Published online 20 November 2010 Nucleic Acids Research, 2011, Vol. 39, Database issue D11–D14
doi:10.1093/nar/gkq1168

COMBREX: a project to accelerate the functional annotation of prokaryotic genomes

Richard J. Roberts[1,2,*], Yi-Chien Chang[1], Zhenjun Hu[1], John N. Rachlin[3], Brian P. Anton[1,2], Revonda M. Pokrzywa[4], Han-Pil Choi[5], Lina L. Faller[1], Jyotsna Guleria[4], Genevieve Housman[4], Niels Klitgord[1], Varun Mazumdar[1], Mark G. McGettrick[3], Lais Osmani[4], Rajeswari Swaminathan[4], Kevin R. Tao[4], Stan Letovsky[1], Dennis Vitkup[6,7], Daniel Segrè[1,4,8], Steven L. Salzberg[9], Charles Delisi[1,4], Martin Steffen[1,4,5,*] and Simon Kasif[1,4,10,*]

[1]Bioinformatics Program, Boston University, Boston, MA 02215, [2]New England Biolabs, 240 County Road, Ipswich, MA 01938, [3]Diatom Software LLC, 260 Winter St., Holliston, MA 01746, [4]Department of Biomedical Engineering, Boston University, Boston, MA 02215, [5]Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston University, Boston, MA 021184, [6]Department of Biomedical Informatics, Columbia University, New York, NY 10032, [7]Center for Computational Biology and Bioinformatics, Columbia University, New York, 10032, [8]Department of Biology, Boston University, Boston MA 02215, [9]Department of Computer Science, University of Maryland, College Park, MD, 20742 and [10]Children's Hospital Informatics Program, Harvard-MIT Division of Health Sciences and Technology (CHIP@HST), Cambridge, MA 02139, USA

Received October 15, 2010; Revised October 31, 2010; Accepted November 1, 2010

ABSTRACT

COMBREX (http://combrex.bu.edu) is a project to increase the speed of the functional annotation of new bacterial and archaeal genomes. It consists of a database of functional predictions produced by computational biologists and a mechanism for experimental biochemists to bid for the validation of those predictions. Small grants are available to support successful bids.

INTRODUCTION

In the last 15 years, since the determination of the complete sequence of the *Haemophilus influenzae* strain Rd genome (1), there has been a rapid increase in the number of prokaryotic genomes that are being sequenced each year. With the cost of DNA sequencing continuing to drop, this has led to an explosion in the number of genes that are predicted computationally, but for which no solid functional annotation can be provided (2). This is illustrated in Table 1, which shows that in a selection of genomes, at best, maybe 70% of the genes have either known, experimentally validated functions or can be assigned function computationally on the basis of sequence similarity, but often with varying or unknown degrees of confidence. With each new genome typically containing anywhere from 500 to 1000 new genes of unknown function, we face the daunting challenge of determining those functions so that the annotation of new genome sequences can be carried out computationally with just a few key functions being tested experimentally. This means that our ability to predict function computationally will need to be quite accurate and must include all genes.

Currently, the quality of computational predictions of function is far from perfect. Indeed, for many of the genes in GenBank the present annotations are either incorrect or so general as to be of little value to the user (3–6). The reason for this is that by far the most common way of making predictions is by checking each newly predicted gene for its similarity to genes annotated in the INSDC (International Nucleotide Sequence Database Consortium) databases (7–9). When a new gene shares high sequence similarity to an annotated gene then it is assigned the same function as that presumed known gene. If they are identical or nearly so, then this method is quite

*To whom correspondence should be addressed. Tel: +1 617 358 1845; Fax: +1 617 353 6766; Email: kasif@bu.edu
Correspondence may also be addressed to Richard J. Roberts. Tel: +1 978 380 7405; Fax: +1 978 380 7406; Email: roberts@neb.com
Correspondence may also be addressed to Martin Steffen. Tel: +1 617 826-9486/7935; Fax: +1 617 414 7073; Email: steffen@bu.edu

Table 1. Distribution of annotated genes in selected genomes

Organism	Publication date	Unknown function[a] (%)	Known or predicted function	Total
Haemophilus influenzae Rd	1995	429 (26)	1228	1657
Methanocaldococcus jannaschii	1996	616 (35)	1155	1771
Helicobacter pylori 26695	1997	552 (35)	1021	1573
Escherichia coli MG1655	1997	550 (13)	3594	4144
Pseudomonas aeruginosa PA7	2010	2653 (42)	3633	6286

[a]Based on genes annotated as unknown or conserved hypothetical in the RefSeq files. These represent lower estimates as the accuracy of the predicted functions is unknown.

reliable. However, when the degree of sequence similarity is poor, or perhaps even when it is reasonably high with only a few key amino acids difference, this can lead to problems, because one can be less sure that the new gene really is an ortholog of the known gene. Perhaps, the new gene encodes a protein with a function that is similar to the known one, but different in some subtle way such that its substrate preference has changed. Unless the sequence differences can be interpreted properly so that the new protein's function is not declared to be identical to that of the old gene product, then a mis-annotation ensues and will likely be propagated (3–6).

