International Tables for Crystallography (2006). Vol. F, Chapter 22.4, pp. 558–574.

22.4. The relevance of the Cambridge Structural Database in protein crystallography

BY F. H. ALLEN, J. C. COLE AND M. L. VERDONK

22.4.1. Introduction

At its inception in the late 1960s, the Cambridge Structural Database (CSD: Allen, Davies *et al.*, 1991; Kennard & Allen, 1993) was one of the first scientific databases for which numerical data were the primary objective of the compilation. Thus, the CSD provides not only a fully retrospective bibliography of the structure determination of organic and metallo-organic compounds, but also gives immediate access to the primary results of each diffraction experiment: the space group, cell dimensions and fractional coordinates that define each structure at atomic resolution. In the late 1960s, the world output of small-molecule structures was just a few hundred per year and it was possible to use existing printed compilations to ensure that the developing CSD was fully retrospective. Despite this comprehensive nature, it has taken time for the CSD to have significant scientific impact as a research tool in its own right, and to be recognized as a source of structural knowledge that is applicable across a broad spectrum of structural chemistry.

There are two reasons for this rather gradual uptake. First, it took time to devise and implement software for the validation and organization of the data. Secondly, and most importantly, it was necessary to develop software for database searching, particularly for locating chemical substructures, and for data analysis and visualization. It was not until the late 1970s that the first comprehensive software systems became available and began to be widely distributed to scientists in academia and industry. Nevertheless, a number of highly influential database analyses were performed prior to 1980, and the proper numerical analysis and statistical treatment of bulk geometrical data began to receive attention (see *e.g.* Murray-Rust & Bland, 1978; Murray-Rust & Motherwell, 1978; Taylor, 1986). This software and its successors at last allowed the types of geometrical surveys, analyses and tabulations carried out manually by early practitioners such as Pauling (1939), Sutton (1956, 1959) and Pimental & McClellan (1960) to be executed automatically in a few minutes of increasingly powerful CPU time.

The early development of applications software simultaneously with methods for the acquisition and validation of new structural data was crucial for the CSD. Developments in structure-determination theory, allied to technological improvements in data collection and the ever increasing speed and capacity of modern computers, led to such a rapid expansion that the archive of May 1999 now contains more than 200000 crystal structures, a total that doubles approximately every seven years. The literature is now so vast, so chemically diverse and so widely spread that it is virtually impossible for individual scientists to maintain current awareness without recourse to database facilities. It is now impossible to carry out viable systematic analyses without recourse to database technology. This chapter focuses primarily on the structural knowledge that is provided by such analyses, and that is relevant to the determination, refinement, validation and systematic study of macromolecular structures. However, the validity of these results depends crucially on two factors: the *completeness* of the archive and the *accuracy* with which the data are recorded. Hence, it is appropriate to preface the chapter with some comparative comment on these fundamental issues as they apply to the small-molecule and macromolecular structure archives.

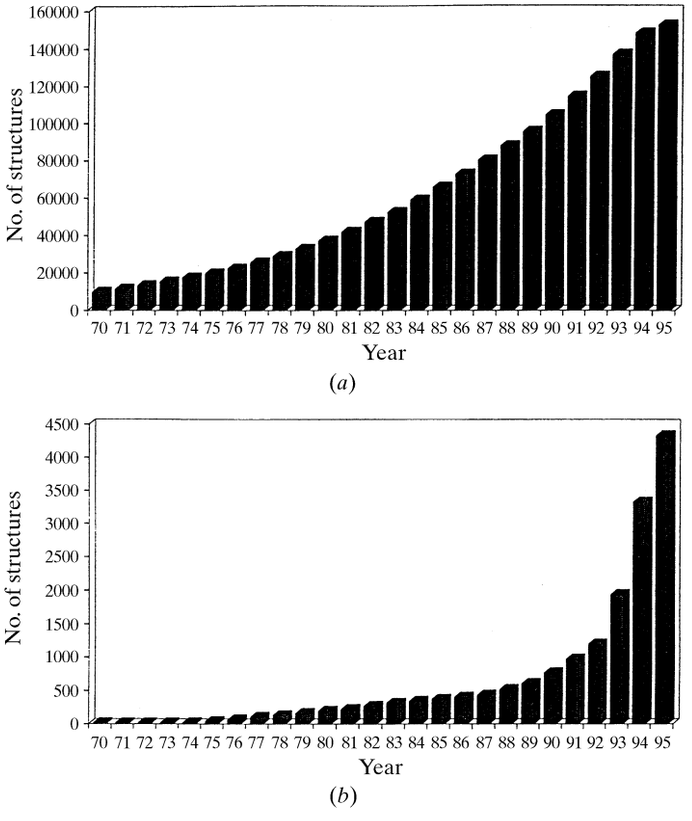


Fig. 22.4.2.1. (*a*) Growth rate of the CSD and (*b*) growth rate of the PDB, in terms of the numbers of structures published per annum for the period 1970–1995.

22.4.2. The CSD and the PDB: data acquisition and data quality

22.4.2.1. *Statistical inferences*

With a current total of 200000 structures and a doubling period of seven years (Fig. 22.4.2.1*a*), we may expect at least half a million small-molecule crystal structures to be in the CSD by the year 2010. The Protein Data Bank (PDB) (Abola *et al.*, 1997; Berman *et al.*, 2000), which began operations in the mid-1970s, and the Nucleic Acid Database (NDB) (Berman *et al.*, 1992) are the international repositories for macromolecular structure information. Input to the PDB was initially slow but is now showing a rapid growth rate reminiscent of the CSD of the 1970s (Fig. 22.4.2.1*b*). The PDB archive has a current total of *ca* 8500 structures (mid-1999) and a doubling period of close to two years. As with the CSD, this early *high rate* of growth will almost certainly decrease, thus increasing the doubling period. Nevertheless, by the year 2010, we might expect the PDB to contain more than 100000 structures.

22.4.2.2. *Data acquisition and completeness*

Given the size and diversity of the CSD, it is amazing that searches for some common chemical substructures often yield far fewer hits than might have been expected. Sometimes, the absence of just a few key CSD entries would have negated a successful systematic analysis: some points in a graph would have been missing and a correlation would not have been detected. Similarly, completeness of the PDB is vital for the future of 'data mining' or 'knowledge engineering' in the macromolecular arena.

Data acquisition by the PDB has always had one valuable advantage in comparison with the CSD. The volume of numerical data generated by a protein structure determination is far too large

for primary publication or hard-copy deposition. Thus, the PDB has always acquired data through direct deposition in electronic form, and authors have usually been involved in the validation of their entries. Further, it is a mandatory requirement of the vast majority of journals, and a clear recommendation of appropriate professional organizations, that prior deposition with the PDB is an essential precursor to primary publication. This key involvement of the PDB in the publication process acts as a vital guarantee of the completeness of the archive. The prior-deposition rule must be rigidly adhered to for the long-term benefit of science.

22.4.2.3. *Standard formats: CIF and mmCIF*

The CSD, on the other hand, reflects the published literature, and much of its data content has been re-keyboarded from hard-copy material. The Cambridge Crystallographic Data Centre (CCDC) is now beginning to receive significant amounts of electronic input, a development that owes much to the rapid international acceptance of an agreed standard electronic interchange format, the crystallographic information file or CIF (Hall *et al.*, 1991), and the rapid incorporation of CIF generators within most major structure solution and refinement packages. The CIF offers many advantages, some of which are only just being addressed within the CSD: (*a*) a clear definition of input data items and their representation; (*b*) a significant reduction in time spent correcting simple typographical errors; and (*c*) the possibility of enhancing the overall database content through the electronic availability of *all* information from the analysis, *i.e.* more than could reasonably be re-typed from hard-copy material. For the PDB, the recent adoption of the macromolecular CIF (mmCIF) as the agreed international standard offers similar advantages. This development, together with advances in communications technology, now make it possible to automate the deposition process more effectively, but the advantages of mmCIF can only be fully realized once it also becomes a standard output format of all of the relevant software packages.

22.4.2.4. *Structure validation*

The value of research results derived from the CSD and the PDB depends crucially on the accuracy of the underlying data [see *e.g.* Hooft *et al.* (1996) with respect to protein data]. As with the early CSD, much current research involves use of data from the developing PDB to establish rules and protocols for the validation of new protein structures (see *e.g.* Laskowski *et al.*, 1993). This activity, in turn, means that earlier entries in the archive may have to be reassessed periodically to bring their representations into line with best current practice. This sequence of events was commonplace in the CSD of the 1970s and, even now, new structure types entering the CSD can still provoke a reassessment of subclasses of earlier entries.

Secondly, it is important that errors and warnings raised by validation software have clear meanings and that validation results are clearly encoded within each entry. The end user can then make informed choices about which entries to include (or not) in any given application. Recent moves to apply a range of agreed and unambiguous primary checks to new data, and to require resolution of any problems prior to the issue of a publication ID code, represent an important development.

22.4.3. Structural knowledge from the CSD

22.4.3.1. *The CSD software system*

Structural knowledge from the CSD is reflected principally in the geometries of individual molecules, extended crystal structures and, most importantly, through systematic studies of the geometrical characteristics of large subsets of related substructural units. Software facilities for search, retrieval, analysis and visualization of CSD information are fully described in Chapter 24.3. The system allows for the calculation of a very wide range of geometrical parameters, both intramolecular and intermolecular. Most importantly, chemical substructural search fragments may be specified using normal covalent bonding definitions (single, double, triple *etc.*), limiting non-covalent contact distances and other geometrical constraints. For each instance of a search fragment located in the CSD, the system will compute a user-defined set of geometrical descriptors. The full matrix, $G(N, p)$, of the p geometrical parameters for each of the N fragments located in the CSD can then be analysed using numerical, statistical and visualization techniques to display individual parameter distributions, to compute medians, means and standard deviations, and to examine the geometrical data for correlations or discrete clusters of observations that may exist in the p-dimensional parameter space.

22.4.3.2. *CSD structures and substructures of relevance to protein studies*

Table 22.4.3.1 presents statistics for the 3137 structures of amino acids and peptides that are available in the CSD of April 1998 (containing 181 309 entries). Although this represents less than 2% of CSD information, some may consider that these are the only entries of real interest in molecular biology. In certain cases, *e.g.* for the derivation of very precise molecular dimensions and for some conformational work, this may be true. However, the real issue concerns the *transferability* of CSD-derived information to the protein environment. It is the biological relevance of a chemical

Table 22.4.3.1. *Summary of amino-acid and peptide structures available in the CSD (April 1998, 181 309 entries)*

(*a*) Overall statistics

Structures	No. of entries
α-Amino acids (any organic) *	3137
Peptides (standard or modified standard α-amino acids) †	1430

(*b*) Peptide statistics

	No. of CSD entries	
No. of residues	Acyclic	Cyclic
2	543	123
3	249	45
4	76	50
5	62	44
6	20	73
7	14	15
8	19	32
10	16	19
11	4	10
12	2	11
13	—	—
14	1	—
15	3	2
16	3	—

* Any organic structure containing the α-amino acid functionality.
† The standard amino acids (those normally found in proteins) may be modified by substitution in these peptides.

Table 22.4.3.2. *CSD entry statistics for selected metal-containing structures*

CSD entries ($R < 0.10$) containing *M* and (N or O). No additional transition metals were allowed to occur in the Na, K, Mg and Ca structures cited.

Metal	No. of CSD entries
Na	1189
K	987
Mg	510
Ca	469
Zn	1996

substructure (inter- or intramolecular) that is important, and this consideration immediately brings much larger subsets of CSD entries into play. Information such as van der Waals radii can be derived from the CSD as a whole, while more specific information concerning, for example, biologically important metal coordination geometries can be derived from appreciable subsets of the total database, as shown in the statistics of Table 22.4.3.2.

22.4.3.3. *Geometrical parameters of relevance to protein studies*

Precise geometrical knowledge from atomic resolution studies of small molecules is important in the macromolecular domain since it provides: (*a*) geometrical restraints and standards to be applied during protein structure determination, refinement and validation; (*b*) model geometries for liganded small molecules and information about their preferred modes of interaction with the host protein; (*c*) details of metal coordination spheres and geometries that are likely to be observed in metalloproteins; and (*d*) information from which force field and other parameters may be derived. Thus, the types of study discussed in this chapter are concerned with retrieving systematic knowledge concerning:

(1) molecular dimensions: bond lengths and valence angles;

(2) conformational features: torsion angles that describe acyclic and cyclic systems;

(3) metal coordination-sphere geometries: coordination numbers, metal–ligand distances and inter-ligand valence angles;

(4) general non-bonded contact distances: van der Waals radii;

(5) hydrogen-bond geometries: distances, angles, directional properties;

(6) other non-bonded interactions: identification and geometrical description;

(7) formation of preferred atomic arrangements or motifs involving non-covalent interactions.

In this short overview, which deals with such a broad range of structural information, our literature coverage is, of necessity, highly selective. In each area, we have tried to cite the more recent papers, from which leading references to earlier studies can be located. We also draw attention to a number of recent monographs in which a variety of CSD analyses are comprehensively cited and discussed: *Structure Correlation* (Bürgi & Dunitz, 1994), *Crystal Structure Analysis for Chemists and Biologists* (Glusker *et al.*, 1994), *Hydrogen Bonding in Biological Structures* (Jeffrey & Saenger, 1991) and *Crystal Engineering: the Design of Organic Solids* (Desiraju, 1989). Finally, we note the CCDC's own database of published research applications of the CSD. The DBUSE database currently contains literature references and short descriptive abstracts for nearly 700 papers. It forms part of each biannual CSD release and is fully searchable using the *Quest3D* program.

22.4.4. Intramolecular geometry

22.4.4.1. *Mean molecular dimensions*

The work of Pauling (1939) represented the first systematic attempt to derive mean values for bond lengths and valence angles from the limited structural data available at that time. This work resulted in the definition of covalent bonding radii for the common elements and had a seminal influence on the development of chemistry over the past half century. Further tabulations appeared sporadically until the publication in 1956 and 1959 of the major compilation *Tables of Interatomic Distances and Configuration in Molecules and Ions,* edited by Sutton (1956, 1959), by The Chemical Society of London. Kennard (1962) extended the available data for bonds between carbon and other elements.

In the mid-1980s, the CCDC and its collaborators compiled updated tables of mean bond lengths for both organic (Allen *et al.*, 1987) and organometallic and metal coordination compounds (Orpen *et al.*, 1989). Both compilations were based on the CSD of September 1985 containing 49 854 entries. Of these, 10 324 organic structures and 9802 organometallics or metal complexes satisfied a variety of secondary selection criteria, and were used in the analysis. For each bond length, both compilations present the mean, its estimated standard deviation and the sample standard deviation, together with the median value of the distribution and its upper and lower quartile values. The organic section describes 682 discrete chemical bond types involving 65 element pairs. Of these, 511 (75%) involve carbon, and 428 (63%) involving only carbon, nitrogen and oxygen are relevant to protein studies. The organometallic and metal complex compilation presents similar statistics for 325 different bond types involving *d*- and *f*-block metals. It is planned to automate and systematize the production of such tabulations, so that they can be dynamically updated in computerized form, as part of CCDC's ongoing development of knowledge-based structural libraries.

