

Demography and Phylogeny of Structural and Sequence Proteomes

S.-H. Kim, S.-R. Jun, G. A. Wu, and G. E. Sims, Univ. of California and Lawrence Berkeley National Laboratory, Berkeley CA 94720 USA

A large number of 3-D structures of proteins became available by the recent development of structure determination technology during the last two decades. There are a few "minimal organisms," for which over 95% of the soluble, globular proteins may now be assigned their structural folds. Near-complete structural proteome of one minimal organism is analyzed as they relate to protein functions and fold usage among functional categories. The structural proteome of the organism is then "mapped" on the protein structure universe, a collection of all known protein structures. This "mapping" reveals features that may be interpretable in terms of evolution of protein structure and function.

Similarly, a large number of whole genome sequences and individual gene sequences became available by the explosive development of sequencing technology during the last two decades. Currently, most comparison methods rely on the multiple sequence alignment of one or more common genes among the population examined, and do not utilize the whole genome sequences. We have developed alignment-free methods to compare and organize whole genome sequences (coding and non-coding regions) or whole genome coding sequences (sequence proteomes). The methods allow us to map "demography" of organisms represented by whole genome sequences or whole genome coding region sequences. Such mapping can provide a global view of distribution and relationship among the demographic groups of organisms, and possibly their evolution.

Funded by NIH (P50 GM62412)