Comparison of Conformations of Small Molecule Structures from the Protein Data Bank with Those Generated by Concord, Cobra, ChemDBS-3D, and Converter and Those Extracted from the Cambridge Structural Database

Eleanor M. Ricketts,* John Bradshaw, Mike Hann, Fiona Hayes, and Neil Tanna
Glaxo Group Research, Greenford, Middlesex UB6 0HE, U.K.
David M. Ricketts
Biosym Technologies Inc., 9685 Scranton Road, San Diego, California 92121

Received June 10, 1993®

This paper explores the ability of the 2D–3D structure conversion packages Concord, Cobra, ChemDBS-3D, and Converter to generate the structures of bound small molecules and compares their conformations with those in ligand complex crystal entries in the Brookhaven Protein Data Bank. ChemDBS-3D is limited by the size of structure that it can handle in its database environment and can only process 62% of the structures, but when these structures are compared with the ligand structures, they are found to most closely approximate the bound conformation. However Converter can be considered to perform better as it can convert 100% of the structures from input 2D diagrams, and its conformations are a good match in most cases. Concord performs well in converting 92% of the structures from input SMILES strings, and the conformations that it produces are good in many cases. The version of Cobra used (1.1) was found to have many problems, which limited its usefulness in this study; Cobra correctly processed only 48% of the structures. Comparisons of the conformations of some of the ligands with entries from the Cambridge Structural Database have mixed success.

INTRODUCTION

In order to accurately model the interaction of a drug with its receptor, it is essential to know the conformation of the drug involved in the interaction. Very little is known about what changes occur either to the drug molecule or to the receptor during this interaction. A molecule may adopt a number of different conformations influenced by its environment; therefore it may be difficult to determine which of the possible conformations to choose when modeling its interaction with the receptor. Since it is the *bound* conformation of a drug which is responsible for activity, studies which focus on solving the crystal structure of the drug–receptor complex would be the best method of determining the conformation of interest. However ligand–macromolecule complexes suitable for crystallographic determination are often not available. The modeler is therefore forced to determine the bound conformation by some other method.

The development of three-dimensional (3D) database searching systems[1] produced a demand to convert information held in two-dimensional (2D) connectivity databases to 3D coordinates. Of the four commercial 2D–3D conversion programs available for testing at the time of this study (1992), the most widely used is Concord.[2,3] This was developed specifically to enable 2D to 3D structure conversion and produces a single low-energy conformation. Concord has been used in the conversion of the Lederle in-house database[4] (88% successfully converted) and to produce 3D coordinates for the Drug Data Report database[5] (82% successfully converted). The other three programs, Cobra,[6–8] ChemDBS-3D,[9] and Converter,[10] produce more than one low-energy conformation for an input structure wherever possible. ChemDBS-3D was developed to provide the ability to convert 2D structures into 3D, taking into account the conformational space available to the structure, and then store and search for these structures in a 3D database system. ChemDBS-3D has been used to convert the Chapman and Hall *Dictionary of Drugs*[11] (87% successfully converted). Cobra was developed for conformational analysis of a structure rather than the rapid conversion of a large number of 2D structures. As part of this procedure, it accepts 2D input and outputs a 3D coordinate structure; therefore it was decided to review its performance in this study. Converter was developed to convert 2D to 3D and can produce multiple conformations that are sterically accessible.

The use of structures generated by these methods in drug–receptor studies is under debate because of the problem of conformational flexibility of drug molecules. Cobra, ChemDBS-3D, and Converter attempt to provide a solution to this problem in that they generate more than one conformation of a structure.

This study investigates the validity of using 2D–3D structure converters to produce 3D conformations of ligands by comparing them with ligands experimentally determined in their bound conformation and stored in the Brookhaven Protein Data Bank (PDB).[12,13] Where they exist, structures with a similar 2D structure to the bound structures but experimentally determined by small molecule crystallography are extracted from the Cambridge Structural Database (CSD)[14,15] for comparison. This work compliments two recent studies which compared 90 X-ray crystallographic structures from the CSD with those generated by Concord and ChemDBS-3D.[16,17] The studies found that Concord exactly matched (with a root mean square (RMS) value of all non-hydrogen atoms of <0.5 Å) 38% of the structures while ChemDBS-3D exactly matched 57% of the structures.

This paper discusses the choice of test sets and the problem with using crystallographic data. The measurements used to compare the generated conformations with those from the PDB are described, and the methodology used by each of the structure generation packages is introduced. The performance

* To whom all correspondence should be addressed at Biosym Technologies Inc., 9685 Scranton Rd., San Diego, CA 92121-3752.

® Abstract published in *Advance ACS Abstracts,* October 15, 1993.

0095-2338/93/1633-0905$04.00/0

of each package in building the test structures, and, in the cases of the database systems CSD and ChemDBS-3D, the search for the test structures is described. Finally, the generated conformations are compared with the PDB conformations and the performance of each package is discussed.