More recently, Engh & Huber (1991) have generated sets of mean bond lengths and valence angles from peptidic structures retrieved from the 80 000 entries then available in the CSD. Their compilations are based on 31 atom types which are most appropriate to the protein environment and are well represented in CSD structures. These authors note that such knowledge, together with torsional and other information, is vital to the determination, refinement and validation of protein structures. Prior to their detailed CSD analysis, some of the parameters used for these purposes had been determined with a lower accuracy than was required by the diffraction data. For this reason, and particularly for use with higher-resolution protein data, they recommend that the most accurate parameters possible should always be used.

Systematic use of CSD data generates mean values together with standard deviations for both the sample and the mean. The sample standard deviations provide information about the spread of each parameter distribution, *i.e.* information about the variability of each parameter which can be parameterized as force constants. Comparative refinements of selected proteins showed that the new CSD-based parameters yielded significant improvements in *R* factors and in geometry statistics. Finally, Engh & Huber (1991) remark that their results should be updated regularly as the quantity and quality of data in the CSD increase with time. Apart from producing more precise estimates of mean values, incorporation of more protein-relevant atom types into the schema should then be possible.

22.4.4.2. *Conformational information*

Torsion angles are the natural measures of conformational relationships within molecules. If we specify a chemical substructure involving a central bond of interest, then the CSD system

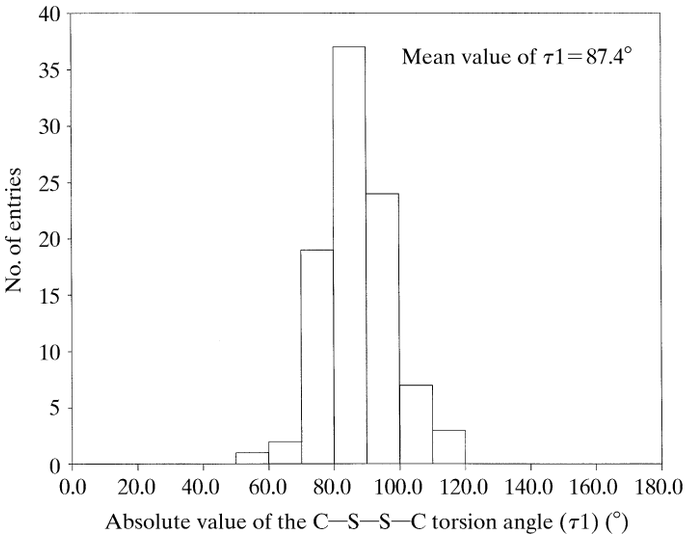


Fig. 22.4.4.1. Distribution of torsion angles in C(sp^3)—S—S—C(sp^3) substructures located in the CSD.

will display the distribution of torsion angles about that bond, computed from the tens, hundreds, or even thousands of instances located in the database. Examination of these *univariate distributions* will reveal any conformational preferences that may exist in small-molecule crystal structures. This approach is illustrated by the histogram of Fig. 22.4.4.1, which shows the torsional distribution about S—S bridge bonds in C(sp^3)—S—S—C(sp^3) substructures located in the CSD. Clearly, there is a preference for a perpendicular conformation in the CS—SC unit. This corresponds well with values observed for cysteine bridges in protein structures, and with theoretical calculations on small model compounds.

The interrelationship between two torsion angles can be visualized by plotting them against each other on a conventional 2D scattergram. In the small-molecule area, the distribution of data points in these scattergrams can reveal conformational interconversion pathways (Rappoport *et al.*, 1990) or show areas of high data density corresponding to conformational preferences (Schweizer & Dunitz, 1982). The best known *bivariate distribution* is the Ramachandran plot of peptidic φ–ψ angles, which is universally used to assess the quality of protein structures and to identify structural features. Ashida *et al.* (1987) performed an extensive analysis of peptide conformations available in the CSD and present torsional histograms, a Ramachandran plot, and a variety of other visual and descriptive statistics that summarize this data set.

It is often necessary to use three or more torsion angles to define the conformation of, *e.g.*, a side chain or flexible ring. Here, *multivariate* statistical techniques (Chatfield & Collins, 1980; Taylor, 1986) have proved valuable for extracting information from the matrix $T(N, k)$ that contains the k torsion angles computed for each of the N examples of the substructure in the CSD. Two methods, both available within the CSD system software described in Chapter 24.3, are commonly used to visualize the k-dimensional data set and to locate natural sub-groupings of data points within it.

Principal component analysis (PCA) (Murray-Rust & Motherwell, 1978; Allen, Doyle & Auf der Heyde, 1991, Allen, Howard & Pitchford, 1996) is a dimension-reduction technique which analyses the variance in $T(N, k)$ in terms of a new set of uncorrelated, orthogonal variables: the principal components, or PCs. The PCs are generated in decreasing order of the percentage of the variance that is explained by each of them. The hope is that the number of PCs, p, that explains most of the variance in the data set is such that $p \ll k$, so that a few pairwise scatter plots with respect to the new PC axes will provide useful visualizations of the complete data set. For cyclic fragments, PCA results are closely related to those obtained using the ring-puckering methodology of Cremer & Pople (1975). Cluster analysis (CA) (Everitt, 1980; Allen, Doyle & Taylor, 1991) is a purely numerical method that attempts to locate discrete groupings of data points within a multivariate data set. CA uses 'distances' or 'dissimilarities' between pairs of points in a k-dimensional space as its working basis, and a very large number of clustering algorithms exist. The mathematical basis of both of these techniques, the modifications that are needed to account for topological symmetry in the search fragment and examples of their application have been reviewed by Taylor & Allen (1994).

Preliminary work using the concepts of machine learning (Carbonell, 1989) for knowledge discovery and classification have also been carried out using the CSD (see *e.g.* Allen *et al.*, 1990; Fortier *et al.*, 1993). In particular, conceptual clustering methods have been applied to a number of substructures (Conklin *et al.*, 1996) and the results compared with those obtained by the statistical and numerical methods described above. Similar techniques are also being used for the classification of protein structures (see *e.g.* Blundell *et al.*, 1987).

22.4.4.3. *Crystallographic conformations and energies*

Crystallographic conformations obviously represent energetically accessible forms. However, for use in molecular-modelling applications, the key question must be asked: Are the condensed-phase crystallographic observations a good guide to conformational preferences in other phases? The indications are that the answer is 'yes' from the types of studies exemplified or cited in the previous section: there appears to be a clear *qualitative* relationship between crystallographic conformer distributions and the low-energy features of the appropriate potential energy hypersurface, although the estimation of absolute energies from the relative populations of these distributions is not appropriate (Bürgi & Dunitz, 1988).

Allen, Harris & Taylor (1996) addressed this question in a systematic manner for a series of 12 one-dimensional (univariate) conformational problems. All of the chosen substructures [simple derivatives of ethane, involving a single torsion angle (τ) about the central C—C bond] were expected to show one symmetric (*anti*, $\tau \simeq 180°$) energy minimum and two symmetry-related asymmetric (*gauche*, $\tau \simeq \pm 60°$) minima. For each substructure, the crystallographic torsional distribution was determined from the CSD and compared with the 1D potential-energy profile, computed using *ab initio* molecular-orbital methods and the 6-31G* basis set. Close agreement was observed between the experimental condensed phase results and the computed *in vacuo* data. Taken over all 12 substructures, the *ab initio* optimized values of the asymmetric (*gauche*) torsion angle vary from $< 55°$ to $> 80°$, and a scatter plot of these optimized values *versus* the mean crystallographic values for *gauche* conformers is linear, with a correlation coefficient of 0.831. Two other results of the study were that: (*a*) torsion angles with higher strain energies (> 4.5 kJ mol^{-1}) are rarely observed in crystal structures ($< 5\%$); and (*b*) taken over many structures, conformational distortions due to crystal packing appear to be the exception rather than the rule.

22.4.4.4. *Conformational libraries*

In essence, the CSD can be regarded as a huge library of individual molecular conformations. However, to be of general value, it is necessary to distil, store and present this knowledge in an ordered manner, in the form of torsional distributions for specific atomic tetrads *A*—*B*—*C*—*D*. Protein-specific libraries of this type derived from high-resolution PDB structures are commonly used as aids to protein structure determination, refinement and validation (Bower *et al.*, 1997; Dunbrack & Karplus, 1993). The information

can either be stored in external databases, or hardwired into the program in the form of rules. However, CSD usage has tended to concentrate on analyses of individual substructures, as noted above, both for their intrinsic interest and to develop novel methods of data analysis. Recently, Klebe & Mietzner (1994) have described the generation of a small library containing 216 torsional distributions derived from the CSD, together with 80 determined from protein–ligand complexes in the PDB. The library was used in a knowledge-based approach for predicting multiple conformer models for putative ligands in the computational modelling of protein–ligand docking. Conformer prediction is accomplished by the computer program *MIMUMBA*. As part of its programme for the development of knowledge-based libraries from the CSD, the CCDC has now embarked on the generation of a more comprehensive torsional library. Here, information is being hierarchically ordered according to the level of specificity of the chemical substructures for which torsional distributions are available in the library.

22.4.4.5. *Metal coordination geometry*

Some 54% of the information content of the CSD relates to organometallics and metal complexes. This reflects the crucial role of single-crystal diffraction analyses in the renaissance of inorganic chemistry since the 1950s, and the fundamental importance of the technique in characterizing the many novel molecules synthesized over the past 40 years. Since ligands containing nitrogen, oxygen and sulfur are ubiquitous, the CSD contains much information that is relevant to the binding of metal ions by proteins [*e.g.* zinc (Miller *et al.*, 1985), calcium (Strynadka & James, 1989) *etc*.]. Some statistics for the occurrence of some common metals having N and/or O ligands are presented in Table 22.4.3.2.

One of the earliest studies (Einspahr & Bugg, 1981) concerned the geometry of Ca–carboxylate binding, with special reference to biological systems. Since that time, a variety of other studies of biologically relevant metal coordination modes have appeared from the laboratories of Glusker, Dunitz and others (see *e.g.* Glusker, 1980; Chakrabarti & Dunitz, 1982; Carrell *et al.*, 1988, 1993; Chakrabarti, 1990*a*,*b*). These studies show, *inter alia*, that α-hydroxycarboxylates and imidazoles such as histidine tend to bind metal ions in their planes, but that alkali metal cations tend to bind carboxylate groups indiscriminately both in-plane and out-of-plane. Chapter 17 of Glusker *et al.* (1994) is a significant source of additional information and leading references to work in this area over the past two decades.

22.4.5. Intermolecular data

Non-bonded interaction geometries observed in small-molecule crystal structures are of great value in the determination and validation of protein structures, in furthering our understanding of protein folding, and in investigating the recognition processes involved in protein–ligand interactions. The CSD continues to provide vital information on all of these topics.

22.4.5.1. *van der Waals radii*

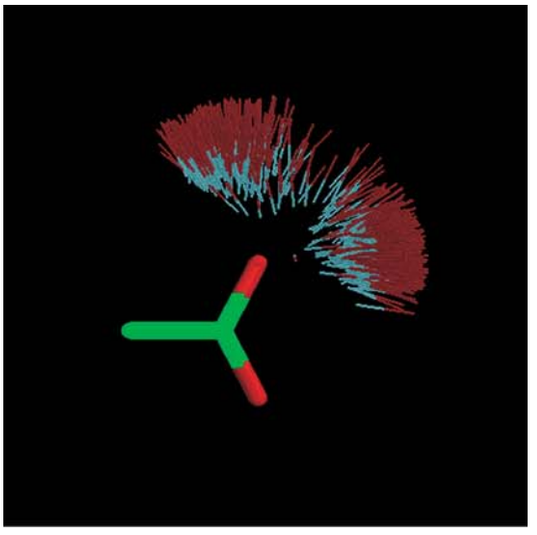
The hard-sphere atomic model is central to chemistry and molecular biology and, to an approximation, atomic van der Waals radii can be regarded as transferable from one structure to another. They are heavily used in assessing the general correctness of all crystal-structure models from metals and alloys to proteins. Pauling (1939) was the first to provide a usable tabulation for a wide range of elements, but the values of Bondi (1964) remain the most highly cited compilation in the modern literature. His values, assembled from a variety of sources including crystal-structure information, were selected for the calculation of molecular volumes and, in his original paper, Bondi (1964) issues a caution about their general validity for the calculation of limiting contact distances in crystals. In view of the huge amount of non-bonded contact information available in the CSD, Rowland & Taylor (1996) recently tested Bondi's statement as it might apply to the common nonmetallic elements, *i.e.* H, C, N, O, F, P, S, Cl, Br and I. They found remarkable agreement (within 0.02 Å) between the crystal-structure data and the Bondi values for S and the halogens, and agreement within 0.05 Å for C, N and O (new values all larger). The only significant discrepancy was for H, where averaged neutron-normalized small-molecule data yield a van der Waals radius of 1.1 Å, 0.1 Å shorter than the Bondi (1964) value. In the specific area of amino-acid structure, Gould *et al.* (1985) have studied the crystal environments and geometries of leucines, isoleucines, valines and phenylalanines. Their work provides estimates of minimum non-bonded contact distances and indicates the preferred van der Waals interactions of these primary building blocks.

22.4.5.2. *Hydrogen-bond geometry and directionality*

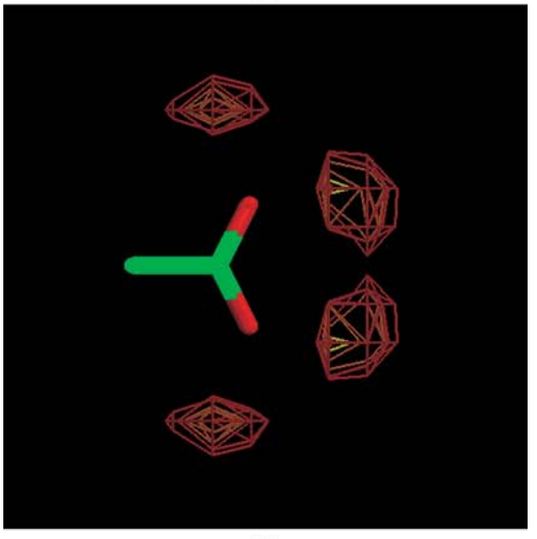
The hydrogen bond is the strongest and most frequently studied of the non-covalent interactions that are observed in crystal structures. As with intramolecular geometries, the first surveys of non-bonded interaction geometries all concerned hydrogen bonds, and were reported long before the CSD existed (Pauling, 1939; Donohue, 1952; Robertson, 1953; Pimentel & McClellan, 1960). The review by Donohue (1952) already contained a plot of N···O distances *versus* C—N···O angles in crystal structures (the C—N groups are terminal charged amino groups), while the review by Pimentel & McClellan (1960) contained histograms of hydrogen-bond distances. Up to the mid-1970s, numerous other studies appeared, *e.g.* Balasubramanian *et al.* (1970), Kroon & Kanters (1974) and Kroon *et al.* (1975), in which all of the statistical analyses were performed manually.

