The Future of Macromolecular Crystallography: Rational Structural Genomics

W. L. Duax, R. Huether, D. Dziak,
Hauptman-Woodward Medical Research Institute, 700 Ellicott St.,
Buffalo, NY 14203

The rational design of enzyme inhibitors that are substrate and species specific depends upon the precision and accuracy of alignment of members of the family. We have developed a novel technique that combines structural knowledge and sequence information to identify and align all members of any major family of proteins We have been able to align 20,000 short chain oxidoreductase enzymes of a subset that has the Rossman fold recognition element TGxxxGxG and the catalytic hexad [NAHD]SYKP[ST] (acronym TGYK) with great accuracy from the N to C termini. The alignment is so accurate that we can separate gram positive from gram negative bacteria and isolate all members of most phylum, class, orders, family and genus. We can correlate variation in 2 positions that determine cofactor recognition with 5 residues that define 300 known or potential substrates with additional residues that determine the details of specific oligomaric aggregation. We can identify a distinction between the two largest TGYK subfamilies, the 1800 member β-keto acyl carrier protein reductase family present in all bacteria and hundreds of eukaryotes and the acetoacetyl CoA reductase family that is present only in α, β and γ proteobacteria on the basis of amino acids in just three positions in the sequence. We achieve exquisitely accurate alignment by locating a few residues (primarily Glycine, Proline, Alanine and Arginine residues) that are fully conserved in all 20,000 members of the family and by determining precisely the location and minimum size of indels required to align all members of family. The GARP residues are critical to the alignment because of their stereochemical properties. Glycines having positive phi values that were embedded early in folded proteins are conserved throughout the evolution of those proteins families. These results support conclusions based upon analysis of multiple open reading frames and codon bias in actinobacteria and deltaproteobacteria that a subset of species in these phylums evolved at a time when the defined genetic code was composed of only triples that end in G and C.

Support in part by: Mr Roy Carver, Stafford Graduate Fellowship, Caerus Forum Fund and The East Hill Foundation.