METHODS

The Test Sets. The test sets were extracted from the PDB and comprise small molecule structures as substrates, inhibitors, cofactors, or prosthetic groups bound to protein structures. To identify the relevant PDB entries, the entire database was searched for those entries with heteroatom records (HET) which detail any nonstandard groups (e.g., nonstandard residues, prosthetic groups, inhibitors, solvent molecules, etc.). Entries were discarded if the HET structures were DNA or RNA; if they were surface binding ligands such as FAD or ATP or had a surface binding molecule together with another small molecule; if they were coenzymes, sugars, immunoglobulins, active site ions, solvents (unless they were with an inhibitor), or particularly large ligands or ligands with nonorganic elements. Other entries were removed if they were too complex or there were atoms missing from the crystal structure. Finally structures were removed if the structure diagrams of the small molecules, assigned to the atoms in the 3D coordinate data set with reference to the Fine Chemicals Directory,[18] did not tally with the PDB name. PDB assigns standard nomenclature and ordering of the atoms in some HET structures, but not in all, which may lead to problems of identification.[12]

After applying the above selection criteria, 52 entries remained. The chosen PDB entries are referred to by their PDB code. The entry for 3DFR was found to contain two small molecule structures of interest—NADPH and methotrexate. Both of these molecules were considered in this study and are differentiated by the addition of the suffix "M" to the four letter PDB code for the methotrexate structure (3DFRM). The NADPH structure retains the original PDB code (3DFR). This increases the number of small molecule entries to 53. Two test sets were generated from these entries. The first test set comprises 18 entries which, when the 2D atomic structure was input as a query to the CSD, resulted in hits that had structures very similar to the PDB query structure. The second test set has 35 PDB structures which did not produce suitable CSD hits.

Table I lists the PDB small molecules in each test set; it gives their PDB assigned name, their function (assigned in most cases by reference to the original paper, but in some cases assumed by the authors), and the protein molecules with which they are complexed. The first test set has 14 *unique* 2D structures out of 18 (Figure 1); there is more than 1 PDB entry for methotrexate (3DFR and 4DFR), flavin mononucleotide (1FCB, 1FX1, and 1GOX) and benzamidine (3PTB and 2TRM) in the set. In the second test set there are 34 unique structures out of 35 (Figure 2). The 2D structures for the two thymidine inhibitors (1SNC and 2SNS) are the same, although their systematic names given in the PDB entries are different. The two Rhinovirus antiviral agents compound I(R) (2RR1) and compound I(S) (2RS1) differ only in chirality at their one chiral center but were treated independently. The *Searching and Building* section will therefore describe the attempts to build the 14 unique structures in the first test set and the 34 unique structures in the second test set.

In some of the PDB files used in this study there are two occurrences of the same small molecule structure bound to different protein chains in one entry. In the results tables these multiple occurrences within one PDB entry are differentiated by adding the prefix "CX" to the four letter PDB code. For example, the entry for 8ATC has two chains, chain A and chain C. The two phosphonacetyl-L-aspartate structures bound to these chains are called CA8ATC and CC8ATC, respectively. This means that at the *Conformation Comparison* stage there are a total of 26 small molecule conformations in the first test set and 38 small molecule conformations in the second test set that can be compared.

Crystallographic Problems. There are problems in using crystallographic data in a test set because the reliability of the atomic positions is dependent on the conditions of the study.[19] This study requires a quantitative figure of the reliability of the atomic coordinates in the bound small molecule. Various figures generated during the structure determination, such as resolution, *R*-factor, temperature factor, and RMS deviations for several distances, are stored in the PDB entry and may be used to determine the reliability of the structure determination. Unfortunately these figures, with the exception of the temperature factor, apply to the whole entry and will therefore be dominated by the effect of the macromolecular structure.

The temperature factor is calculated for each atom during refinement, and the greater its value, the less localized the atom is in the crystal.[20] However, bound small molecules may have their coordinates solved from the electron density map or they may be fitted by using a CSD structure or a structure obtained from a molecular modeling program. If they are fitted, there may be little or no further refinement of the coordinates which may produce an invalid temperature factor. As the crystal structure determinations used in this study do not report the origin of the small molecule coordinates, it was decided that the temperature factor could not be used as a reliable means of estimating the deviation of the atomic positions.

Due to these problems it was decided that none of the figures stored with the PDB entry could provide the quantitative figure required, therefore other measurements of accuracy were considered. Glusker and Trueblood[21] recommend that all interatomic distances are regarded as only being accurate to the nearest 0.01 Å because even in the most careful work, systematic errors may occur. However, this study uses torsion angles rather than interatomic distances to compare the conformations; thus the figure cannot be accurately employed.

