COMBREX: COMputational BRidge to EXperiments

Richard J. Roberts[1]
New England Biolabs, 240 County Road, Ipswich, MA 01938, U.S.A.

Abstract

COMBREX (computational bridges to experimentation) is a project to engage the biological community in providing better functional annotation of genomes. In essence, the project involves the generation by computational biologists of a database of predicted functions for genes in bacterial genomes. Those genes for which no functional assignments have been proven experimentally are then open for bids by biochemists to test the predicted functions. High-priority genes are those for which no previous functional assignment has been made as well as those where uncharacterized examples are present in many genomes. A pilot project is running that focuses on bacterial and archaeal genomes.

Introduction

Although the first bacterial genome was sequenced in 1995, we are still at the very beginning of the era of genomics. It is now routine to obtain complete and accurate sequences for the genomes of small organisms such as bacteria, archaea and even the lower eukaryotes such as *Saccharomyces cerevisiae*. These complete DNA sequences are of great importance because they contain the genetic blueprint for the organism. Importantly, not only do they tell us what genes are present in the organism, but also they tell us which are not. However, to properly interpret this information, not only do we need to identify exactly which parts of the DNA sequence encode the genes, but we also need to know what the genes are responsible for. What proteins do they encode and what are the biochemical functions of those proteins? Only when we have this information will we have the fundamental building blocks upon which to base our understanding of how the organism works. That basic understanding will also enable us to manipulate those organisms in a sensible fashion. For example, a good understanding of the human genome will be the basis for personalized medicine. From plant genomes, we can probably design new pathways that will lead to better foods and perhaps even energy production. From the myriad of organisms in the oceans, we can expect to find many novel genes that may lead to new medicines and also may impact energy. Increasingly we are realizing that all the large organisms we typically think of live within a milieu of bacteria and archaea that affect, in a very significant fashion, all of life on earth. Their genomes offer particular promise in areas such as health, food and energy and have the great advantage that being small one might hope to gain a good understanding of how these organisms work.

The problem

Thanks to the technological advances that began with the development of DNA sequencing methods in the mid-1970s, DNA sequencing is becoming faster and cheaper, but, disturbingly, the accumulating sequences are greatly exceeding our ability to interpret the final products. This is well illustrated by looking at any recent bacterial genome sequence and seeing that the number of predicted genes that have unknown, and in most cases unpredictable, function is often 20–40 % of the total. This is about the same number that we found in the mid-1990s when the first bacterial genomes were sequenced. Thus, whereas sequencing techniques have vastly improved, the interpretation of those sequences continues to make progress at a snail's pace. We have spent a huge amount of money developing better sequencing methods and drastically reducing the costs of sequencing, but there has been no such investment in understanding the functions of those sequences. Increasingly, we have been content to rely on computers to look for similarity between old genes and new genes and simply assumed that if the two genes were similar, then they must have the same function. This has led to many known cases where an initial assignment was incorrect and now hundreds of assignments are similarly incorrect because the original problem was propagated. Even more disturbingly, for those genes with no known function and no easy way to predict a function, they have simply been ignored for the most part. There has been no systematic effort either to predict or to determine experimentally the function of those genes. One might ask why this is the case? There are several reasons. First, *de novo* prediction or testing for function is not a simple matter. Predictions can be made in many ways such as locating small characteristic elements within a sequence that might suggest interaction with a possible substrate. Sometimes the neighbourhood of the gene can be helpful,

Key words: biological community, computational bridge to experiments (COMBREX), DNA sequence, functional assignment, genomics.

Abbreviations used: COMBREX, COMputational BRidge to EXperiments; RFP, request for proposals.

[1] email roberts@neb.com

as when a gene lies in a well-known operon in which the functions of neighbouring genes are known. Occasionally the structure of a protein can give some clues to function. But in all these cases the biochemical challenge remains that often a good deal of experimentation is required to test and hopefully validate the predicted function. More significantly in the current research climate, there are no really good high-throughput methods that allow the function of large numbers of genes to be tested experimentally. The result has been that we still rely on the serendipitous discovery of new functions for genes instead of adopting a dedicated and systematic approach to fill in the void.