With the advent of the CSD and its developing software system, these kinds of studies became much more accessible and easier to perform, although the non-bonded search facility was only generalized and fully integrated within *Quest3D* in 1992. Thus, Taylor and colleagues reported studies on N—H···O=C hydrogen bonds (Taylor & Kennard, 1983; Taylor *et al.*, 1983, 1984*a*,*b*), Jeffrey and colleagues reported detailed studies on the O—H···O hydrogen bond (Ceccarelli *et al.*, 1981), hydrogen bonds in amino acids (Jeffrey & Maluszynska, 1982; Jeffrey & Mitra, 1984), and hydrogen bonding in nucleosides and nucleotides, barbiturates, purines and pyrimidines (Jeffrey & Maluszynska, 1986), while Murray-Rust & Glusker (1984) studied the directionalities of O—H···O hydrogen bonds to ethers and carbonyls. These studies indicated that hydrogen bonds are often very directional. For example, the distribution of the O—H···O hydrogen-bond angle, after correction for a geometrical factor, peaks at 180° (*i.e.* there is a clear preference for linear hydrogen bonds) and, in carbonyls and carboxylate groups, hydrogen bonds tend to form along the lone-pair directions of the O-atom acceptors (Fig. 22.4.5.1). For ethers, however, lone-pair directionality is not observed, as is illustrated in Fig. 22.4.5.2.

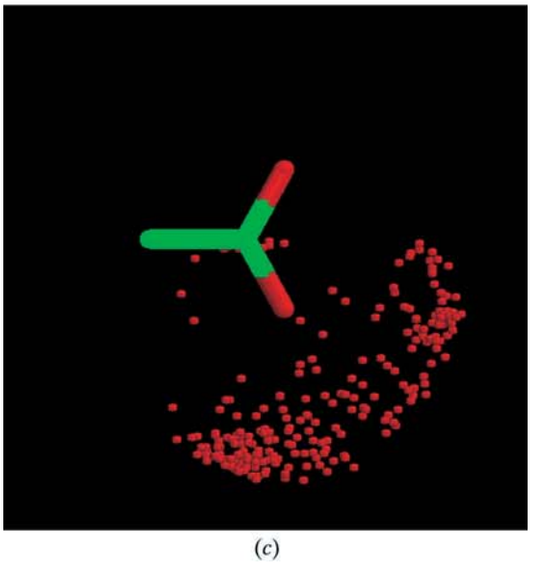
Software availability has facilitated CSD studies of a wide range of individual hydrogen-bonded systems in the recent literature, including studies of resonance-assisted hydrogen bonds (Bertolasi *et al.*, 1996) and resonance-induced hydrogen bonding to sulfur (Allen, Bird *et al.*, 1997*a*). These statistical studies are often combined with molecular-orbital calculations of interaction energies. Some of these studies are cited in this chapter, but the monograph of Jeffrey & Saenger (1991) and the CCDC's DBUSE database are valuable reference sources.



(*a*)



(*b*)



(*c*)

Fig. 22.4.5.1. The IsoStar knowledge-based library of intermolecular interactions: interaction of O—H donors (contact groups) with one of the >C=O acceptors of a carboxylate group (the central group). (*a*) Direct scatter plot derived from CSD data, (*b*) contoured scatter plot derived from CSD data and (*c*) direct scatter plot derived from PDB data.

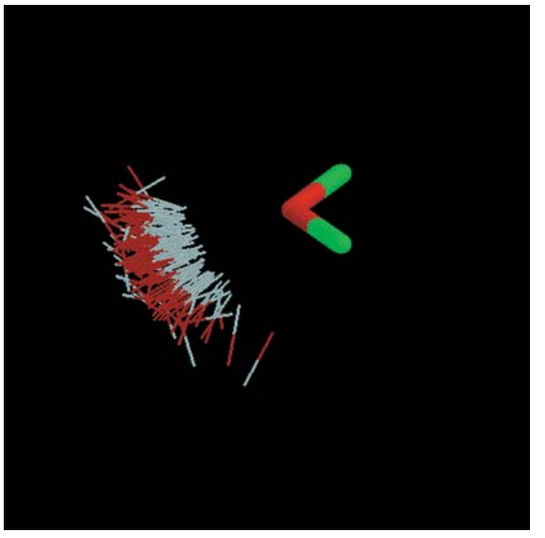


Fig. 22.4.5.2. Distribution of O—H donors around ether oxygen acceptors (CSD data from the IsoStar library, see text).

22.4.5.3. *C—H···X hydrogen bonds*

An important and often underestimated interaction in biological systems is the C—H···*X* hydrogen bond. These bonds have been extensively studied in small-molecule crystal structures, especially in relation to the ongoing discussion as to whether or not they should be called hydrogen bonds. Although Donohue (1968) concluded that the question 'The C—H···O hydrogen bond: what is it?' had only one answer: 'It isn't', a survey of 113 neutron-diffraction structures showed clear statistical evidence for an attractive interaction between C—H groups and oxygen and nitrogen acceptors (Taylor & Kennard, 1982). Later, more evidence for this hypothesis was found, and it was even shown that some C—H···O interactions are directional (Berkovitch-Yellin & Leiserowitz, 1984; Desiraju, 1991; Steiner & Saenger, 1992; Desiraju *et al.*, 1993; Steiner *et al.*, 1996). A continuing area of interest has been to establish the relative donor abilities of C—H in different chemical environments, since spectroscopic data had indicated that donor ability decreased in the order C(sp)—H > C(sp^2)—H > C(sp^3)—H. This general hydrogen-acidity requirement was noted by Taylor & Kennard (1982), and systematically addressed using CSD information by Desiraju & Murty (1987), and by Pedireddi & Desiraju (1992), who derived a novel scale of carbon acidity based on C···O separations in a wide variety of systems containing C—H···O hydrogen bonds. A recent paper (Derewenda *et al.*, 1995) highlights the importance of C—H···O=C bonds in stabilizing protein secondary structure.

22.4.5.4. *O—H···π and N—H···π hydrogen bonds*

Spectroscopic evidence for the existence of N,O—H···π hydrogen bonding to acetylenic, olefinic and aromatic acceptors is well documented (Joris *et al.*, 1968). To our knowledge, the first survey of these interactions in the CSD was carried out by Levitt & Perutz (1988), prompted by observations made in protein structures. A more recent CSD survey of this type of bonding (Viswamitra *et al.*, 1993) has shown that intermolecular examples are clearly observed and that these bonds, although very weak, can be both structurally and energetically significant. Recently, Steiner *et al.* (1995) have presented novel crystal structures, database evidence and quantum-chemical calculations on C≡C—H···π(C≡C) and π(phenyl) bonding. They cite H···C≡C (midpoint) distances as short as 2.51 Å and observe hydrogen-bond cooperativity in extended systems with hydrogen-bond energies in the range 4.2–

9.2 kJ mol^{-1}. Finally, we note that electron-rich transition metals can act as proton acceptors in hydrogen-bond interactions with O—H, N—H and C—H donors. Brammer *et al.* (1995) have reviewed progress in this developing area.

22.4.5.5. *Other non-covalent interactions*

The hydrogen bond, $X(\delta-)$—$H(\delta+)\cdots Y(\delta-)$—$Z(\delta+)$, can be viewed as an (almost) linear dipole–dipole interaction, whose ubiquity in nature is due to the presence of many donor–hydrogen dipoles. In a recent review of supramolecular synthons and their application in crystal engineering, Desiraju (1995) illustrates the structural importance of a wide range of attractive non-bonded interactions that do not involve hydrogen mediacy, and notes the long-term value of the CSD in identifying and characterizing these interactions. The area of weak intermolecular interactions is now a burgeoning one in which the combination of CSD analysis and high-level *ab initio* molecular-orbital calculations is proving important in establishing both preferred geometries and estimates of interaction energies. In this context, the intermolecular perturbation theory (IMPT) of Hayes & Stone (1984), a methodology which is free of basis-set superposition errors, is proving particularly useful.

Some of the earliest CSD studies concerned the geometry and directionality of approach of N and O nucleophiles to carbonyl centres, leading to the mapping of (dynamic) reaction pathways through systematic analysis of many examples of related (static) crystal structures (see Bürgi & Dunitz, 1983, 1994). This work was also extended to a study of the directional preferences of non-bonded atomic contacts at sulfur atoms, initially using S in amino acids but later including other examples of divalent sulfur (Rosenfield *et al.*, 1977). It was shown that C—S—C groups tend to bind positively charged electrophiles in directions that are approximately perpendicular to the C—S—C plane, while negatively charged nucleophiles prefer to bind to S along an extension of one of the C—S bonds.

The strong tendency for halogens X = Cl, Br and I to form short contacts to other halogens, and especially to electronegative O and N atoms (Nyburg & Faerman, 1985) is well known (Price *et al.*, 1994). Recent combined CSD/IMPT studies of C—$X\cdots$O=C (Lommerse *et al.*, 1996) and C—$X\cdots$O(nitro) (Allen, Lommerse *et al.*, 1997) systems showed a marked preference for the $X\cdots$O interaction to form along the extension of the C—X bond, with interaction energies in the range -7 to -10 kJ mol^{-1}. These interactions have been used (Desiraju, 1995) to engineer a variety of novel small-molecule crystal structures, and the few $X\cdots$O interactions observed in protein structures generally conform to the geometrical preferences observed in small-molecule studies.

Interactions involving other functional groups are also of importance, and Taylor *et al.* (1990) used CSD information to construct composite crystal-field environments for carbonyl and nitro groups in their search for isosteric replacements in modelling protein–ligand interactions. Their work showed that many of the short intermolecular contacts made by carbonyl groups are to other carbonyl groups in the extended crystal structure. More recently, Maccallum *et al.* (1995*a*,*b*) have demonstrated the importance of Coulombic interactions between the C and O atoms of proximal CONH groups in proteins as an important factor in stabilizing α-helices, β-sheets and the right-hand twist often observed in β-strands. Their calculations indicate an attractive carbonyl–carbonyl interaction energy of about -8 kJ mol^{-1} in specific cases, and they remark that these interactions are *ca* 80% as strong as the CO$\cdots$HN hydrogen bonds within their computational model. Allen, Baalham *et al.* (1998) have used combined CSD/IMPT analysis in a more detailed study of carbonyl–carbonyl interactions and have shown that (*a*) the interaction is commonly observed in small-molecule structures; (*b*) that the preferred interaction geometry is a dimer motif involving two antiparallel C$\cdots$O interactions, although numerous examples of a perpendicular motif (one C$\cdots$O interaction) were also observed; and that (*c*) the total interaction energies for the antiparallel and perpendicular motifs are about -20 and -8 kJ mol^{-1}, respectively, the latter value being comparable to that computed by Maccallum *et al.* (1995*a*,*b*). In studies with protein structures, it has also been noted that carbonyl–carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid (Deane *et al.*, 1999)

22.4.5.6. *Intermolecular motif formation in small-molecule crystal structures*

Desiraju (1995) has stressed that the design process in crystal engineering depends crucially on the high probabilities of formation of certain well known intermolecular motifs, *e.g.* the hydrogen-bonded dimer frequently formed by pairs of carboxylate groups. By analogy with molecular synthesis, he describes these general non-covalent motifs (which often contain strong hydrogen bonds) as supramolecular *synthons*, and points to their importance in supramolecular chemistry as a whole (see *e.g.* Lehn, 1988; Whitesides *et al.*, 1995). Since protein–protein and protein–ligand interactions are also supramolecular phenomena, it follows that information about common interaction motifs is also of importance in structural biology. A computer program is now being written at the CCDC to establish the topologies, chemical constitutions and probabilities of formation of intermolecular motifs directly from the CSD. Initial results (Allen, Raithby *et al.*, 1998; Allen *et al.*, 1999) provide statistics for the most common cyclic hydrogen-bonded motifs, and it is likely that motif information will be included in the developing IsoStar knowledge-based library described in Section 22.4.5.8.

22.4.5.7. *The answer 'no'*

Previous sections have illustrated the location and characterization of some important non-covalent interactions. Equally important is a knowledge of when such interactions *do not* occur although chemical sensibility might indicate that they should. We provide four examples from the CSD: (*a*) only 4.8% of more than 1000 thioether S atoms form hydrogen-atom contacts that are within van der Waals limits, despite the obvious analogy with the potent

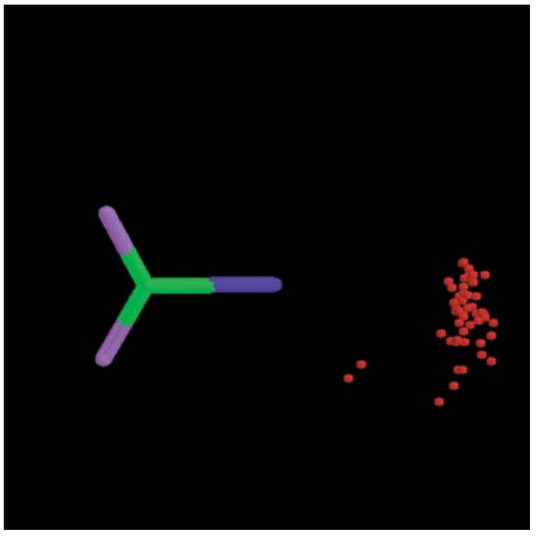


Fig. 22.4.5.3. Distribution of oxygen atoms around C(aromatic)—I (CSD data from the IsoStar library, see text).

acceptor C—O—C (Allen, Bird *et al.*, 1997*b*); (*b*) of 118 instances in which a furan ring coexists with N—H or O—H donors, the O atom forms hydrogen bonds on only three occasions (Nobeli *et al.*, 1997); (*c*) the ester oxygen $(R_1)(O{=})C{-}\underline{O}{-}R_2$ almost never forms strong hydrogen bonds, although the adjunct carbonyl oxygen atom is well known as a highly potent acceptor (Lommerse *et al.*, 1997); and (*d*) covalently bound fluorine atoms rarely form hydrogen bonds (Dunitz & Taylor, 1997).

22.4.5.8. *IsoStar: a library of non-bonded interactions*

The previous sections show that the amount of data in the CSD on intermolecular geometries is vast, and CSD-derived information for a number of specific systems is available in the literature at various levels of detail. If not, the CSD must be searched for contacts between the relevant functional groups. To provide structured and direct access to a more comprehensive set of derived information, a knowledge-based library of non-bonded interactions (IsoStar: Bruno *et al.*, 1997) has been developed at the CCDC since 1995. IsoStar is based on experimental data, not only from the CSD but also from the PDB, and contains some theoretical results calculated using the IMPT method. Version 1.1 of IsoStar, released in October 1998, contains information on non-bonded interactions formed between 310 common functional groups, referred to as *central groups*, and 45 *contact groups*, *e.g.* hydrogen-bond donors, water, halide ions *etc.* Information is displayed in the form of scatter plots for each interaction. Version 1.1 contains about 12000 scatter plots: 9000 from the CSD and 3000 from the PDB. IsoStar also reports results for 867 theoretical potential-energy minima.

