Prokaryote Progress

EMSL resources help researchers shed light on organizational structure of bacterial genome

Understanding how DNA-encoded information is organized to give rise to protein-mediated processes will unlock biotechnology breakthroughs in energy production, environmental protection, and medical science. To augment current knowledge of how genetic information is controlled at the molecular level, researchers have developed and applied a new systems-level approach for generating the comprehensive transcription unit architecture of *Escherichia coli* K-12 MG1655. Integrating multiple "-omics" measurements, including proteomics data obtained at the Department of Energy's EMSL, a team of researchers from the University of California San Diego and Virginia Commonwealth University identified roughly five times more transcription units of *E. coli* than were previously known, and improved translation start site mapping over earlier annotations. The team's research was featured on the November 2009 cover of *Nature Biotechnology*.

Cover Story *A graphical representation of part of the* E. coli *genome, the organization of which was elucidated by EMSL users.*

To define the transcription units of *E. coli* and shed light on its genome-wide organizational structure, researchers applied a pioneering, four-step approach to map genetic information to protein function: (1) RNA polymerase binding assays helped the team identify promoters, (2) microarrays revealed gene expression profiles under various growth conditions, (3) a new sequencing method identified transcriptional start sites, and (4) proteomics data were used to determine translational start and stop sites. The proteomics data, obtained with high-throughput mass spectrometry capabilities at EMSL, both improved upon and validated data from previous studies—the team observed that only 64% of translation start sites had been accurately identified by annotation tools, but greater than 99% of translation stop sites were matched correctly. In total, the researchers' integrated process was able to identify 4,661 transcription units—far more than the 875 that had previously been identified through experiments. This new wealth of data for *E. coli* has great potential to allow researchers to manipulate the bacterium's functions in valuable ways, such as for assisting in bioremediation, optimizing industrial chemical production, and protecting human health.

Scientific impact: The team's integrated systems-level approach, which connects measurements made across the bacterium's genome, transcriptome, and proteome, is broadly applicable to many organisms. The resulting knowledge of *E. coli* K-12 MG1655 accelerates the construction of its complete transcriptional and translational regulatory network—opening the door to unprecedented understanding of complex biological functions. Moreover, this work supports EMSL's goal to predict biological functions from molecular and chemical data.

Societal impact: Understanding genetic information flow can lead to significant biotechnology advancements. Armed with new knowledge, researchers may be able to modify *E. coli* to efficiently convert biomass into biofuels, manipulate *E. coli* to consume environmental contaminants, or develop drug therapies that inhibit bacterial infection without damaging the host.

Reference: Cho BK, K Zengler, Y Qiu, YS Park, EM Knight, CL Barrett, Y Gao, and BO Palsson. 2009. "The Transcription Unit Architecture of the *Escherichia coli* Genome." *Nature Biotechnology* 27(11):1043-1051. doi: 10.1038/nbt.1582

Acknowledgment: This work was supported by the US National Institutes of Health and DOE's Office of Science.
User Proposal: 25660