Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs)

ABSTRACT

Special-purpose databases organized on the basis of phylogenetic analysis and carefully curated with respect to known and predicted protein functions provide for a significant improvement in genome annotation. A differential genome display approach helps in a systematic investigation of common and distinct features of gene repertoires and in some cases reveals unexpected connections that may be indicative of functional similarities between phylogenetically distant organisms and of lateral gene exchange.

INTRODUCTION

Background Functional annotation of genomes is a critical aspect of the genomics enterprise. Without reliable assignment of gene function at the appropriate level of specificity, new genome sequences are plainly useless. The primary methodology used for genome annotation is the sequence database search, the results of which allow transfer of functional information from experimentally characterized genes (proteins) to their uncharacterized homologs in newly sequenced genomes. However, general-purpose, archival sequence databases are not particularly suited for the purpose of genome annotation. The quality of the annotation of a new genome produced using a particular database critically depends on the reliability and completeness of the annotations in the database itself. As far as annotation is concerned, the purpose of primary sequence databases is to faithfully preserve the description attached to each sequence by its submitter. In their capacity as sequence archives, such databases include no detailed documentation in support of the functional annotations. Furthermore, primary sequence databases are not explicitly structured by either evolutionary or functional criteria. These features, which are inevitable in archival databases, seriously impede their utility as resources for genome annotation, particularly when an automated or semi-automated approach is attempted. At its worst, this situation results in a notorious vicious circle of error amplification - an inadequately annotated database is used to produce an error-ridden and incomplete annotation of a new genome, which in turn makes the database even less useful. One way out of this 'Catch-22' situation is to use a different type of database for genome annotation, namely databases in which sequence information is organized by structural, functional or phylogenetic criteria, or a combination thereof. For example, the KEGG and WIT databases are primarily function-oriented and organize protein sequences from completely and partially sequenced genomes according to their known or predicted roles in biochemical pathways, although WIT also provides a phylogenetic classification. In contrast, the SMART database is organized on a structural principle and provides a searchable collection of common protein domains. All these databases share a fundamental common feature - they encapsulate carefully verified knowledge on protein structure, function and/or evolutionary relationships, and therefore, at least in principle, provide for a more robust mode of genome annotation than general-purpose databases and may serve as a stronger foundation for partially automated approaches to genome analysis. The database of Clusters of Orthologous Groups of proteins (COGs) is a phylogenetic classification of proteins encoded in completely sequenced genomes. An attempt has been made to organize these proteins into groups of orthologs, direct evolutionary counterparts related by vertical descent. Because of lineage-specific duplications, orthologous relationships in many cases exist between gene (protein) families, rather than between individual proteins, hence 'orthologous groups' (including only lineage-specific duplications in a COG is the principle of this analysis; in practice, because of insufficient resolution of sequence comparisons, certain COGs may include ancestral duplications). The principal phylogenetic classification in the COG database is overlaid with functional classification and annotation based on detailed sequence and structure analysis and published experimental data. The COG system has been designed as a platform for evolutionary analyses and for phylogenetic and functional annotation of genomes. The COGNITOR program associated with the COGs allows one to fit new proteins into existing COGs. The central tenet of this analysis is that, if it can be shown that the protein under analysis is an ortholog of functionally characterized proteins from other genomes, this functional information can be transferred to the analyzed protein with considerable confidence. In addition to COGNITOR, the COG system includes certain higher-level functionalities, such as analysis of phylogenetic patterns and co-occurrence of genomes in COGs. The current (as of 1 June, 2000) system consists of 2,112 COGs that encompass about 27,000 proteins from 21 completely sequenced genomes. Here we describe the application of the COGs to the systematic annotation and evolutionary analysis of two recently sequenced archaeal genomes, those of the euryarchaeon Pyrococcus abyssi and the crenarchaeon Aeropyrum pernix. These genomes were selected to compare the utility of the COGs for the annotation of two types of genomes - one that is closely related to another genome already included in the system, as Pyrococcus abyssi is to P. horikoshii, and one that represents a group previously not covered by the COGs, the Crenarchaeota. We show here the relatively low error rate of the COG-assisted analysis and its contribution to a significant number of new functional predictions. Emphasis is on using the COG approach to identify features of the A. pernix genome that are shared among all Archaea and those that distinguish Crenarchaeota from Euryarchaeota. Thus this work had a dual focus: first, to explore the potential of the COG system for genome annotation; and second, to use the COG approach to reveal important trends in archaeal genome evolution. It should not be construed as a comprehensive analysis of any particular genome or a comprehensive comparative and evolutionary study; addressing each of these tasks would require the use of several additional methodologies.

CONCLUSION

Conclusions The annotation of a new genome is likely to be as good as the database(s) to which it is compared. The COG database was constructed on the phylogenetic principle of protein classification, namely clustering by (probable) orthology. In addition, considerable effort has been invested in the functional characterization and classification of the COGs. As a result, using the COGs for annotating new genomes of organisms that do not belong to already well-characterized groups provides for numerous functional predictions that are not readily attained in more routine annotation protocols. Furthermore, taking advantage of the structure of the COG database, it is possible to reveal the main functional systems of an organism and its probable evolutionary affinities, and to systematically uncover sets of genes whose presence or absence in the given genome is unexpected and informative from an evolutionary standpoint.