MS22.03.05 THE PROTEIN DATA BANK AND THE CAMBRIDGE STRUCTURAL DATABASE: INTER-RELATIONSHIPS IN CONSTRUCTION AND UTILIZATION.

Frank H. Allen, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 lEZ, England

Proteins are chemically simple but structurally complex, while small molecules are chemically complex but structurally simple. Because of, or even despite, these (rather generalised) differences, there are a number of areas in which the two Centres can collaborate in database construction, and many areas in which the two databases can be used together in research applications.

Both databases are growing at speed. The present PDB doubling period is ca. 3 years - similar to that of the CSD in the early 1970's - while the CSD currently doubles every 6-7 years. This is a crucial time for database creators, as experimentalists, journals and databases are brought ever closer by rapid communications. Novel publication routes are emerging, but we must ensure that data capture and integrity are also improved by these new mechanisms. Both databases are automating direct data deposition procedures based on CIF, mmCIF and MIF. More specifically, the PDB and CSD are collaborating to encode atomic level chemical connection tables for protein-bound ligand molecules, which will then be available for 2D-substructure and 3D-structure searching by suitable software.

At the applications level, small molecule data are used to enhance macromolecular model-building, structure refinement and interpretation. Protein data, particularly from protein-ligand complexes, can indicate the chemical complexity needed to design novel biologically active ligands. Knowledge derived from small molecule structures can, in its turn, indicate possible binding modes for proposed new actives. These synergies will be illustrated in the talk.

MS22.03.06 IMPROVING THE QUALITY OF NMR AND CRYSTALLOGRAPHIC PROTEIN STRUCTURES BY MEANS OF A CONFORMATIONAL DATABASE POTENTIAL DERIVED FROM STRUCTURE DATABASES.

G. Marius Clore, John Kuszewski, Angela M. Gronenborn, Laboratory of Chemical Physics, Building 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 208920520

A new conformational database potential involving dihedral angle relationships in databases of high resolution highly refined protein crystal structures is presented as a method for improving the quality of structures generated from NMR data. The rationale for this procedure is based on the observation that uncertainties in the description of the non-bonded contacts present a key limiting factor in the attainable accuracy of protein NMR structures. The idea behind the conformational database potential is to restrict sampling during simulated annealing refinement to conformations that are likely to be energetically possible by effectively limiting the choices of dihedral angles to those that are known to be physically realizable. In this manner, the variability in the structures produced by this method is primarily a function of the experimental restraints, rather than an artifact of a poor non-bonded interaction model. We tested this approach with the experimental NMR data (comprising an average of about 30 restraints per residue and consisting of interproton distances, torsion angles, $^3J_{HNa}$ coupling constants, and ^{13}C chemical shifts) used to previously calculate the solution structure of reduced human thioredoxin. Incorporation of the conformational database potential into the target function used for refinement (which also includes terms for the experimental restraints, covalent geometry, and non-bonded interactions in the form of either a repulsive, repulsive-attractive or 6-12 Lennard-Jones potential) results in a significant improvement in various quantitative measures of quality (Ramachadran plot, sidechain torsion angles, overall packing). This is achieved without compromising the agreement with the experimental restraints and the deviations from idealized covalent geometry which remain within experimental error, and the agreement between calculated and observed ^{1}H chemical shifts which provides an independent NMR parameter of accuracy. The method is equally applicable to crystallographic refinement, and should be particular useful during the early stages of either an NMR or crystallographic structure determination and in cases where relatively few experimental restraints can be derived from the measured data (due, for example to broad lines in the NMR spectra or to poorly diffracting crystals).

MS22.03.07 POLYMER NUCLEOTIDES, PEPTIDES, SACCHARIDES AND THE PDB.

Wolfram Saenger, Institut fur Kristallographie, Freie Universitat Berlin Takusk.6, 14195 Berlin, FRG

In modern days, publication of the crystal or NMR structure of a biological macromolecule is linked with deposition of atomic parameters in the PDB. Problems arise with older data in the literature which were not deposited, in those days, with the data bank. Notorious are data from X-ray fiber diffraction analyses on polynucleotides, -peptides and -saccharides. Another problem may be faced if a crystal structure determination is only available at such low resolution that a complete chain tracing is not possible. Presented here are our own experiences with cellulose and photosystem I.

PS22.03.08 STRUCTURAL DIVERSITY OF SEQUENTIALLY IDENTICAL SUBSEQUENCES OF PROTEINS

Sucha Sudarsanam and Subhashini Srinivasan, Department of Protein Chemistry, Immunex Corporation, Seattle, WA 98101

Spectroscopic studies indicate that protein folding occurs through the formation of stable secondary structures as intermediates. Consequently, tertiary structure prediction algorithms have approached protein folding by attempting to predict secondary structures and assembling them to form tertiary structures. The success of these algorithms depend on the accurate prediction of local structures encoded by subsequences, herein termed n-mers, of proteins. In this context, one of the important questions is: what is the minimum number of amino acids needed to form unique structures that are stabilized only by local interactions? The most recent analysis (Cohen et al., Prot. Sci. **2**, 2134-2145, 1993) using the July 1990 release of the PDB found that identical 6-mers can have dissimilar structures. Given the explosive growth of the PDB since then, we have revisited this question.

We have analyzed unrelated protein structures (as measured by pairwise sequence identity after an optimal global sequence alignment) in the most recent release of the PDB for identical n-mers. A database consisting of sequences of polypeptide chains along with their backbone dihedral angles f_{i+1}, y_i, where i = 1, m - 1 and m is the length of a polypeptide chain, was constructed using procedures described earlier (Sudarsanam et al., Prot. Sci., **4**, 1412-1420, 1995). This database can be thought of as a "condensed" version of the PDB with sequence and structural information for backbone conformations. For each polypeptide chain n-mers, where n >= 5, were searched against the database for identical matches. Structural similarity of a pair of n-mers was measured by backbone root mean square deviation.

We find the population of 6-mers with identical sequences but dissimilar structures have increased since the last study. In addition, we find at least one pair of identical 7-mers with dissimilar structures. The ability of identical n-mers to adopt different conformations emphasizes the complex interplay of short and long range interactions in protein folding which will be discussed.