For a given contact between between a central group (*A*) and a contact group (*B*), CSD search results were transformed into an easily visualized form by overlaying the *A* moieties. This results in a 3D distribution (scatter plot) showing the experimental distribution of *B* around *A*. Fig. 22.4.5.1(*a*) shows an example of a scatter plot: the distribution of OH groups around carboxylate anions, illustrating hydrogen-bond formation along the lone-pair directions of the carboxylate oxygens. The IsoStar software provides a tool that enables the user to inspect quickly the original crystal structures in which the contacts occur *via* a hyperlink to the original CSD entries. This is very helpful in identifying outliers, motifs and biases. Another tool generates contoured surfaces from scatter plots, which show the density distribution of the contact groups. A similar approach was first used by Rosenfield *et al.* (1984). Contouring aids the interpretation of the scatter plot and the analysis of preferred geometries. Fig. 22.4.5.1(*b*) shows the contoured surface of the scatter plot in Fig. 22.4.5.1(*a*); the lone-pair directionality now becomes even more obvious.

The fact that carboxylate anions form hydrogen bonds along their lone-pair directions may be well known, although force fields do not always use this information. However, the IsoStar library also contains information on many less well understood functional groups. The interaction between aromatic halo groups and oxygen atoms (Lommerse *et al.*, 1996) is referred to above, and Fig. 22.4.5.3 shows the distribution of oxygen acceptor atoms around aromatic iodine groups. It is clear that the contact O atoms are preferentially observed along the elongation of the C—I bond.

The PDB scatter plots in IsoStar only involve interactions between non-covalently bound ligands and proteins, *i.e* side chain–side chain interactions are excluded. Similar work was presented by Tintelnot & Andrews (1989), but at that time the PDB contained only 40 structures of protein–ligand complexes. The IsoStar library contains data derived from almost 800 complexes having a resolution better than 2.5 Å. Fig. 22.4.5.1(*c*) shows an example of a scatter plot from the PDB (the distribution of OH groups around carboxylate groups). Here, although the hydrogen atoms are missing in the PDB plot, the close similarity between Figs. 22.4.5.1(*c*) (PDB) and 22.4.5.1(*a*) (CSD) is obvious.

22.4.5.9. *Protein–ligand binding*

The reluctance to use data from the CSD because they do not relate directly to biological systems has been noted earlier. However, in principle, the same forces that drive the inclusion of a new molecule into a growing crystal should also apply to the binding of a ligand to a protein. In both cases, molecule and target need to be de-solvated first (although in the first case not necessarily from a water environment) and then interact in the most favourable way.

Nicklaus and colleagues suggested that on average, the conformational energy of ligands in the protein-bound state is 66 (48) kJ mol^{-1} above that of the global minimum-energy conformation *in vacuo* (Nicklaus *et al.*, 1995). This result was based on 33 protein–ligand complexes from the PDB for which the ligand also occurs in a small-molecule structure in the CSD. The same investigation also showed that, although ligand conformations in the protein-bound state are generally different from those observed in small-molecule crystal structures, on average the conformational energy of the ligand in the CSD crystal-structure conformation is 66 (47) kJ mol^{-1} above that of the global minimum-energy conformation *in vacuo*, although Boström *et al.* (1998) have shown that these conformational energies are much lower if calculated in a water environment. The computational work indicates that the forces that affect the conformation of a ligand are of comparable magnitude at a protein binding site to those in a small-molecule crystal-structure environment. Thus, if small-molecule crystal-structure statistics tell us that a given structure fragment can only adopt one conformation, generally there is no reason to believe that a ligand that contains this fragment will adopt a different conformation when it binds to a protein.

In principle, the information on non-bonded interactions derived from the CSD and assembled in the IsoStar library should be very important for the understanding and prediction of interaction geometries. However, in light of the comments above, it is important to know whether these data *are* generally relevant to interactions that occur in the protein binding site. Work by Klebe (1994) indicated that, at least for a limited set of test cases, the geometrical distributions derived from ligand–protein complexes are similar to those derived from small-molecule crystal structures. Since the IsoStar library contains information from both the PDB and the CSD, it provides the ultimate basis for establishing similarities (or not) between the interaction geometries observed in small-molecule crystal structures and those observed in protein–ligand complexes. Comparing CSD scatter plots with their corresponding plots from the PDB is an obvious way of establishing the relevance of non-bonded interaction data from small-molecule crystal structures to biological systems.

A full systematic comparison of PDB and CSD scatter plots or, more accurately, of PDB and CSD *density maps* has recently been performed by Verdonk (1998). He calculated residual densities, obtained by subtracting one density map from the other, for each pair of density maps. It appears that, in general, CSD and PDB plots (and thus interaction geometries) are very similar indeed: the average residual density is only 10 (10)%, indicating that 90% of the density in the PDB map is also observed in the CSD map. In Fig. 22.4.5.4(*a*), the average residual densities of each PDB–CSD comparison are plotted *versus* the average concentration of contact groups in the scatter plot. The filled circles represent comparisons for which the protonation state of the central group is unambiguous (*i.e.* carboxylic acid, imidazole *etc.* were excluded). It appears that the residual density decreases with the amount of data in the plots,

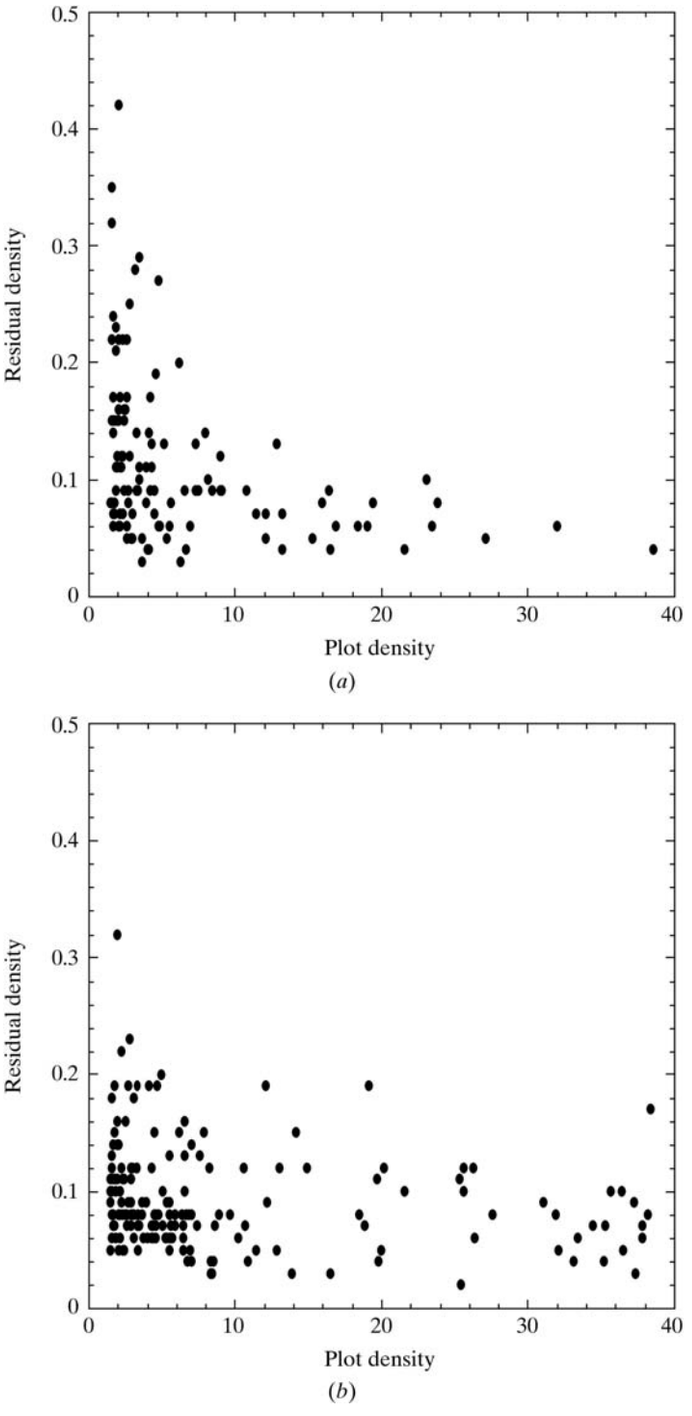


Fig. 22.4.5.4. Pairwise comparison of intermolecular-interaction density maps from the CSD and the PDB. Plots of *residual density* $|\rho(\text{CSD}) - \rho(\text{PDB})|$ *versus plot density*, *i.e.* the average density in the least dense situation (CSD or PDB), for situations where the protonation state of the central group is (*a*) unambiguous, and (*b*) ambiguous.

obviously caused by the more accurate calculation of the residual density. The 'true' residual density seems to be as low as about 6%.

Fig. 22.4.5.4(*b*) shows a similar graph, but now for those density maps in which the protonation state of the central group is ambiguous. As expected, the spread in the calculated residual densities is much higher, even for very dense plots. By comparing the density map from the PDB with the CSD maps for the different protonation states of the central group, the most frequent protonation state of this central group in the protein structures can be predicted. In Table 22.4.5.1, for example, the residual densities for protein carboxylic acid groups are shown, compared with the CSD plots of the neutral carboxylic acid and with those of the charged carboxylate anion. In all cases, the residual density is lower if the PDB map is compared with the CSD map for charged carboxylate anions. This indicates that the majority of glutamate and aspartate side chains are charged, which is consistent with other evidence.

Table 22.4.5.1. *Residual densities for carboxylic acid groups*

The PDB density maps are compared with the CSD maps for uncharged carboxylic acid and for charged carboxylate anions.

	Residual density (CCO_2H)	Residual density ($CCOO^-$)
Any (N,O,S)—H	0.06	0.04
Any N—H nitrogen	0.07	0.05
Any O—H oxygen	0.07	0.05
Non-donating oxygen	0.12	0.04
Carbonyl oxygen	0.13	0.07
Carbonyl carbon	0.12	0.04
Water oxygen	0.07	0.05
Any aliphatic C—H carbon	0.08	0.06

22.4.5.10. *Modelling applications that use CSD data*

Predicting binding modes of ligands at protein binding sites is a problem of paramount importance in drug design. One approach to this problem is to attempt to dock the ligand directly into the binding site. There are several protein–ligand docking programs available, *e.g. DOCK* (see Kuntz *et al.*, 1994), *GRID* (Goodford, 1985), *FLExX* and *FLExS* (Rarey *et al.*, 1996; Lemmen & Lengauer, 1997), and *GOLD* (Jones *et al.*, 1995, 1997). The docking program *GOLD*, developed by the University of Sheffield, Glaxo Wellcome and the CCDC, and which has the high docking success rate of 73%, uses a small torsion library, based on the data from the CSD, to explore the conformational space of the ligand. Its hydrogen-bond geometries and fitness functions are also partly based on CSD data. In the future, we intend to create a more direct link between the crystallographic data and the docking program, *via* IsoStar and the developing torsion library.

Another approach to the prediction of binding modes is to calculate the energy fields for different probes at each position of the binding site, for instance using the *GRID* program (Goodford, 1985). The resulting maps can be displayed as contoured surfaces which can assist in the prediction and understanding of binding modes of ligands. CCDC is developing a program called *SuperStar* (Verdonk *et al.*, 1999) which uses a similar approach to that of the *X-SITE* program (Singh *et al.*, 1991; Laskowski *et al.*, 1996). However, *SuperStar* uses non-bonded interaction data from the CSD rather than the protein side chain–side chain interaction data employed in *X-SITE*. Thus, for a given binding site and contact group (probe), *SuperStar* selects the appropriate scatter plots from the IsoStar library, superimposes the scatter plots on the relevant functional groups in the binding site, and transforms them into one composite probability map. Such maps can then, for example, be used to predict where certain functional groups are likely to interact with the binding site. The strength of *SuperStar* is that it is based entirely on experimental data (although this is also the cause of some limitations). The fields simply represent what has been observed in crystal strucures. We are currently verifying *SuperStar* on a test set of more than 100 protein–ligand complexes from the PDB and preliminary results are encouraging.

Finally, CSD data are used in several *de novo* design programs. These types of programs, *e.g. LUDI* (Böhm, 1992*a*,*b*), predict novel ligands that will interact favourably with a given protein and use hydrogen-bond geometries from the CSD (indirectly) to position their structural fragments in the binding site.

22.4.6. Conclusion

This chapter has summarized the vast range of structural knowledge that can be derived from the small-molecule data contained in the CSD. We have attempted to show that much of this knowledge is directly transferable and applicable to the protein environment. Far from being discrete, structural studies of small molecules and proteins have a natural synergy which, if exploited creatively, will lead to significant advances in both areas. It is therefore unsurprising that some of these CSD studies have been prompted by initial observations made on proteins.

As a result of this activity, it is now very clear that software access to the information stored in the CSD and the PDB must be at two levels: a raw-data level and a derived-knowledge level. The onward development of structural knowledge bases from the underlying data provides for the preservation and storage of the results of data-mining experiments, thus avoiding repetition of standard experiments and providing instant access to complex derivative information. Most importantly, a suitably structured knowledge base can be acted on by software tools that are designed to solve complex problems in structural chemistry (see *e.g.* Thornton & Gardner, 1989; Allen *et al.*, 1990; Bruno *et al.*, 1997; Jones *et al.*, 1997). The availability of knowledge bases derived from experimental observations is likely to be a crucial factor in the solution of those two analogous, and currently intractable, problems in the small-molecule and protein-structure domains: crystal structure and polymorph prediction on the one hand, and protein folding on the other.

References

22.1

Acharya, R., Fry, E., Logan, D., Stuart, D., Brown, F., Fox, G. & Rowlands, D. (1990). *The three-dimensional structure of foot-and-mouth disease virus. New aspects of positive-strand RNA viruses*, edited by M. A. Brinton & S. X. Heinz, pp. 319–327. Washington DC: American Society for Microbiology.

Arnold, E. & Rossmann, M. G. (1990). *Analysis of the structure of a common cold virus, human rhinovirus 14, refined at a resolution of 3.0 Å. J. Mol. Biol.* **211**, 763–801.

Baker, E. N. & Hubbard, R. E. (1984). *Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol.* **44**, 97–179.

Bernal, J. D. & Finney, J. L. (1967). *Random close-packed hard-sphere model II. Geometry of random packing of hard spheres. Discuss. Faraday Soc.* **43**, 62–69.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Structure of hen egg-white lysozyme, a three-dimensional Fourier synthesis at 2 Å resolution. Nature (London)*, **206**, 757–761.

Bondi, A. (1964). *van der Waals volumes and radii. J. Phys. Chem.* **68**, 441–451.

