medigraphic Artemisa en línea

Revista Latinoamericana de MICROBIOLOGÍA
ALAM

Vol. 47, No. 1
January - April. 2006
pp. 44-45

COMENTARIO

The post-genomic era – What comes next?

Isabel María López-Lara,* Christian Sohlenkamp,* Otto Geiger*

Over the last decade, the genomes of several hundreds of organisms have been sequenced (http://www.ncbi.nlm.nih.gov/Genomes/) providing a tremendous amount of data that need to be interpreted and decorated with functions. However, in any newly sequenced bacterial genome, as many as 30-40% of the genes do not have an assigned function.[1] This figure is even higher for archaeal and eukaryotic genomes and for the relative large genomes of bacteria with complex life styles, such as rhizobia or *Streptomyces*. Remarkably, species- or genus-specific genes comprise a relatively small fraction of the uncharacterized genes. The majority of such "hypothetical" genes without any assigned function has a wider phylogenetic distribution and therefore are usually referred to as "conserved hypothetical". Presently available wholistic approaches such as micro- and macroarrays, protein-protein interaction analyses, etc., can provide important information about regulation and interactions and provide valuable clues for possible functions. However, the majority of genes code for enzymes and presently available whole-scale approaches, such as metabolomics, will reveal new enzymatic functions in a few cases only.

A significant proportion of the remaining 60-70% of genes in genomes, for which functional annotations have been made, are often imprecisely described or assigned with vague functions (i.e. described simply as "putative dehydrogenase"), and some are likely to be annotated wrong. The functional annotations are in most cases derived by inference rather than by experiment, through the observation of some level of sequence identity to a characterized gene product from another organism. It is Genomics itself which provides that rare opportunity in science where the boundaries of current knowledge can be clearly defined.

Much of the presently available annotation information is provided by computer programs that predict the functions of newly sequenced genes on the basis of their similarity to genes (or gene products) of known (or predicted) function. There are at least two problems associated with the method: 1) the small size of the core foundational set of genes with experimentally established functions and 2) there are difficulties to define which level of identity, similarity, or E-value is required to establish that two proteins have the same function or otherwise related functions. In extreme cases, dissimilar proteins might catalyze the same reaction (Fig. 7 in reference 5) whereas proteins with 98% identity might be functionally different.[7] Therefore, it is problematic to establish such threshold values because they probably change with each functional group of proteins. Similarities suggest a certain function but they don't show it and in order to obtain better predictions and annotations a larger data base with many more experimentally verified functional assignments is needed. In this sense, a recent report from the American Academy of Microbiology calls for an annotation initiative that is strongly backed by experimental evidence.[6]

It is also worrisome that, as many as 5-10%[6] of predicted gene functions may be incorrect. The source of such mistakes might have different reasons, but in fact such mistakes are rapidly propagated when genomes are annotated. Examples for such misannotations that currently exist in data bases are the *N*-acyltransferase OlsB and the acyl carrier protein phosphodiesterase AcpH.

The function of the *N*-acyltransferase OlsB[2] was discovered after the original annotation of the encoding gene. The original annotation for a protein of the cluster of orthologous group of proteins COG3176 was for PhyA which has a haemolysin function. However, even though the similarity between OlsB and PhyA is low, OlsB and its homologues are still commonly annotated in genomes as conserved hypothetical protein or as putative haemolysin. But the function of OlsB is not that of a haemolysin but that of an *N*-acyltransferase, catalyzing the first biosynthesis step for ornithine-containing lipids[2] which are widespread in bacteria.[2,4]

The enzymatic activity of acyl carrier protein phosphodiesterase was initially discovered in the sixties of the last century and cleaves a 4'-phosphopantetheine group from the constitutive acyl carrier protein AcpP. During the nineties, acyl carrier protein phosphodiesterase from *E. coli* was partially purified and an N-terminal sequence was obtained from this fraction. Based on this sequence, the function of acyl carrier protein phosphodiesterase AcpD was assigned to the corresponding ORF in the *E. coli* ge-

* Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México.

Rev Latinoam Microbiol 2006; 48 (1): 44-45

nome and this wrong assignment was propagated to other genome annotations. The ORF annotated as acyl carrier protein phosphodiesterase AcpD is an azoreductase and does not code for an acyl carrier protein phosphodiesterase. Recently, the structural gene for acyl carrier protein phosphodiesterase has been described and its product is now termed AcpH.[8]

Clearly one big task in this post-genomic era consists in much curation work in which genome annotations are continuously updated. Another major task will be the search for functions for the numerous COGs. One clear lead is provided by the fact that, for numerous experimentally characterized enzymes, there is still no available sequence information.[3] For as many as 1,400 known enzymes no corresponding genes have been identified in the sequence databases, and as many as for 36% of the EC numbers no protein sequence for that enzyme activity is known. The "conserved hypothetical" genes are the pool where biologist can fish for these "homeless" activities. The SRI International group has begun the project called the Enzyme Genomics Initiative, the goal of which is to find the genes associated with known enzymatic functions (http://bioinformatics.ai.sri.com/enzyme-genomics/).

The assignment of functions to ORFans as well as finding the sequence of orphan enzymes are big challenges ahead and a close collaboration is needed between bioinformaticians and experimentalists to achieve more meaningful annotations and more complete coverage. It is evident that the true value of the genome sequence information will only be realized after at least one meaningful function has been assigned to all of the encoded proteins.

ACKNOWLEDGEMENTS

Work in our laboratories was supported by grants from DGAPA/UNAM (IN200806), CONACyT-Mexico (42578/A-1 and 46020-N), and the Howard Hughes Medical Institute (HHMI 55003675).

REFERENCES

1. Bork, P. 2000 Powers and pitfalls in sequence analysis: the 70% hurdle. Genome Res. 10:398-400.
2. Gao, J.-L., B. Weissenmayer, A. Taylor, J. Thomas-Oates, I.M. Lopez-Lara & O. Geiger. 2004. Identification of a gene required for the formation of lyso-ornithine lipid, an intermediate in the biosynthesis of ornithine-containing lipids. Mol. Microbiol. 53:1757-1770.
3. Karp, P.D. 2004. Call for an enzyme genomic initiative. Genome Biol. 5:401.
4. López-Lara, I.M., C. Sohlenkamp & O. Geiger. 2003. Membrane lipids in plant-associated bacteria: their biosyntheses and possible functions. Mol. Plant-Microbe Interact. 16: 567-579.
5. Martínez-Morales, F., M. Schobert, I.M. López-Lara & O. Geiger. 2003. Pathways for phosphatidylcholine biosynthesis in bacteria. Microbiology 149:3461-3471.
6. Roberts, R.J., P. Karp, S. Kasif, S. Linn & M.R. Buckley. 2004. A report from the American Academy of Microbiology: An experimental approach to genome annotation. American Society for Microbiology, Washington, D.C.
7. Seffernick, J.L., M.L. de Souza, M.J. Sadowsky & L.P. Wackett. 2001. Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. J. Bacteriol. 183:2405-2410.
8. Thomas, J. & J.E. Cronan. 2005. The enigmatic acyl carrier protein phosphodiesterase of *Escherichia coli*. J. Biol. Chem. 280:34675-34683.

Correspondence to:

Dr. Otto Geiger
Centro de Ciencias Genómicas
Universidad Nacional Autónoma de México
otto@ccg.unam.mx

edigraphic.com