The solution

In 2004, I suggested a solution to this dilemma in which high throughput might be achieved by parallelization of low-throughput traditional biochemical approaches [1]. That is, a large number of individual biochemical laboratories might be corralled to test specific computational predictions that lie squarely in their area of expertise. I proposed that a database of predicted functions could be assembled by the computational biologists and that expert biochemists could then browse those predictions and find ones that were easily testable in their laboratories. They could then be awarded a small grant, perhaps $5–10 000, to support a student who, under the expert tutelage of the laboratory, could quickly test the prediction. This idea was sufficiently appealing to NIH (National Institutes of Health) that I was encouraged to organize a workshop in Washington to explore it further. My colleague, Simon Kasif, a computational biologist from Boston University, and I organized such a workshop in 2004 and a report was issued by the American Academy of Microbiology. Many grant administrators came to this meeting, expressed interest in the idea, but told me that they were ill-equipped to administer small grants and so the idea was shelved. However, in 2009 an RFP (request for proposals) was issued by NIH, under their Grand Challenges Initiative, that asked for novel approaches to the functional annotation of genomes. At this point, Simon Kasif and Martin Steffen, from Boston University, and I decided to pursue my original idea and respond to the RFP. We were fortunate to get it funded and the result is COMBREX (COMputational BRidge to EXperiments), a project that aims to provide functional annotation of bacterial and archaeal genomes through a grand collaboration of biochemists and bioinformaticians.

Under the leadership of Simon Kasif, the computational challenge of building the database of predictions is already well under way. A number of expert bioinformaticians have already joined the project and more are expected in the future. At the same time, Martin Steffen and I are organizing the biochemical testing of these predictions. We already have a few collaborations under way and we anticipate that there will be many more in the future. The COMBREX website (http://combrex.bu.edu) is now open to the scientific community and we are receiving 'bids' from biochemists who have the expertise to test predictions. Under the current funding model, a biochemist in the U.S.A. wishing to test a prediction would receive a small grant to pay for the incremental cost of bringing a student, perhaps a rotation student, or even an undergraduate, into their laboratory to perform the necessary biochemistry. The proposal that the biochemist would submit would be fairly brief and would outline the experimental approach to be pursued to test the prediction. At that time, the gene or genes chosen would be off limits for 6 months and no other bids would be allowed until the first laboratory reported its results. Those results will be posted on the COMBREX website, irrespective of whether they are positive or negative, and we would also encourage publication of the results in the scientific literature. Indeed, we are contemplating initiating a journal specifically dedicated to such efforts.

It is important to note that because this initiative is currently being funded with U.S. funds, only U.S. laboratories have the opportunity of obtaining funding through this mechanism. However, the possibility of picking a gene to test will be open to everybody, including laboratories in Europe and the rest of the world. We hope that many U.K. scientists will join us! We are encouraging other funding agencies both within and outside of the U.S.A. to also make funds available to support this project and we have found some interest from both the Wellcome Trust in the U.K. and the Howard Hughes Medical Institute in the U.S.A. As time goes on and if the model proves effective, we anticipate that a number of non-U.S. government funding agencies might also be prepared to throw small amounts of money into the ring to support this effort.

At the present stage, the effort is limited to genes in bacteria and archaea mainly because these are more straightforward to identify. Functional predictions are often easier to make and certainly producing the proteins encoded by these genes in fully functional form and hence available for biochemical testing is generally quite facile. However, if the project is as successful as we imagine, then it could easily be extended to eukaryotic organisms. A more complete description of COMBREX is available [2].

It is obvious from the previous description that what we are doing under the auspices of this project is not just providing experimental determination of gene function, but we are also building a community of bioinformaticians and biochemists eager to collaborate on this important project, which will impact all aspects of biology. Importantly, as more functions are elucidated, so our ability to make better predictions will increase. There will be a synergy between the two fields that should benefit everybody. I am reminded of my own early days in biology when collaboration was the norm. Because biology was still a very small field at the time, everyone was quite keen to share results, reagents and ideas so that progress could be made quickly. It is our hope that this COMBREX project can revive some of this collaborative spirit which was so successful in driving the early developments in molecular biology.

Acknowledgements

I am grateful to Simon Kasif and Martin Steffen as well as all the members of the COMBREX team (listed as authors in [2]) for their enthusiastic help with this project.

Funding

This project was supported by a GO grant from National Institute of General Medical Sciences [grant number 1RC2GM092602–01].

References

1 Roberts, R.J. (2004) Identifying protein function: a call for community action. PLoS Biol. **2**, 293–294

2 Roberts, R.J., Chang, Y.-C., Hu, Z., Rachlin, J.N., Anton, B.P., Pokrzywa, R.M., Choi, H.-P., Faller, L.L., Guleria, J., Housman, G. et al. (2011) COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. Nucleic Acids Res. **39**, D11–D14

Received 4 December 2010
doi:10.1042/BST0390581