Bondi, A. (1968). *Molecular crystals, liquids and glasses*. New York: Wiley.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem.* **4**, 187–217.

Chandler, D., Weeks, J. D. & Andersen, H. C. (1983). *van der Waals picture of liquids, solids, and phase transformations. Science*, **220**, 787–794.

Chapman, M. S. (1993). *Mapping the surface properties of macromolecules. Protein Sci.* **2**, 459–469.

Chapman, M. S. (1994). *Sequence similarity scores and the inference of structure/function relationships. Comput. Appl. Biosci. (CABIOS)*, **10**, 111–119.

Chothia, C. (1975). *Structural invariants in protein folding. Nature (London)*, **254**, 304–308.

Chothia, C. (1976). *The nature of the accessible and buried surfaces in proteins. J. Mol. Biol.* **105**, 1–12.

Chothia, C. & Janin, J. (1975). *Principles of protein–protein recognition. Nature (London)*, **256**, 705–708.

Connolly, M. (1986). *Measurement of protein surface shape by solid angles. J. Mol. Graphics*, **4**, 3–6.

Connolly, M. L. (1983). *Analytical molecular surface calculation. J. Appl. Cryst.* **16**, 548–558.

Connolly, M. L. (1991). *Molecular interstitial skeleton. Comput. Chem.* **15**, 37–45.

Diamond, R. (1974). *Real-space refinement of the structure of hen egg-white lysozyme. J. Mol. Biol.* **82**, 371–391.

Dunfield, L. G., Burgess, A. W. & Scheraga, H. A. (1979). *J. Phys. Chem.* **82**, 2609.

Edelsbrunner, H., Facello, M. & Liang, J. (1996). *On the definition and construction of pockets in macromolecules*, pp. 272–287. Singapore: World Scientific.

Edelsbrunner, H., Facello, M., Ping, F. & Jie, L. (1995). *Measuring proteins and voids in proteins. Proc. 28th Hawaii Intl Conf. Sys. Sci.* pp. 256–264.

Edelsbrunner, H. & Mucke, E. (1994). *Three-dimensional alpha shapes. ACM Trans. Graphics*, **13**, 43–72.

Eisenberg, D. & McLachlan, A. D. (1986). *Solvation energy in protein folding and binding. Nature (London)*, **319**, 199–203.

Fauchere, J.-L. & Pliska, V. (1983). *Hydrophobic parameters π of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. Eur. J. Med. Chem. Chim. Ther.* **18**, 369–375.

Finkelstein, A. (1994). *Implications of the random characteristics of protein sequences for their three-dimensional structure. Curr. Opin. Struct. Biol.* **4**, 422–428.

Finney, J. L. (1975). *Volume occupation, environment and accessibility in proteins. The problem of the protein surface. J. Mol. Biol.* **96**, 721–732.

Finney, J. L., Gellatly, B. J., Golton, I. C. & Goodfellow, J. (1980). *Solvent effects and polar interactions in the structural stability and dynamics of globular proteins. Biophys. J.* **32**, 17–33.

Fritz-Wolf, K., Schnyder, T., Wallimann, T. & Kabsch, W. (1996). *Structure of mitochondrial creatine kinase. Nature (London)*, **381**, 341–345.

Gelin, B. R. & Karplus, M. (1979). *Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. Biochemistry*, **18**, 1256–1268.

Gellatly, B. J. & Finney, J. L. (1982). *Calculation of protein volumes: an alternative to the Voronoi procedure. J. Mol. Biol.* **161**, 305–322.

Gerstein, M. (1992). *A resolution-sensitive procedure for comparing surfaces and its application to the comparison of antigen-combining sites. Acta Cryst.* A**48**, 271–276.

Gerstein, M. & Chothia, C. (1996). *Packing at the protein–water interface. Proc. Natl Acad. Sci. USA*, **93**, 10167–10172.

Gerstein, M., Lesk, A. M., Baker, E. N., Anderson, B., Norris, G. & Chothia, C. (1993). *Domain closure in lactoferrin: two hinges produce a see-saw motion between alternative close-packed interfaces. J. Mol. Biol.* **234**, 357–372.

Gerstein, M., Lesk, A. M. & Chothia, C. (1994). *Structural mechanisms for domain movements. Biochemistry*, **33**, 6739–6749.

Gerstein, M. & Lynden-Bell, R. M. (1993*a*). *Simulation of water around a model protein helix. 1. Two-dimensional projections of solvent structure. J. Phys. Chem.* **97**, 2982–2991.

22.1 (*cont.*)

Gerstein, M. & Lynden-Bell, R. M. (1993*b*). *Simulation of water around a model protein helix. 2. The relative contributions of packing, hydrophobicity, and hydrogen bonding*. *J. Phys. Chem.* **97**, 2991–2999.

Gerstein, M. & Lynden-Bell, R. M. (1993*c*). *What is the natural boundary for a protein in solution? J. Mol. Biol.* **230**, 641–650.

Gerstein, M., Sonnhammer, E. & Chothia, C. (1994). *Volume changes on protein evolution*. *J. Mol. Biol.* **236**, 1067–1078.

Gerstein, M., Tsai, J. & Levitt, M. (1995). *The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra*. *J. Mol. Biol.* **249**, 955–966.

Grant, J. A. & Pickup, B. T. (1995). *A Gaussian description of molecular shape. J. Phys. Chem.* **99**, 3503–3510.

Greer, J. & Bush, B. L. (1978). *Macromolecular shape and surface maps by solvent exclusion. Proc. Natl Acad. Sci. USA*, **75**, 303–307.

Harber, J., Bernhardt, G., Lu, H.-H., Sgro, J.-Y. & Wimmer, E. (1995). *Canyon rim residues, including antigenic determinants, modulate serotype-specific binding of polioviruses to mutants of the poliovirus receptor. Virology*, **214**, 559–570.

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Volume changes on protein folding*. *Structure*, **2**, 641–649.

Hermann, R. B. (1977). *Use of solvent cavity area and number of packed solvent molecules around a solute in regard to hydrocarbon solubilities and hydrophobic interactions*. *Proc. Natl Acad. Sci. USA*, **74**, 4144–4195.

Hubbard, S. J. & Argos, P. (1994). *Cavities and packing at protein interfaces*. *Protein Sci.* **3**, 2194–2206.

Hubbard, S. J. & Argos, P. (1995). *Evidence on close packing and cavities in proteins*. *Curr. Opin. Biotechnol.* **6**, 375–381.

Kapp, O. H., Moens, L., Vanfleteren, J., Trotman, C. N. A., Suzuki, T. & Vinogradov, S. N. (1995). *Alignment of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume*. *Protein Sci.* **4**, 2179–2190.

Kauzmann, W. (1959). *Some factors in the interpretation of protein denaturation. Adv. Protein Chem.* **14**, 1–63.

Kelly, J. A., Sielecki, A. R., Sykes, B. D., James, M. N. & Phillips, D. C. (1979). *X-ray crystallography of the binding of the bacterial cell wall trisaccharaide NAM-NAG-NAM to lysozymes. Nature (London)*, **282**, 875–878.

Kim, K. H., Willingmann, P., Gong, Z. X., Kremer, M. J., Chapman, M. S., Minor, I., Oliveira, M. A., Rossmann, M. G., Andries, K., Diana, G. D., Dutko, F. J., McKinlay, M. A. & Pevear, D. C. (1993). *A comparison of the anti-rhinoviral drug binding pocket in HRV14 and HRV1A. J. Mol. Biol.* **230**, 206–227.

Kleywegt, G. J. & Jones, T. A. (1994). *Detection, delineation, measurement and display of cavities in macromolecular structures*. *Acta Cryst.* D**50**, 178–185.

Kocher, J. P., Prevost, M., Wodak, S. J. & Lee, B. (1996). *Properties of the protein matrix revealed by the free energy of cavity formation*. *Structure*, **4**, 1517–1529.

Kraulis, P. J. (1991). *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst.* **24**, 946–950.

Kuhn, L. A., Siani, M. A., Pique, M. E., Fisher, C. L., Getzoff, E. D. & Tainer, J. A. (1992). *The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures*. *J. Mol. Biol.* **228**, 13–22.

Lee, B. & Richards, F. M. (1971). *The interpretation of protein structures: estimation of static accessibility*. *J. Mol. Biol.* **55**, 379–400.

Leicester, S. E., Finney, J. L. & Bywater, R. P. (1988). *Description of molecular surface shape using Fourier descriptors*. *J. Mol. Graphics*, **6**, 104–108.

Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). *Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution*. *Comput. Phys. Comm.* **91**, 215–231.

Lewis, M. & Rees, D. C. (1985). *Fractal surfaces of proteins*. *Science*, **230**, 1163–1165.

Lim, V. I. & Ptitsyn, O. B. (1970). *On the constancy of the hydrophobic nucleus volume in molecules of myoglobins and hemoglobins*. *Mol. Biol. (USSR)*, **4**, 372–382.

Madan, B. & Lee, B. (1994). *Role of hydrogen bonds in hydrophobicity: the free energy of cavity formation in water models with and without the hydrogen bonds*. *Biophys. Chem.* **51**, 279–289.

Matthews, B. W., Morton, A. G. & Dahlquist, F. W. (1995). *Use of NMR to detect water within nonpolar protein cavities*. (Letter.) *Science*, **270**, 1847–1849.

Merritt, E. A. & Bacon, D. J. (1997). *Raster3D: photorealistic molecular graphics. Methods Enzymol.* **277**, 505–525.

Molecular Structure Corporation (1995). *Insight II user guide*. Biosym/MSI, San Diego.

Nemethy, G., Pottle, M. S. & Scheraga, H. A. (1983). *Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions and hydrogen bond interactions for the naturally occurring amino acids*. *J. Phys. Chem.* **87**, 1883–1887.

Nicholls, A. (1992). *GRASP: graphical representation and analysis of surface properties.* New York: Columbia University.

Nicholls, A. & Honig, B. (1991). *A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. J. Comput. Chem.* **12**, 435–445.

Nicholls, A., Sharp, K. & Honig, B. (1991). *Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins*, **11**, 281–296.

Olson, N., Kolatkar, P., Oliveira, M. A., Cheng, R. H., Greve, J. M., McClelland, A., Baker, T. S. & Rossmann, M. G. (1993). *Structure of a human rhinovirus complexed with its receptor molecule. Proc. Natl Acad. Sci. USA*, **90**, 507–511.

O'Rourke, J. (1994). *Computational geometry in C*. Cambridge University Press.

Palmenberg, A. C. (1989). *Sequence alignments of picornaviral capsid proteins*. In *Molecular aspects of picornavirus infection and detection*, edited by B. L. Semler & E. Ehrenfeld, pp. 211–241. Washington DC: American Society for Microbiology.

Pattabiraman, N., Ward, K. B. & Fleming, P. J. (1995). *Occluded molecular surface: analysis of protein packing*. *J. Mol. Recognit.* **8**, 334–344.

Pauling, L. (1960). *The nature of the chemical bond*, 3rd ed. Ithaca: Cornell University Press.

Peters, K. P., Fauck, J. & Frommel, C. (1996). *The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria*. *J. Mol. Biol.* **256**, 201–213.

Petitjean, M. (1994). *On the analytical calculation of van der Waals surfaces and volumes: some numerical aspects*. *J. Comput. Chem.* **15**, 1–10.

Pontius, J., Richelle, J. & Wodak, S. J. (1996). *Deviations from standard atomic volumes as a quality measure for protein crystal structures*. *J. Mol. Biol.* **264**, 121–136.

Procacci, P. & Scateni, R. (1992). *A general algorithm for computing Voronoi volumes: application to the hydrated crystal of myoglobin*. *Int. J. Quant. Chem.* **42**, 151–152.

Rashin, A. A., Iofin, M. & Honig, B. (1986). *Internal cavities and buried waters in globular proteins*. *Biochemistry*, **25**, 3619–3625.

Reynolds, J. A., Gilbert, D. B. & Tanford, C. (1974). *Empirical correlation between hydrophobic free energy and aqueous cavity surface area*. *Proc. Natl Acad. Sci. USA*, **71**, 2925–2927.

Richards, F. M. (1974). *The interpretation of protein structures: total volume, group volume distributions and packing density*. *J. Mol. Biol.* **82**, 1–14.

Richards, F. M. (1977). *Areas, volumes, packing, and protein structure*. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Richards, F. M. (1979). *Packing defects, cavities, volume fluctuations, and access to the interior of proteins. Including some general comments on surface area and protein structure*. *Carlsberg Res. Commun.* **44**, 47–63.

22.1 (*cont.*)

Richards, F. M. (1985). *Calculation of molecular volumes and areas for structures of known geometry*. *Methods Enzymol.* **115**, 440–464.

Richards, F. M. & Lim, W. A. (1994). *An analysis of packing in the protein folding problem*. *Q. Rev. Biophys.* **26**, 423–498.

Richmond, T. J. (1984). *Solvent accessible surface area and excluded volume in proteins: analytical equations for overlapping spheres and implications for the hydrophobic effect. J. Mol. Biol.* **178**, 63–89.

Richmond, T. J. & Richards, F. M. (1978). *Packing of alpha-helices: geometrical constraints and contact areas*. *J. Mol. Biol.* **119**, 537–555.

Rossmann, M. G. (1989). *The canyon hypothesis. J. Biol. Chem.* **264**, 14587–14590.

Rossmann, M. G. & Palmenberg, A. C. (1988). *Conservation of the putative receptor attachment site in picornaviruses. Virology*, **164**, 373–382.

Rowland, R. S. & Taylor, R. (1996). *Intermolecular nonbonded contact distances in organic crystal structures: comparison with distances expected from van der Waals radii*. *J. Phys. Chem.* **100**, 7384–7391.

Sgro, J.-Y. (1996). *Virus visualization*. In *Encyclopedia of virology plus* (CD-ROM version), edited by R. G. Webster & A. Granoff. San Diego: Academic Press.

Sharp, K. A., Nicholls, A., Fine, R. F. & Honig, B. (1991). *Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. Science*, **252**, 107–109.

Sherry, B., Mosser, A. G., Colonno, R. J. & Rueckert, R. R. (1986). *Use of monoclonal antibodies to identify four neutralization immunogens on a common cold picornavirus, human rhinovirus 14. J. Virol.* **57**, 246–257.

Sherry, B. & Rueckert, R. (1985). *Evidence for at least two dominant neutralization antigens on human rhinovirus 14. J. Virol.* **53**, 137–143.

Shrake, A. & Rupley, J. A. (1973). *Environment and exposure to solvent of protein atoms. Lysozyme and insulin*. *J. Mol. Biol.* **79**, 351–371.

Sibbald, P. R. & Argos, P. (1990). *Weighting aligned protein or nucleic acid sequences to correct for unequal representation*. *J. Mol. Biol.* **216**, 813–818.

Singh, R. K., Tropsha, A. & Vaisman, I. I. (1996). *Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues*. *J. Comput. Biol.* **3**, 213–222.