A number of now classic examples of such mis-annotations have been noted in the literature and only when biochemical experiments were carried out, could the annotation be corrected. One classic example was the family of genes labeled hemK. These genes had been annotated as either a protoporphyrinogen oxidase or a DNA methyltransferase. It later turned out that the hemK gene in *Escherichia coli* actually encoded a protein methyltransferase, a finding of some considerable interest because the hemK gene is widely conserved from humans all the way to bacteria (10,11), although further testing on remote homolog's would still be appropriate. This emphasizes the need for biochemical characterization of gene products whenever possible, but certainly when the sequence distance to a known gene product is insufficiently high to be certain of the assignment. The degree of caution necessary often varies, depending on the level of conservation or its location. In some cases, one or a few amino acid changes in a region of a protein responsible for substrate recognition can completely alter its function. Unfortunately, we often do not know in which regions of a protein we should look for such changes and the computer blindly labels the new gene incorrectly.

THE COMBREX DATABASE

Once a particular gene's function is characterized biochemically, then that function can be propagated with some degree of certainty to the likely orthologous genes in other organisms, although remote homolog's will require experimental validation. It is precisely this combination of computational prediction and biochemical validation of function that the COMBREX (COMputational BRidges to EXperiments) project, recently funded by NIGMS, is all about. At the heart of COMBREX is a database (http://combrex.bu.edu) of computational predictions of gene function. This database, which is currently under construction, contains all of the annotated genes present in the bacterial and archaeal genomes section of the NCBI's Protein RefSeq Database (12,13). We take advantage of the Clusters Database (14) in which these genes have been sorted into families of similar sequences to organize the genes for predictive purposes. However, in addition to the annotations contained in the Clusters Database, which are themselves of course predictions for the most part, the COMBREX database also contains functional predictions made by other groups of computational biologists. A major goal is to provide reliability estimates for those predictions. We already have some ongoing collaborations and it is our hope that we can involve many others in the bioinformatics community who are making gene predictions and who are willing to share them through the medium of the COMBREX database where they will be publicly and freely available. The reason for doing so is that we are recruiting biochemists to test those predictions experimentally.

The involvement of the biochemists is the other leg of this project. We are inviting them to browse the COMBREX database and identify predictions that match their own laboratory's biochemical expertise. They can then make a bid to test the function of any high-value predictions lying within their area of expertise. For instance, if an unknown protein is predicted to hydrolyze carbohydrates, then we would look for a laboratory that is expert in such hydrolases that would have a large range of carbohydrate substrates and suitable assays to detect hydrolytic activity. The idea is that they would make some of the protein from the gene in question and run it through their battery of assays. For US laboratories, we are able to offer small grants through COMBREX that are typically in the range of $5000 to $10 000 to support his work, which might be carried out by an individual such as a graduate student, a supervised rotation student or anyone with sufficient proven expertise to complete the task. There should be opportunities here for teaching colleges to participate as well as for some of the top universities to incorporate this approach into their normal curriculum. The successful bid would guarantee six months of sole access to that gene product through COMBREX and at the end of that time, a report would be written and would be available on the COMBREX website describing the experiments that were performed and the results, either a positive validation or a failure to detect the predicted activity. We would encourage the laboratory to publish these results in the peer reviewed literature, and they would be featured on the COMBREX web site. They would also be propagated to the appropriate databases of the INSDC.

A GOLD STANDARD DATABASE OF PROTEINS

One important problem that has been identified during the initial stage of the COMBREX project, is that the

successful propagation of annotations from one gene to another, depends critically on knowing when a protein of known sequence has an experimentally demonstrated biochemical function, because this becomes a standard against which similar proteins can have their functions predicted. It turns out that this information is not always easily deciphered. For many proteins the biochemical characterization was carried out on a purified protein from a bacterial strain many years before the gene for that protein was cloned and sequenced. In this case the cloned gene may well come from a strain that is different from the one in which the original characterization was done and there may be subtle, but important, differences in sequence. Sometimes accurate strain information can be found in the major databases, because both the sequence and the characterization were described in a single publication. But often the necessary pedigree information is not easily traced. Recognizing that this is a major problem, not just for COMBREX, but for all groups trying to propagate annotation, we have now undertaken a project in collaboration with the RefSeq Database at NCBI (12) and the UniProt Knowledge Database at EBI (14) to identify a 'Gold Standard Set of Proteins' for which the function is known and the exact sequence of the gene/protein on which the functional tests were done is known. This gold standard project is currently being managed by COMBREX and a pipeline has been set up whereby candidate genes/proteins for gold standard status are identified and distributed to individual annotators who will check in detail that they do indeed, have gold standard properties. The final gold standard database will be maintained and distributed from both NCBI and UniProt.

