EVALUATION OF SEQUENCE/ACTIVITY RELATIONSHIPS FOR MORE THAN 50 PROTEINS: IMPLICATIONS FOR NATURAL AND DIRECTED EVOLUTION, PROTEIN ENGINEERING AND MACHINE LEARNING ALGORITHMS

David Estell, Genencor Technology Center, IFF
dave.estell@iff.com

Key Words: Evolution, Directed Evolution, Protein Engineering, Sequence/Activity, Mutation frequency.

At Genencor, now a division of IFF, inc., we have evaluated more than 600,000 single mutations for multiple properties, in more than 50 different proteins. These proteins include several different classes of enzymes, non-enzyme proteins and antibodies. For every property, the performance of each variant was measured and compared with a parent enzyme to determine a performance index (PI) and a corresponding $\Delta\Delta G_{app} = -RT\ln(P_{variant}/P_{parent})$ for that property and variant.

The frequency of occurrence of up mutations, down mutations and deleterious mutations were determined for each property and each protein. Correlation coefficients were also determined for all properties for each protein. The distributions of $\Delta\Delta G_{app}$ for all properties were determined, and could all be modeled as Gaussian. These data can be used to evaluate and improve strategies for protein engineering and directed evolution, and to guide machine learning strategies. They also shed light on the results of natural evolution, and structural constraints on evolution.