 30, 47–49.
- Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA* 95, 5857–5864.
- Sonnhammer, E., Eddy, S., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28, 405–420.
- Tatusov, R., Galperin, M., Natale, D., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28, 33–36.
- Thomas, G. (1999). Completing the *E. coli* proteome: a database of gene products characterized since the completion of the genome sequence. *Bioinformatics* 15, 860–861.
- Valencia, A. (2005). Automatic annotation of protein function. *Curr Opin Struct Biol* 15, 267–274.
- Wu, C., Huang, H., Yeh, L., and Barker, W. (2003). Protein family classification and functional annotation. *Comp Biol Chem* 27, 37–47.

Address reprint requests to:

Eugene Kolker
Seattle Children's Hospital Research Institute
1900 9th Avenue
Seattle, WA, 98101-1304

E-mail: eugene.kolker@seattlechildrens.org