COMPUTING THE EMERGENCE AND HISTORY OF MODERN BIOCHEMISTRY AND CELLULAR LIFE F. Aziz[1], A. Nasir[1], K.M. Kim[3], J.E. Mittenthal[2] and G. Caetano-Anollés[1] [1]Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, and [2]Department of Cell and Developmental Biology, University of Illinois, Urbana, IL 61801, USA; [2]Microbial Resource Center, KRIBB, Daejeon 305-806, Korea.

Background: The origin and evolution of modern biochemistry and cellular life is a complex problem that has puzzled scientists for almost a century. The entire history of life can in principle be studied using myriad sequences generated by genomic research. This includes the appearance of the first cells and the emergence of diversified domains of life. However, the use of molecular sequence information for deep phylogenetic exploration is limited by technical (e.g., the problem of Markov convergence and phylogenetic character independence) and biological considerations (e.g. mutation saturation, intra-molecular interactions, horizontal transfer). In contrast, macromolecular structures are evolutionary modules that are highly conserved and diverse enough to enable deep historical inquiry. We have dissected the emergence of the very early macromolecules that populated primordial cells using ideographic (historical, retrodictive) approaches, with the hope that this knowledge will unfold the mechanistic basis of genetics and its links to molecular functions. This is needed for future endeavors in synthetic biology and biotechnology.

Results: Deep evolutionary signals were retrieved from a census of molecular structures and functions in thousands of nucleic acids and millions of proteins using powerful phylogenomic methods. Phylogenies describing the evolution of proteins, proteomes and molecular functions were built from a genomic census of protein structural domains defined at different levels of structural abstraction and associated Gene Ontology (GO) terms in thousands of organisms (e.g. [1-3]). Phylogenomic trees of domains unfolded molecular clocks and timelines of domain appearance, with time spanning from the origin of proteins to the present. Trees of domains and trees of molecular functions revealed: (1) a primordial protein world, (2) the rise by reductive evolution of viruses and Archaea from a primordial stem line of descent, (3) the appearance of structural innovations unique to the diversified domains of life, first in Bacteria and then in Archaea and Eukarya , and (4) an explosion of functions and structures in Eukarya. The process of domain gain and loss was pervasive, with gains exceeding losses along the entire timeline [1]. Reductive forces resulted in compact and more flexible protein structures with short domain linkers. Primordial metabolic domains evolved earlier than informational domains involved in translation and transcription, supporting the metabolism-first hypothesis and an ancient protein-RNA world rather than the canonical RNA world scenario. Universal trees of proteomes consistently supported the very early cellular origin of viruses and the late appearance of capsids. An analysis of protein domain organization and RNA structure confirms the validity of these evolutionary patterns and a graph theoretical view of domain combination in proteins uncovers complex accretion pathways culminating in a 'big bang' of domain rearrangement.

Conclusions: Clock-like signals revealed that modern biochemistry resulted from gradual accretion and coevolution of molecular parts and molecules. This was made evident in the study of individual molecules (e.g. tRNA , RNase P RNA or rRNA) and macromolecular complexes such as the ATPase synthase (Fig. 1). While the first biochemical functions were metabolic, translation and the genetic code appeared quite late as 'exacting' mechanisms that enhanced protein folding speed and flexibility, impacting the structural make up of proteins and benefiting the search for new molecular functions. Our timelines reveal that genetic memory unfolded only after the rise of viruses but prior to the appearance of diversified archaeal microbes. Remarkably, its debut coincided with the rise of nucleotide and amino acid biosynthetic pathways.

Selected references: [1] Nasir A. et al. (2014) *PLoS Comput. Biol., 10,* e1003452. [2] Nath N. et al. (2014) *PLoS Comput. Biol., 10,* e1003642. [3] Kim K.M. et al. (2014) *J. Mol. Evol., 79,* 79: 240-262. [4] Caetano-Anollés G. and Seufferheld M.J. (2013) *J. Mol. Microbiol. Biotechnol., 23,* 152-177.

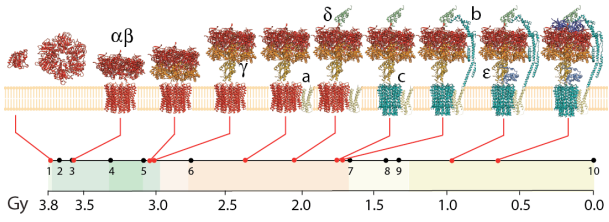

Fig. 1. Structural evolution of the F_1/F_0 ATP synthase complex (from [4]). The phylogenomic-based model describes the most likely and parsimonious scenario of origin and evolution of the complex. Gy, billions of years ago.