Sreenivasan, U. & Axelsen, P. H. (1992). *Buried water in homologous serine proteases*. *Biochemistry*, **31**, 12785–12791.

Tanford, C. (1997). *How protein chemists learned about the hydrophobicity factor. Protein Sci.* **6**, 1358–1366.

Tanford, C. H. (1979). *Interfacial free energy and the hydrophobic effect*. *Proc. Natl Acad. Sci. USA*, **76**, 4175–4176.

Ten Eyck, L. F. (1977). *Efficient structure-factor calculation for large molecules by the fast Fourier transform. Acta Cryst.* A**33**, 486–492.

Tsai, J., Gerstein, M. & Levitt, M. (1996). *Keeping the shape but changing the charges: a simulation study of urea and its isosteric analogues*. *J. Chem. Phys.* **104**, 9417–9430.

Tsai, J., Gerstein, M. & Levitt, M. (1997). *Estimating the size of the minimal hydrophobic core*. *Protein Sci.* **6**, 2606–2616.

Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). *The packing density in proteins: standard radii and volumes*. *J. Mol. Biol.* **290**, 253–266.

Tsai, J., Voss, N. & Gerstein, M. (2001). *Voronoi calculations of protein volumes: sensitivity analysis and parameter database*. *Bioinformatics*. In the press.

Voronoi, G. F. (1908). *Nouvelles applications des paramétres continus à la théorie des formes quadratiques*. *J. Reine Angew. Math.* **134**, 198–287.

Williams, M. A., Goodfellow, J. M. & Thornton, J. M. (1994). *Buried waters and internal cavities in monomeric proteins*. *Protein Sci.* **3**, 1224–1235.

Wodak, S. J. & Janin, J. (1980). *Analytical approximation to the accessible surface areas of proteins. Proc. Natl Acad. Sci. USA*, **77**, 1736–1740.

Xie, Q. & Chapman, M. S. (1996). *Canine parvovirus capsid structure, analyzed at 2.9 Å resolution. J. Mol. Biol.* **264**, 497–520.

Zhou, G., Somasundaram, T., Blanc, E., Parthasarathy, G., Ellington, W. R. & Chapman, M. S. (1998). *Transition state structure of arginine kinase: implications for catalysis of bimolecular reactions. Proc. Natl Acad. Sci. USA*, **95**, 8449–8454.

22.2

Adman, E., Watenpaugh, K. D. & Jensen, L. H. (1975). *N—H···S hydrogen bonds in Peptococcus aerogenes ferredoxin, Clostridium pasteurianum rubredoxin and Chromatium high potential iron protein*. *Proc. Natl Acad. Sci. USA*, **72**, 4854–4858.

Alber, T., Dao-pin, S., Wilson, K., Wozniak, J. A., Cook, S. P. & Matthews, B. W. (1987). *Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. Nature (London)*, **330**, 41–46.

Artymiuk, P. J. & Blake, C. C. F. (1981). *Refinement of human lysozyme at 1.5 Å resolution. Analysis of non-bonded and hydrogen-bonded interactions. J. Mol. Biol.* **152**, 737–762.

Baker, E. N. (1995). *Solvent interactions with proteins as revealed by X-ray crystallographic studies.* In *Protein–solvent interactions*, edited by R. B. Gregory, pp. 143–189. New York: Marcel Dekker Inc.

Baker, E. N. & Hubbard, R. E. (1984). *Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol.* **44**, 97–179.

Blundell, T., Barlow, D., Borkakoti, N. & Thornton, J. (1983). *Solvent-induced distortions and the curvature of α-helices. Nature (London)*, **306**, 281–283.

Bordo, D. & Argos, P. (1994). *The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins. J. Mol. Biol.* **243**, 504–519.

Burley, S. K. & Petsko, G. A. (1986). *Amino–aromatic interactions in proteins. FEBS Lett.* **203**, 139–143.

Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R. & Doudna, J. A. (1996). *Crystal structure of a group I ribozyme domain: principles of RNA packing. Science*, **273**, 1678–1685.

Derewenda, Z. S., Derewenda, U. & Kobos, P. (1994). *(His)Cε—H···O=C< hydrogen bond in the active site of serine hydrolases. J. Mol. Biol.* **241**, 83–93.

Derewenda, Z. S., Lee, L. & Derewenda, U. (1995). *The occurrence of C—H···O hydrogen bonds in proteins. J. Mol. Biol.* **252**, 248–262.

Edwards, R. A., Baker, H. M., Whittaker, M. M., Whittaker, J. W. & Baker, E. N. (1998). *Crystal structure of Escherichia coli manganese superoxide dismutase at 2.1 Å resolution. J. Biol. Inorg. Chem.* **3**, 161–171.

Fersht, A. R. & Serrano, L. (1993). *Principles in protein stability derived from protein engineering experiments. Curr. Opin. Struct. Biol.* **3**, 75–83.

Fersht, A. R., Shi, J.-P., Knill-Jones, J., Lowe, D. M., Wilkinson, A. J., Blow, D. M., Brick, P., Carter, P., Waye, M. M. Y. & Winter, G. (1985). *Hydrogen bonding and biological specificity analysed by protein engineering. Nature (London)*, **314**, 235–238.

Flocco, M. M. & Mowbray, S. L. (1995). *Strange bedfellows: interactions between acidic side-chains in proteins. J. Mol. Biol.* **254**, 96–105.

Gregoret, L. M., Rader, S. D., Fletterick, R. J. & Cohen, F. E. (1991). *Hydrogen bonds involving sulfur atoms in proteins. Proteins Struct. Funct. Genet.* **9**, 99–107.

Hagler, A. T., Huler, E. & Lifson, S. (1974). *Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. J. Am. Chem. Soc.* **96**, 5319–5327.

Harper, E. T. & Rose, G. D. (1993). *Helix stop signals in proteins and peptides: the capping box. Biochemistry*, **32**, 7605–7609.

Huggins, M. L. (1971). *50 years of hydrogen bonding theory. Angew. Chem. Int. Ed. Engl.* **10**, 147–208.

22.2 (*cont.*)

Ippolito, J. A., Alexander, R. S. & Christianson, D. W. (1990). *Hydrogen bond stereochemistry in protein structure and function. J. Mol. Biol.* **215**, 457–471.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen bonding in biological structures*. New York: Springer-Verlag.

Kauzmann, W. (1959). *Some factors in the interpretation of protein denaturation. Adv. Protein Chem.* **14**, 1–64.

Legon, A. C. & Millen, D. J. (1987). *Directional character, strength, and nature of the hydrogen bond in gas-phase dimers. Acc. Chem. Res.* **20**, 39–45.

Levitt, M. & Perutz, M. F. (1988). *Aromatic rings act as hydrogen bond acceptors. J. Mol. Biol.* **201**, 751–754.

McDonald, I. K. & Thornton, J. M. (1994*a*). *Satisfying hydrogen bonding potential in proteins. J. Mol. Biol.* **238**, 777–793.

McDonald, I. K. & Thornton, J. M. (1994*b*). *The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. Protein Eng.* **8**, 217–224.

Matthews, B. W. (1972). *The γ turn. Evidence for a new folded conformation in proteins. Macromolecules*, **5**, 818–819.

Mitchell, J. B. O., Nandi, C. L., McDonald, I. K., Thornton, J. M. & Price, S. L. (1994). *Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding? J. Mol. Biol.* **239**, 315–331.

Nemethy, G. & Printz, M. P. (1972). *The γ turn, a possible folded conformation of the polypeptide chain. Comparison with the β turn. Macromolecules*, **5**, 755–758.

Pauling, L. (1960). *The nature of the chemical bond*, 3rd ed. Ithaca: Cornell University Press.

Pauling, L. & Corey, R. B. (1951). *Configurations of polypeptide chains with favoured orientations around single bonds: two new pleated sheets. Proc. Natl Acad. Sci. USA*, **37**, 729–740.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). *The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl Acad. Sci. USA*, **37**, 205–211.

Pley, H. W., Flaherty, K. M. & McKay, D. B. (1994). *Three-dimensional structure of a hammerhead ribozyme. Nature (London)*, **372**, 68–74.

Presta, L. G. & Rose, G. D. (1988). *Helix signals in proteins. Science*, **240**, 1632–1641.

Richardson, J. S., Getzoff, E. D. & Richardson, D. C. (1978). *The β-bulge: a common small unit of nonrepetitive protein structure. Proc. Natl Acad. Sci. USA*, **75**, 2574–2578.

Richardson, J. S. & Richardson, D. C. (1988). *Amino acid preferences for specific locations at the ends of α-helices. Science*, **240**, 1648–1652.

Savage, H. J., Elliott, C. J., Freeman, C. M. & Finney, J. L. (1993). *Lost hydrogen bonds and buried surface area: rationalising stability in globular proteins. J. Chem. Soc. Faraday Trans.* **89**, 2609–2617.

Schellman, C. (1980). *The alpha-L conformation at the ends of helices.* In *Protein folding*, edited by R. Jaenicke, pp. 53–61. Amsterdam: Elsevier.

Stickle, D. F., Presta, L. G., Dill, K. A. & Rose, G. D. (1992). *Hydrogen bonding in globular proteins. J. Mol. Biol.* **226**, 1143–1159.

Sutor, D. J. (1962). *The C—H···O hydrogen bond in crystals. Nature (London)*, **195**, 68–69.

Taylor, R., Kennard, O. & Versichel, W. (1983). *Geometry of the N—H···O=C hydrogen bond. 1. Lone pair directionality. J. Am. Chem. Soc.* **105**, 5761–5766.

Thanki, N., Thornton, J. M. & Goodfellow, J. M. (1988). *Distribution of water around amino acids in proteins. J. Mol. Biol.* **202**, 637–657.

Thanki, N., Umrania, Y., Thornton, J. M. & Goodfellow, J. M. (1991). *Analysis of protein main-chain solvation as a function of secondary structure. J. Mol. Biol.* **221**, 669–691.

Wahl, M. C. & Sundaralingam, M. (1997). *C—H···O hydrogen bonding in biology. Trends Biochem. Sci.* **22**, 97–102.

Williams, M. A., Goodfellow, J. M. & Thornton, J. M. (1994). *Buried waters and internal cavities in monomeric proteins. Protein Sci.* **3**, 1224–1235.

22.3

Antosiewicz, J., McCammon, J. A. & Gilson, M. K. (1994). *Prediction of pH-dependent properties of proteins. J. Mol. Biol.* **238**, 415–436.

Åqvist, J., Luecke, H., Quiocho, F. A. & Warshel, A. (1991). *Dipoles localized at helix termini of proteins stabilize charges. Proc. Natl Acad. Sci. USA*, **88**, 2026–2030.

Bacquet, R. & Rossky, P. (1984). *Ionic atmosphere of rodlike polyelectrolytes. A hypernetted chain study. J. Phys. Chem.* **88**, 2660.

Bashford, D. & Karplus, M. (1990). *pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry*, **29**, 10219–10225.

Beroza, P., Fredkin, D., Okamura, M. & Feher, G. (1991). *Protonation of interacting residues in a protein by a Monte-Carlo method. Proc. Natl Acad. Sci. USA*, **88**, 5804–5808.

Bharadwaj, R., Windemuth, A., Sridharan, S., Honig, B. & Nicholls, A. (1994). *The fast multipole boundary-element method for molecular electrostatics – an optimal approach for large systems. J. Comput. Chem.* **16**, 898–913.

Bruccoleri, R. E., Novotny, J., Sharp, K. A. & Davis, M. E. (1996). *Finite difference Poisson–Boltzmann electrostatic calculations: increased accuracy achieved by harmonic dielectric smoothing and charge anti-aliasing. J. Comput. Chem.* **18**, 268–276.

Davis, M. E. (1994). *The inducible multipole solvation model – a new model for solvation effects on solute electrostatics. J. Chem. Phys.* **100**, 5149–5159.

Davis, M. E. & McCammon, J. A. (1989). *Solving the finite difference linear Poisson Boltzmann equation: comparison of relaxational and conjugate gradient methods. J. Comput. Chem.* **10**, 386–395.

Gilson, M. (1993). *Multiple-site titration and molecular modeling: 2. Rapid methods for computing energies and forces for ionizable groups in proteins. Proteins Struct. Funct. Genet.* **15**, 266–282.

Gilson, M., Davis, M., Luty, B. & McCammon, J. (1993). *Computation of electrostatic forces on solvated molecules using the Poisson–Boltzmann equation. J. Phys. Chem.* **97**, 3591–3600.

Gilson, M. & Honig, B. (1986). *The dielectric constant of a folded protein. Biopolymers*, **25**, 2097–2119.

Gilson, M., McCammon, J. & Madura, J. (1995). *Molecular dynamics simulation with continuum solvent. J. Comput. Chem.* **16**, 1081–1095.

Gilson, M., Sharp, K. A. & Honig, B. (1988). *Calculating the electrostatic potential of molecules in solution: method and error assessment. J. Comput. Chem.* **9**, 327–335.

Holst, M., Kozack, R., Saied, F. & Subramaniam, S. (1994). *Protein electrostatics – rapid multigrid-based Newton algorithm for solution of the full nonlinear Poisson–Boltzmann equation. J. Biomol. Struct. Dyn.* **11**, 1437–1445.

Holst, M. & Saied, F. (1993). *Multigrid solution of the Poisson–Boltzmann equation. J. Comput. Chem.* **14**, 105–113.

Jayaram, B., Fine, R., Sharp, K. A. & Honig, B. (1989). *Free energy calculations of ion hydration: an analysis of the Born model in terms of microscopic simulations. J. Phys. Chem.* **93**, 4320–4327.

Jayaram, B., Sharp, K. A. & Honig, B. (1989). *The electrostatic potential of B-DNA. Biopolymers*, **28**, 975–993.

Jean-Charles, A., Nicholls, A., Sharp, K., Honig, B., Tempczyk, A., Hendrickson, T. & Still, C. (1990). *Electrostatic contributions to solvation energies: comparison of free energy perturbation and continuum calculations. J. Am. Chem. Soc.* **113**, 1454–1455.

Langsetmo, K., Fuchs, J. A., Woodward, C. & Sharp, K. A. (1991). *Linkage of thioredoxin stability to titration of ionizable groups with perturbed pK_a. Biochemistry*, **30**, 7609–7614.

Lee, F., Chu, Z. & Warshel, A. (1993). *Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the Polaris and Enzymix programs. J. Comput. Chem.* **14**, 161–185.

22.3 (*cont.*)

Lee, F. S., Chu, Z. T., Bolger, M. B. & Warshel, A. (1992). *Calculations of antibody–antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to Mcpc603. Protein Eng.* **5**, 215–228.

Misra, V., Hecht, J., Sharp, K., Friedman, R. & Honig, B. (1994). *Salt effects on protein–DNA interactions: the lambda cI repressor and Eco R1 endonuclease. J. Mol. Biol.* **238**, 264–280.

Misra, V. & Honig, B. (1995). *On the magnitude of the electrostatic contribution to ligand–DNA interactions. Proc. Natl Acad. Sci. USA*, **92**, 4691–4695.