This gold standard project is another community-based project in which individuals from around the world can help either by identifying potential gold standard genes or by helping in the manual curation. In a similar fashion, we hope that COMBREX will also become a community-based project around the world so that experimental characterization of function for COMBREX predictions can be carried out in laboratories in many countries. It is worth noting that this approach can be used as a teaching tool for students learning how to do hands-on biochemistry, while at the same time making a valuable contribution to biology. There are many small laboratories with appropriate biochemical expertise to test certain specific predictions within the area of their expertise and for whom a small amount of money can make a large difference. Already a number of collaborators are helping with this project. We hope that by engaging the larger community of computational biologists and biochemists, we can build momentum in a project that has the potential to greatly increase the accuracy of genome annotation. It should enable bioinformaticians to make better predictions by providing a set of reference points in the gold standard set of proteins and facilitating the biochemical testing of their predictions. The feedback from this validation process should then impact their ability to make more reliable predictions. The involvement of the experimental community will highlight the importance of their discipline while also providing some funds that might help train the future generation. A successful outcome to the project should mean that the number of genes in need of experimental verification will diminish with time, because the computer predictions can be assessed by rigorously documenting their distance to a gold standard protein.

The size of the functional annotation problem is enormous and it is essential that it be tackled if we are to keep pace with the genome and metagenome projects that are currently underway. It is one where collaboration is demanded and competition would merely serve to waste money. While the original idea was proposed in 2004 (15), before the current 'Wiki' approaches were popular, COMBREX can be seen as essentially a similar community-based approach to the problem of functional annotation of prokaryotic genomes. We anticipate that if this project is successful, the model will be expanded from its initial focus on bacteria and archaea to cover genomes in all kingdoms of life. It will also provide some more of the raw data on function that may one day permit systems biology to really view an organism as a system. Furthermore, it will also have an impact on many of the databases in this issue. For a relatively small investment and a massively parallel human effort, it should be possible to achieve high throughput.

FUNDING

GO grant from National Institute of General Medical Sciences (NIGMS) (1RC2GM092602-01 to COMBREX). The open access publication charge for this paper has been waived by Oxford University Press - *NAR* Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.

Conflict of interest statement. None declared.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
2. Galperin,M.Y. and Koonin,E.V. (2010) From complete genome sequence to 'complete' understanding? *Trends Biotechnol*., **28**, 398–406.
3. Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet*., **15**, 132–133.
4. Schnoes,A.M., Brown,S.D., Dodevski,I. and Babbitt,P.C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comp. Biol*., **5**, e1000605.
5. Green,M.L. and Karp,P.A. (2005) Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res*., **33**, 4035–4039.
6. Hsiao,T.L., Revelles,O., Chen,L., Sauer,U. and Vitkup,D. (2010) Automatic policing of biochemical annotations using genomic correlations. *Nat. Chem. Biol*., **6**, 34–40.
7. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) Genbank. *Nucleic Acids Res*., **38**, D46–D51.
8. Kaminuma,E., Mashima,J., Kodama,Y., Gojobori,T., Ogasawara,O., Okubo,K., Takagi,T. and Nakamura,Y. (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res*., **38**, D33–D38.

9. Leinonen,R., Akhtar,R., Birney,E., Bonfield,J., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N., Gobson,R. *et al*. (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res*., **38**, D39–D45.
10. Heurgue-Hamard,V., Champ,S., Engstrom,A., Ehrenberg,M. and Buckingham,R.H. (2002) The hemK gene in *Escherichia coli* encodes the N(5)-glutamine methyltransferase that modifies peptide release factors. *EMBO J*., **21**, 769–778.
11. Nakahigashi,K., Kubo,N., Narita,S., Shimaoka,T., Goto,S., Oshima,T., Mori,H., Maeda,M., Wada,C. and Inokuchi,H. (2002) HemK, a class of protein methyl transferase with similarity to DNA methyl transferases, methylates polypeptide chain release factors, and hemK knockout induces defects in translational termination. *Proc. Natl Acad. Sci. USA*, **99**, 1473–1478.
12. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–65.
13. Klimke,W., Agarwala,R., Badretdin,A., Chetvernin,S., Ciufo,S., Fedorov,B., Kiryutin,B., O'Neill,K., Resch,W., Resenchuk,S. *et al*. (2009) The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res*., **37**, D216–223.
14. The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*., **38**, D142–D148.
15. Roberts,R.J. (2004) Identifying protein function–a call for community action. *PLoS Biol*., **2**, E42.