

Reading guide for Wednesday, September 12, 2007. Bioc 565, Dr. Cordes
Methods for comparing and classifying protein structures

primary literature reading:

CATH- a hierarchic classification of protein domain structures

C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells and J.M. Thornton
Structure (1997) **5**, 1093-1108.

additional optional literature reading:

Protein structure comparison by alignment of distance matrices

L. Holm and C. Sander
J Mol Biol (1993) **233**, 123-138.

Overview:

Your primary reading for today describes CATH, one of the three major systems for hierarchically classifying the known protein structures. CATH is what is known as a “curated” classification system in that it is not fully automated, but also incorporates the judgment of human experts as to how a protein should be classified. Not only does this paper offer a nice window on how people think about differences and similarities in structures, it also has figures that give a nice visual idea of the diversity present in the "universe" of proteins.

The additional suggested reading describes the DALI algorithm, which is an automated method for structure comparison using the concept of a “distance matrix”, which is basically the same thing as a "contact map" or "diagonal plot" we covered in the section on the anatomy of protein structure. DALI is the basis for FSSP, another of the three major classification systems. The two papers make a nice compare/contrast study, so try to read them both if you can find the time.

Refer to the Powerpoint notes for broader information to place the papers in context. DALI is not covered in the main Powerpoint notes, but there is a Powerpoint appendix posted which goes over DALI in some detail.

For the primary reading, I'd like you to think about the following questions:

Isn't this boring? Why care about classifying proteins of known structure?

What does CATH stand for?

What is the fundamental unit of structure classification used? Is it entire proteins or just parts of them?

CATH is a “semi-automatic” structure classification system. What part or parts of the classification system are done manually?

Are all protein domains with the same fold (topology) homologous?

Do all protein domains which have same fold have the same function?

What is a “superfold”?

What is the difference between a “homologous superfamily” and a “sequence family”?