COMBREX: A Project to Accelerate the Functional Annotation of Prokaryotic Genomes

https://hdl.handle.net/2144/3036
Boston University
COMBREX: a project to accelerate the functional annotation of prokaryotic genomes

Richard J. Roberts1,2,*, Yi-Chien Chang1, Zhenjun Hu1, John N. Rachlin3, Brian P. Anton1,2, Revonda M. Pokrzywa4, Han-Pil Choi5, Lina L. Faller1, Jyotsna Guleria4, Genevieve Housman4, Niels Klitgord1, Varun Mazumdar1, Mark G. McGettrick3, Lais Osman4, Rajeswari Swaminathan4, Kevin R. Tao4, Stan Letovsky1, Dennis Vitkup6,7, Daniel Segrè1,4,8, Steven L. Salzberg9, Charles Delisi1,4, Martin Steffen1,4,5,* and Simon Kasif1,4,10,*

1Bioinformatics Program, Boston University, Boston, MA 02215, 2New England Biolabs, 240 County Road, Ipswich, MA 01938, 3Diatom Software LLC, 260 Winter St., Holliston, MA 01746, 4Department of Biomedical Engineering, Boston University, Boston, MA 02215, 5Department of Pathology and Laboratory Medicine, Boston University School of Medicine, Boston University, Boston, MA 02118, 6Department of Biomedical Informatics, Columbia University, New York, NY 10032, 7Center for Computational Biology and Bioinformatics, Columbia University, New York, 10032, 8Department of Biology, Boston University, Boston MA 02215, 9Department of Computer Science, University of Maryland, College Park, MD, 20742 and 10Children’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...
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 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 new classical 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
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