Misra, V., Sharp, K., Friedman, R. & Honig, B. (1994). *Salt effects on ligand–DNA binding: minor groove antibiotics. J. Mol. Biol.* **238**, 245–263.

Mohan, V., Davis, M. E., McCammon, J. A. & Pettitt, B. M. (1992). *Continuum model calculations of solvation free energies – accurate evaluation of electrostatic contributions. J. Phys. Chem.* **96**, 6428–6431.

Murthy, C. S., Bacquet, R. J. & Rossky, P. J. (1985). *Ionic distributions near polyelectrolytes. A comparison of theoretical approaches. J. Phys. Chem.* **89**, 701.

Nakamura, H., Sakamoto, T. & Wada, A. (1988). *A theoretical study of the dielectric constant of a protein. Protein Eng.* **2**, 177–183.

Nicholls, A. & Honig, B. (1991). *A rapid finite difference algorithm utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. J. Comput. Chem.* **12**, 435–445.

Oberoi, H. & Allewell, N. (1993). *Multigrid solution of the nonlinear Poisson–Boltzmann equation and calculation of titration curves. Biophys. J.* **65**, 48–55.

Olmsted, M. C., Anderson, C. F. & Record, M. T. (1989). *Monte Carlo description of oligoelectrolyte properties of DNA oligomers. Proc. Natl Acad. Sci. USA*, **86**, 7766–7770.

Olmsted, M. C., Anderson, C. F. & Record, M. T. (1991). *Importance of oligoelectrolyte end effects for the thermodynamics of conformational transitions of nucleic acid oligomers. Biopolymers*, **31**, 1593–1604.

Pack, G., Garrett, G., Wong, L. & Lamm, G. (1993). *The effect of a variable dielectric coefficient and finite ion size on Poisson–Boltzmann calculations of DNA–electrolyte systems. Biophys. J.* **65**, 1363–1370.

Pack, G. R. & Klein, B. J. (1984). *The distribution of electrolyte ions around the B- and Z-conformers of DNA. Biopolymers*, **23**, 2801.

Pack, G. R., Wong, L. & Prasad, C. V. (1986). *Counterion distribution around DNA. Nucleic Acids Res.* **14**, 1479.

Rashin, A. A. (1990). *Hydration phenomena, classical electrostatics and the boundary element method. J. Phys. Chem.* **94**, 725–733.

Record, T., Olmsted, M. & Anderson, C. (1990). *Theoretical studies of the thermodynamics of ion interaction with DNA.* In *Theoretical biochemistry and molecular biophysics.* New York: Adenine Press.

Reiner, E. S. & Radke, C. J. (1990). *Variational approach to the electrostatic free energy in charged colloidal suspensions. J. Chem. Soc. Faraday Trans.* **86**, 3901.

Schaeffer, M. & Frommel, C. (1990). *A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. J. Mol. Biol.* **216**, 1045–1066.

Sharp, K. & Honig, B. (1990). *Calculating total electrostatic energies with the non-linear Poisson–Boltzmann equation. J. Phys. Chem.* **94**, 7684–7692.

Sharp, K. A., Friedman, R., Misra, V., Hecht, J. & Honig, B. (1995). *Salt effects on polyelectrolyte–ligand binding: comparison of Poisson–Boltzmann and limiting law counterion binding models. Biopolymers*, **36**, 245–262.

Simonson, T. & Brooks, C. L. (1996). *Charge screening and the dielectric-constant of proteins: insights from molecular-dynamics. J. Am. Chem. Soc.* **118**, 8452–8458.

Simonson, T. & Brünger, A. (1994). *Solvation free energies estimated from a macroscopic continuum theory. J. Phys. Chem.* **98**, 4683–4694.

Simonson, T. & Perahia, D. (1995). *Internal and interfacial dielectric properties of cytochrome c from molecular dynamics in aqueous solution. Proc. Natl Acad. Sci. USA*, **92**, 1082–1086.

Sitkoff, D., Sharp, K. & Honig, B. (1994). *Accurate calculation of hydration free energies using macroscopic solvent models. J. Phys. Chem.* **98**, 1978–1988.

Slagle, S., Kozack, R. E. & Subramaniam, S. (1994). *Role of electrostatics in antibody–antigen association: anti-hen egg lysozyme/lysozyme complex (HyHEL–5/HEL). J. Biomol. Struct. Dyn.* **12**, 439–456.

Smith, P., Brunne, R., Mark, A. & van Gunsteren, W. (1993). *Dielectric properties of trypsin inhibitor and lysozyme calculated from molecular dynamics simulations. J. Phys. Chem.* **97**, 2009–2014.

Still, C., Tempczyk, A., Hawley, R. & Hendrickson, T. (1990). *Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc.* **112**, 6127–6129.

Takashima, S. & Schwan, H. P. (1965). *Dielectric constant measurements on dried proteins. J. Phys. Chem.* **69**, 4176.

Warshel, A. & Åqvist, J. (1991). *Electrostatic energy and macromolecular function. Annu. Rev. Biophys. Biophys. Chem.* **20**, 267–298.

Warshel, A. & Russell, S. (1984). *Calculations of electrostatic interactions in biological systems and in solutions. Q. Rev. Biophys.* **17**, 283.

Warwicker, J. (1994). *Improved continuum electrostatic modelling in proteins, with comparison to experiment. J. Mol. Biol.* **236**, 887–903.

Warwicker, J. & Watson, H. C. (1982). *Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. J. Mol. Biol.* **157**, 671–679.

Wendoloski, J. J. & Matthew, J. B. (1989). *Molecular dynamics effects on protein electrostatics. Proteins*, **5**, 313.

Yang, A., Gunner, M., Sampogna, R., Sharp, K. & Honig, B. (1993). *On the calculation of pK_a in proteins. Proteins Struct. Funct. Genet.* **15**, 252–265.

Yoon, L. & Lenhoff, A. (1992). *Computation of the electrostatic interaction energy between a protein and a charged suface. J. Phys. Chem.* **96**, 3130–3134.

Zauhar, R. & Morgan, R. J. (1985). *A new method for computing the macromolecular electric potential. J. Mol. Biol.* **186**, 815–820.

Zhou, H. X. (1994). *Macromolecular electrostatic energy within the nonlinear Poisson–Boltzmann equation. J. Phys. Chem.* **100**, 3152–3162.

22.4

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). *Protein Data Bank archives of three-dimensional macromolecular structures. Methods Enzymol.* **277**, 556–571.

Allen, F. H., Baalham, C. A., Lommerse, J. P. M. & Raithby, P. R. (1998). *Carbonyl–carbonyl interactions can be competitive with hydrogen bonds. Acta Cryst.* B**54**, 320–329.

Allen, F. H., Bird, C. M., Rowland, R. S. & Raithby, P. R. (1997*a*). *Resonance-induced hydrogen bonding at sulfur acceptors in $R_1R_2C{=}S$ and $R_1CS_2^-$ systems. Acta Cryst.* B**53**, 680–695.

Allen, F. H., Bird, C. M., Rowland, R. S. & Raithby, P. R. (1997*b*). *Hydrogen-bond acceptor and donor properties of divalent sulfur. Acta Cryst.* B**53**, 696–701.

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *The development of versions 3 and 4 of the Cambridge Structural Database system. J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H., Doyle, M. J. & Auf der Heyde, T. P. E. (1991). *Automated conformational analysis from crystallographic data. 6. Principal-component analysis for n-membered carbocyclic rings ($n = 4, 5, 6$): symmetry considerations and correlations with ring-puckering parameters. Acta Cryst.* B**47**, 412–424.

Allen, F. H., Doyle, M. J. & Taylor, R. (1991). *Automated conformational analysis from crystallographic data. 3. Three-dimensional pattern recognition within the Cambridge Structural Database system: implementation and practical examples. Acta Cryst.* B**47**, 50–61.

22.4 (*cont.*)

Allen, F. H., Harris, S. E. & Taylor, R. (1996). *Comparison of conformer distributions in the crystalline state with conformational energies calculated by ab initio techniques*. *J. Comput.-Aided Mol. Des.* **10**, 247–254.

Allen, F. H., Howard, J. A. K. & Pitchford, N. A. (1996). *Symmetry-modified conformational mapping and classification of the medium rings from crystallographic data. IV. Cyclooctane and related eight-membered rings*. *Acta Cryst.* B**52**, 882–891.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds*. *J. Chem. Soc. Perkin Trans.* 2, pp. S1–S19.

Allen, F. H., Lommerse, J. P. M., Hoy, V. J., Howard, J. A. K. & Desiraju, G. R. (1997). *Halogen···O(nitro) supramolecular synthon in crystal engineering: a combined crystallographic database and ab initio molecular orbital study*. *Acta Cryst.* B**53**, 1006–1016.

Allen, F. H., Motherwell, W. D. S., Raithby, P. R., Shields, G. P. & Taylor, R. (1999). *Systematic analysis of the probabilities of formation of bimolecular hydrogen bonded ring motifs in organic crystal structures*. *New J. Chem.* **23**, 25–34.

Allen, F. H., Raithby, P. R., Shields, G. P. & Taylor, R. (1998). *Probabilities of formation of bimolecular cyclic hydrogen bonded motifs in organic crystal structures: a systematic database study*. *Chem. Commun.* pp. 1043–1044.

Allen, F. H., Rowland, R. S., Fortier, S. & Glasgow, J. I. (1990). *Knowledge acquisition from crystallographic databases: towards a knowledge-based approach to molecular scene analysis*. *Tetrahedron Comput. Methodol.* **3**, 757–774.

Ashida, T., Tsunogae, Y., Tanaka, I. & Yamane, T. (1987). *Peptide chain structure parameters, bond angles and conformation angles from the Cambridge Structural Database*. *Acta Cryst.* B**43**, 212–218.

Balasubramanian, R., Chidambaram, R. & Ramachandran, G. N. (1970). *Potential functions for hydrogen-bond interactions. II. Formulation of an empirical potential function*. *Biochim. Biophys. Acta*, **221**, 196–206.

Berkovitch-Yellin, Z. & Leiserowitz, L. (1984). *The role played by C—H···O and C—H···N interactions in determining molecular packing and conformation*. *Acta Cryst.* B**40**, 159–165.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database. A comprehensive relational database of three-dimensional structures of nucleic acids*. *Biophys. J.* **63**, 751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank. Nucleic Acids Res.* **28**, 235–242.

Bertolasi, V., Gilli, P., Ferretti, V. & Gilli, G. (1996). *Resonance-assisted O—H···O hydrogen bonding. Its role in the crystalline self-recognition of beta-diketone enols and its structural and IR characterisation*. *Chem. Eur. J.* **2**, 925–934.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987). *Knowledge-based prediction of protein structures and the design of novel molecules*. *Nature (London)*, **326**, 347–352.

Böhm, H.-J. (1992*a*). *The computer program LUDI: a new method for the de novo design of enzyme inhibitors*. *J. Comput.-Aided Mol. Des.* **6**, 61–78.

Böhm, H.-J. (1992*b*). *LUDI: rule-based automatic design of new substituents for enzyme inhibitors*. *J. Comput.-Aided Mol. Des.* **6**, 593–606.

Bondi, A. (1964). *van der Waals volumes and radii*. *J. Phys. Chem.* **68**, 441–452.

Boström, J., Norrby, P.-O. & Liljefors, T. (1998). *Conformational energy penalties of protein-bound ligands*. *J. Comput.-Aided Mol. Des.* **12**, 383–396.

Bower, M., Cohen, F. E. & Dunbrack, R. L. Jr (1997). *Prediction of protein side-chain rotamers from a backbone-dependent rotamer library*. *J. Mol. Biol.* **267**, 1268–1282.

Brammer, L., Zhao, D., Ladipo, F. T. & Braddock-Wilking, J. (1995). *Hydrogen bonds involving transition metal centres – a brief review*. *Acta Cryst.* B**51**, 632–640.

Bruno, I. J., Cole, J. C., Lommerse, J. P. M., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). *IsoStar: a library of information about nonbonded interactions*. *J. Comput.-Aided Mol. Des.* **11**, 525–537.

Bürgi, H.-B. & Dunitz, J. D. (1983). *From crystal statics to chemical dynamics*. *Acc. Chem. Res.* **16**, 153–161.

Bürgi, H.-B. & Dunitz, J. D. (1988). *Can statistical analysis of structural parameters from different crystal environments lead to quantitative energy relationships? Acta Cryst.* B**44**, 445–451.

Bürgi, H.-B. & Dunitz, J. D. (1994). *Structure correlation*. Weinheim: VCH Publishers.

Carbonell, J. (1989). Editor. *Machine learning – paradigms and methods*. Amsterdam: Elsevier.

Carrell, A. B., Shimoni, L., Carrell, C. J., Bock, C. W., Murray-Rust, P. & Glusker, J. P. (1993). *The stereochemistry of the recognition of nitrogen-containing heterocycles by hydrogen bonding and by metal ions*. *Receptor*, **3**, 57–76.

Carrell, C. J., Carrell, H. L., Erlebacher, J. & Glusker, J. P. (1988). *Structural aspects of metal ion–carboxylate interactions*. *J. Am. Chem. Soc.* **110**, 8651–8656.

Ceccarelli, C., Jeffrey, G. A. & Taylor, R. (1981). *A survey of O—H···O hydrogen bond geometries determined by neutron diffraction*. *J. Mol. Struct.* **70**, 255–271.

Chakrabarti, P. (1990*a*). *Interaction of metal ions with carboxylic and carboxamide groups in protein structures*. *Protein. Eng.* **4**, 49–56.

Chakrabarti, P. (1990*b*). *Geometry of interaction of metal ions with histidine residues in protein structures. Protein Eng.* **4**, 57–63.

Chakrabarti, P. & Dunitz, J. D. (1982). *Directional preferences of ether O-atoms towards alkali and alkaline earth cations. Helv. Chim. Acta*, **65**, 1482–1488.

Chatfield, C. & Collins, A. J. (1980). *Introduction to multivariate analysis*. London: Chapman & Hall.

Conklin, D., Fortier, S., Glasgow, J. I. & Allen, F. H. (1996). *Conformational analysis from crystallographic data using conceptual clustering*. *Acta Cryst.* B**52**, 535–549.

Cremer, D. & Pople, J. A. (1975). *A general definition of ring puckering coordinates*. *J. Am. Chem. Soc.* **97**, 1354–1358.

Deane, C. M., Allen, F. H., Taylor, R. & Blundell, T. L. (1999). *Carbonyl–carbonyl interactions stabilise the partially allowed Ramachandran conformations of asparagine and aspartic acid. Protein Eng.* **12**, 1025–1028.

Derewenda, Z. S., Lee, L. & Derewenda, U. (1995). *The occurrence of C—H···O hydrogen bonds in proteins*. *J. Mol. Biol.* **252**, 248–262.

Desiraju, G. R. (1989). *Crystal engineering: the design of organic solids*. New York: Academic Press.

Desiraju, G. R. (1991). *The C—H···O hydrogen bond in crystals. What is it? Acc. Chem. Res.* **24**, 270–276.

Desiraju, G. R. (1995). *Supramolecular synthons in crystal engineering – a new organic synthesis. Angew. Chem. Int. Ed. Engl.* **34**, 2311–2327.

Desiraju, G. R., Kashino, S., Coombs, M. M. & Glusker, J. P. (1993). *C—H···O packing motifs in some cyclopenta[a]phenanthrenes*. *Acta Cryst.* B**49**, 880–892.

Desiraju, G. R. & Murty, B. N. (1987). *Correlation between crystallographic and spectroscopic properties for C—H···O bonds in terminal acetylenes. Chem. Phys. Lett.* **139**, 360–361.

Donohue, J. (1952). *The hydrogen bond in organic crystals. J. Phys. Chem.* **56**, 502–510.

Donohue, J. (1968). *Selected topics in hydrogen bonding*. In *Structural chemistry and molecular biology*, edited by W. Davidson & E. Rich, pp. 443–465. San Francisco: W. H. Freeman

Dunbrack, R. L. Jr & Karplus, M. (1993). *Backbone-dependent rotamer library for proteins: applications to side-chain prediction. J. Mol. Biol.* **230**, 534–571.

Dunitz, J. D. & Taylor, R. (1997). *Organic fluorine hardly ever accepts hydrogen bonds. Chem. Eur. J.* **3**, 83–90.

22.4 (*cont.*)

Einspahr, H. & Bugg, C. E. (1981). *The geometry of calcium–carboxylate interactions in crystalline complexes. Acta Cryst.* B**37**, 1044–1052.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

Everitt, B. (1980). *Cluster analysis*. New York: Wiley.

Fortier, S., Castleden, I., Glasgow, J., Conklin, D., Walmsley, C., Leherte, L. & Allen, F. H. (1993). *Molecular scene analysis: the integration of direct-methods and artificial-intelligence strategies for solving protein crystal structures. Acta Cryst.* D**49**, 168–178.

Glusker, J. P. (1980). *Citrate conformation and chelation: enzymatic implications. Acc. Chem. Res.* **13**, 345–352.

Glusker, J. P., Lewis, M. & Rossi, M. (1994). *Crystal structure analysis for chemists and biologists*. Weinheim: VCH Publishers.

Goodford, P. J. (1985). *A computational procedure for determining energetically favourable binding sites on biologically important molecules. J. Med. Chem.* **28**, 849–857.

Gould, R. O., Gray, A. M., Taylor, P. & Walkinshaw, M. D. (1985). *Crystal environments and geometries of leucine, isoleucine, valine and phenylalanine provide estimates of minimum nonbonded contact and preferred van der Waals interaction distances. J. Am. Chem. Soc.* **107**, 5921–5927.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The crystallographic information file (CIF): a new standard archive file for crystallography. Acta Cryst.* A**47**, 655–685.

Hayes, I. C. & Stone, A. J. (1984). *An intermolecular perturbation theory for the region of moderate overlap. J. Mol. Phys.* **53**, 83–105.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Errors in protein structures. Nature (London)*, **381**, 272.

Jeffrey, G. A. & Maluszynska, H. (1982). *A survey of hydrogen bond geometries in the crystal structures of amino acids. Int. J. Biol. Macromol.* **4**, 173–185.

Jeffrey, G. A. & Maluszynska, H. (1986). *A survey of the geometry of hydrogen bonds in the crystal structures of barbiturates, purines and pyrimidines. J. Mol. Struct.* **147**, 127–142.

Jeffrey, G. A. & Mitra, J. (1984). *Three-centre (bifurcated) hydrogen bonding in the crystal structures of amino acids. J. Am. Chem. Soc.* **106**, 5546–5553.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen bonding in biological structures*. Berlin: Springer Verlag.

Jones, G., Willett, P. & Glen, R. C. (1995). *Molecular recognition of receptor sites using a genetic algorithm with a description of solvation. J. Mol. Biol.* **245**, 43–53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). *Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol.* **267**, 727–748.

Joris, L., Schleyer, P. & Gleiter, R. (1968). *Cyclopropane rings as proton acceptor groups in hydrogen bonding. J. Am. Chem. Soc.* **90**, 327–336.

Kennard, O. (1962). In *International tables for X-ray crystallography*, Vol. II. Birmingham: Kynoch Press.

Kennard, O. & Allen, F. H. (1993). *The Cambridge Crystallographic Data Centre. Chem. Des. Autom. News*, **8**, 1, 31–37.

Klebe, G. (1994). *The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands. J. Mol. Biol.* **237**, 212–235.

Klebe, G. & Mietzner, T. (1994). *A fast and efficient method to generate biologically relevant conformations. J. Comput.-Aided Mol. Des.* **8**, 583–594.

Kroon, J. & Kanters, J. A. (1974). *Non-linearity of hydrogen bonds in molecular crystals. Nature (London)*, **248**, 667–669.

Kroon, J., Kanters, J. A., van Duijneveldt-van de Rijdt, J. G. C. M., van Duijneveldt, F. B. & Vliegenthart, J. A. (1975). *O—H···O hydrogen bonds in molecular crystals: a statistical and quantum chemical analysis. J. Mol. Struct.* **24**, 109–129.

Kuntz, I. D., Meng, E. C. & Stoichet, B. K. (1994). *Structure-based molecular design. Acc. Chem. Res.* **27**, 117–122.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.

Laskowski, R. A., Thornton, J. M., Humblet, C. & Singh, J. (1996). *X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. J. Mol. Biol.* **259**, 175–201.

Lehn, J.-M. (1988). *Perspectives in supramolecular chemistry – from molecular recognition towards molecular information processing and self-organization. Angew. Chem. Int. Ed. Engl.* **27**, 90–112.

Lemmen, C. & Lengauer, T. (1997). *Time-efficient flexible superposition of medium sized molecules. J. Comput.-Aided Mol. Des.* **11**, 357–368.

Levitt, M. & Perutz, M. (1988). *Aromatic rings act as hydrogen bond acceptors. J. Mol. Biol.* **201**, 751–754.

Lommerse, J. P. M., Price, S. L. & Taylor, R. (1997). *Hydrogen bonding of carbonyl, ether and ester oxygen atoms with alkanol hydroxyl groups. J. Comput. Chem.* **18**, 757–780.

Lommerse, J. P. M., Stone, A. J., Taylor, R. & Allen, F. H. (1996). *The nature and geometry of intermolecular interactions between halogens and oxygen or nitrogen. J. Am. Chem. Soc.* **118**, 3108–3116.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995*a*). *Coulombic interactions between partially charged main-chain atoms not hydrogen bonded to each other influence the conformations of α-helices and antiparallel β-sheets. J. Mol. Biol.* **248**, 361–373.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995*b*). *Coulombic interactions between partially charged main-chain atoms stabilise the right-handed twist found in most β-strands. J. Mol. Biol.* **248**, 374–384.

Miller, J., McLachlan, A. D. & Klug, A. (1985). *Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J.* **4**, 1609–1614.

Murray-Rust, P. & Bland, R. (1978). *Computer retrieval and analysis of molecular geometry. II. Variance and its interpretation. Acta Cryst.* B**34**, 2527–2533.

Murray-Rust, P. & Glusker, J. P. (1984). *Directional hydrogen bonding to sp^2 and sp^3 hybridized oxygen atoms and its relevance to ligand–macromolecule interactions. J. Am. Chem. Soc.* **105**, 1018–1025.

Murray-Rust, P. & Motherwell, S. (1978). *Computer retrieval and analysis of molecular geometry. III. Geometry of the β-1′-aminofuranoside fragment. Acta Cryst.* B**34**, 2534–2546.

Nicklaus, M. C., Wang, S., Driscoll, J. S. & Milne, G. W. A. (1995). *Conformational changes of small molecules binding to proteins. Bioorg. Med. Chem.* **3**, 411–428.

Nobeli, I., Price, S. L., Lommerse, J. P. M. & Taylor, R. (1997). *Hydrogen bonding properties of oxygen and nitrogen acceptors in aromatic heterocycles. J. Comput. Chem.* **18**, 2060–2074.

Nyburg, S. C. & Faerman, C. H. (1985). *A revision of van der Waals atomic radii for molecular crystals: N, O, F, S, Cl, Se, Br and I bonded to carbon. Acta Cryst.* B**41**, 274–279.

Orpen, A. G., Brammer, L., Allen, F. H., Kennard, O., Watson, D. G. & Taylor, R. (1989). *Tables of bond lengths determined by X-ray and neutron diffraction. Part II: organometallic compounds and coordination complexes of the d- and f-block metals. J. Chem. Soc. Dalton Trans.* pp. S1–S83.

Pauling, L. (1939). *The nature of the chemical bond*. Ithaca: Cornell University Press.

Pedireddi, V. R. & Desiraju, G. R. (1992). *A crystallographic scale of carbon acidity. Chem. Commun.* pp. 988–990.

Pimentel, G. C. & McClellan, A. L. (1960). *The hydrogen bond*. San Francisco: W. H. Freeman.

Price, S. L., Stone, A. J., Lucas, J., Rowland, R. S. & Thornley, A. E. (1994). *The nature of —Cl···Cl— intermolecular interactions. J. Am. Chem. Soc.* **116**, 4910–4918.

Rappoport, Z., Biali, S. E. & Kaftory, M. (1990). *Application of the structure correlation method to ring-flip processes in benzophenone. J. Am. Chem. Soc.* **112**, 7742–7750.

Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996). *Predicting receptor-ligand interactions by an incremental construction algorithm. J. Mol. Biol.* **261**, 470–481.

22.4 (*cont.*)

Robertson, J. M. (1953). *Organic crystals and molecules*. Ithaca: Cornell University Press.

Rosenfield, R. E., Parthasarathy, R. & Dunitz, J. D. (1977). *Directional preferences of non-bonded atomic contacts with divalent sulphur. 1. Electrophiles and nucleophiles. J. Am. Chem. Soc.* **99**, 4860–4862.

Rosenfield, R. E. Jr, Swanson, S. M., Meyer, E. F. Jr, Carrell, H. L. & Murray-Rust, P. (1984). *Mapping the atomic environment of functional groups: turning 3D scatterplots into pseudo-density contours. J. Mol. Graphics*, **2**, 43–46.

Rowland, R. S. & Taylor, R. (1996). *Intermolecular nonbonded contact distances in organic crystal structures: comparison with distances expected from van der Waals radii. J. Phys. Chem.* **100**, 7384–7391.

Schweizer, W. B. & Dunitz, J. D. (1982). *Structural characteristics of the carboxylic acid ester group. Helv. Chim. Acta*, **65**, 1547–1552.

Singh, J., Saldanha, J. & Thornton, J. M. (1991). *A novel method for the modelling of peptide ligands to their receptors. Protein Eng.* **4**, 251–261.

Steiner, T., Kanters, J. A. & Kroon, J. (1996). *Acceptor directionality of sterically unhindered C—H···O=C hydrogen bonds donated by acidic C—H groups. J. Chem. Soc. Chem. Commun.* **11**, 1277–1278.

Steiner, T. & Saenger, W. (1992). *Geometry of C—H···O hydrogen bonds in carbohydrate crystal structures. Analysis of neutron diffraction data. J. Am. Chem. Soc.* **114**, 10146–10154.

Steiner, T., Starikov, E. B., Amado, A. M. & Teixeira-Dias, J. J. C. (1995). *Weak hydrogen bonding. Part 2. The hydrogen-bonding nature of short C—H··· • contacts. Crystallographic, spectroscopic and quantum mechanistic studies of some terminal alkynes. J. Chem. Soc. Perkin Trans.* 2, pp. 1312–1326.

Strynadka, N. C. J. & James, M. N. G. (1989). *Crystal structures of the helix-loop-helix calcium-binding proteins. Annu. Rev. Biochem.* **58**, 951–998.

Sutton, L. E. (1956). *Tables of interatomic distances and configuration in molecules and ions*. Special Publication No. 11. London: The Chemical Society.

Sutton, L. E. (1959). *Tables of interatomic distances and configuration in molecules and ions (supplement)*. Special Publication No. 18. London: The Chemical Society.

Taylor, R. (1986). *The Cambridge Structural Database in molecular graphics: techniques for the rapid identification of conformational minima. J. Mol. Graphics*, **4**, 123–131.

Taylor, R. & Allen, F. H. (1994). *Statistical and numerical methods of data analysis*. In *Structure correlation*, edited by H.-B. Bürgi & J. D. Dunitz. Weinheim: VCH Publishers.

Taylor, R. & Kennard, O. (1982). *Crystallographic evidence for the existence of C—H···O, C—H···N and C—H···Cl hydrogen bonds. J. Am. Chem. Soc.* **104**, 5063–5070.

Taylor, R. & Kennard, O. (1983). *Comparison of X-ray and neutron diffraction results for the N—H···O=C hydrogen bond. Acta Cryst.* B**39**, 133–138.

Taylor, R., Kennard, O. & Versichel, W. (1983). *Geometry of the N—H···O=C hydrogen bond. 1. Lone-pair directionality. J. Am. Chem. Soc.* **105**, 5761–5766.

Taylor, R., Kennard, O. & Versichel, W. (1984*a*). *Geometry of the N—H···O=C hydrogen bond. 2. Three-centre (bifurcated) and four-centre (trifurcated) bonds. J. Am. Chem. Soc.* **106**, 244–248.

Taylor, R., Kennard, O. & Versichel, W. (1984*b*). *Geometry of the N—H···O=C hydrogen bond. 3. Hydrogen-bond distances and angles. Acta Cryst.* B**40**, 280–288.

Taylor, R., Mullaley, A. & Mullier, G. W. (1990). *Use of crystallographic data in searching for isosteric replacements: composite field environments of nitro and carbonyl groups. Pestic. Sci.* **29**, 197–213.

Thornton, J. M. & Gardner, S. P. (1989). *Protein motifs and database searching. Trends Biochem. Sci.* **14**, 300–304.

Tintelnot, M. & Andrews, P. (1989). *Geometries of functional group interactions in enzyme–ligand complexes: guides for receptor modelling. J. Comput.-Aided Mol. Des.* **3**, 67–84.

Verdonk, M. L. (1998). Unpublished results.

Verdonk, M. L., Cole, J. C. & Taylor, R. (1999). *SuperStar: a knowledge-based approach for identifying interaction sites in proteins. J. Mol. Biol.* **289**, 1093–1108.

Viswamitra, M. A., Radhakrishnan, R., Bandekar, J. & Desiraju, G. R. (1993). *Evidence for O—H···C and N—H···C hydrogen bonding. J. Am. Chem. Soc.* **115**, 4868–4869.

Whitesides, G. M., Simanek, E. E., Mathias, J. P., Seto, C. T., Chin, D. N., Mammen, M. & Gordon, D. M. (1995). *Non-covalent synthesis: using physical-organic chemistry to make aggregates. Acc. Chem. Res.* **28**, 37–43.