It was considered that it might be possible to calculate an RMS fit for each PDB small molecule to indicate how similar the coordinates of their atoms might be to those of an "ideal" structure. Unfortunately, the concept of an ideal structure is also not straightforward. For example, an ideal pyridine ring is built by different software packages with different bond lengths:

	C–N (aromatic)	C–C (aromatic)
CSD bond length tables	1.337	1.379
SYBYL	1.349	1.398
Concord	1.340	1.400
Cobra	1.371	1.396
Chem-X	1.340	1.384
Converter	1.340	1.390

It is unclear which of these models should be used to produce the ideal pyridine ring. The pyridine ring in the PDB small molecule structure of pyridoxamine phosphate (2AAT) has the bond lengths

	C–N (aromatic)	C–C (aromatic)
2AAT	1.350/1.327	1.411/1.400/1.397/1.366

Table I. Description of the Small Molecules from the Protein Data Bank Used in the Two Test Sets

PDB code	small molecule	macromolecule	function
(a) Test Set 1			
2AAT	pyridoxamine phosphate	aspartate aminotransferase	cofactor
8ATC	phosphonacetyl-L-aspartate	aspartate carbamoyltransferase	substrate
1CLA	chloramphenicol	chloramphenicol aminotransferase	substrate
3CPP	camphor	cytochrome P450CAM	substrate
3DFRM	methotrexate	dihydrofolate reductase	inhibitor
4DFR	methotrexate	dihydrofolate reductase	inhibitor
7DFR	folate	dihydrofolate reductase	substrate
2DHF	deazafolate	dihydrofolate reductase	inhibitor
1FCB	flavin mononucleotide	flavocytochrome B2	prosthetic group
1FX1	flavin mononucleotide	flavodoxin	prosthetic group
1GOX	flavin mononucleotide	glycolate oxidase	prosthetic group
2GBP	β-D-glucose	D-galactose/D-glucose binding protein	substrate
8LDH	citrate	lactate dehydrogenase	inhibiting effect
3PTB	benzamidine	beta-trypsin	inhibitor
8RSA	acetyl deoxythymidine	ribonuclease A	inhibitor
2TRM	benzamidine	trypsin	inhibitor
4XIA	sorbitol	xylose isomerase	inhibitor
5XIA	xylitol	xylose isomerase	inhibitor
(b) Test Set 2			
4CTS	oxaloacetate	citrate synthase	substrate
3DFR	NADPH	dihydrofolate reductase	cofactor
4ER1	PD125967	endothiapepsin	inhibitor
2EST	trifluoroacetyl-L-lysyl-L-alanyl-*p*-trifluoromethylphenylalanine	elastase	inhibitor
3GAP	cyclic AMP	catabolite gene activator protein	cofactor
7GCH	*N*-acetyl-L-leucyl-L-phenylalanyltrifluoromethyl ketone	gamma chymotrypsin	inhibitor
3GPD	NAD	D-glyceraldehyde-3-phosphate dehydrogenase	cofactor
4PAD	tosyl-methylenyllysyl	papain	inhibitor
5PAD	benzyloxycarbonyl-glycylphenylalanyl-ethylenylglycyl derivative	papain	inhibitor
6PAD	benzyloxycarbonyl-phenylalanyl-methylenylalanyl derivative	papain	inhibitor
2R04	compound IV	rhinovirus 14	antiviral agent
2R06	compound VI	rhinovirus 14	antiviral agent
2R07	compound VII	rhinovirus 14	antiviral agent
1R08	compound VIII	rhinovirus 14	antiviral agent
2RM2	compound II (R/S)	rhinovirus 14	antiviral agent
1RNT	2′-guanylic acid	ribonuclease T1	substrate
2RNT	guanylyl-2′,5′-guanosine	ribonuclease T1	substrate
2RR1	compound I(R)	rhinovirus 14	antiviral agent
2RS1	compound I(S)	rhinovirus 14	antiviral agent
2RS3	compound III(S)	rhinovirus 14	antiviral agent
2RS5	compound V(S)	rhinovirus 14	antiviral agent
9RSA	*N*-acetyldeoxyuridine	ribonuclease A	inhibitor
1SGC	chymostatin A	proteinase A	inhibitor
1SNC	3′,5′-deoxythymidine biphosphate	staphylococcal nuclease	inhibitor
2SNS	2′-deoxy-3′,5′-diphosphothymidine	staphylococcal nuclease	inhibitor
5TLN	HONH-benzylmalonyl-L-alanylglycine-*p*-nitroanilide	thermolysin	inhibitor
7TLN	$CH_2CO(N–OH)LEU–OCH_3$	thermolysin	inhibitor
1TLP	phosphoramidon	thermolysin	inhibitor
1TMN	*N*-1-carboxy-3-phenylpropyl-L-leucyl-L-tryptophan	thermolysin	inhibitor
4TMN	CBZ—PHE═P═LEU—ALA	thermolysin	inhibitor
5TMN	CBZ—GLY═P═LEU—ALA	thermolysin	inhibitor
6TMN	CBZ—GLY═P═(O)—LEU—ALA	thermolysin	inhibitor
1TPP	*p*-amidinophenylpyruvate	β-trypsin	inhibitor
3TS1	tyrosinyl adenylate	tyrosyl-transfer RNA synthetase	reaction intermediate
2YHX	*o*-toluoylglucosamine	yeast hexokinase B	inhibitor

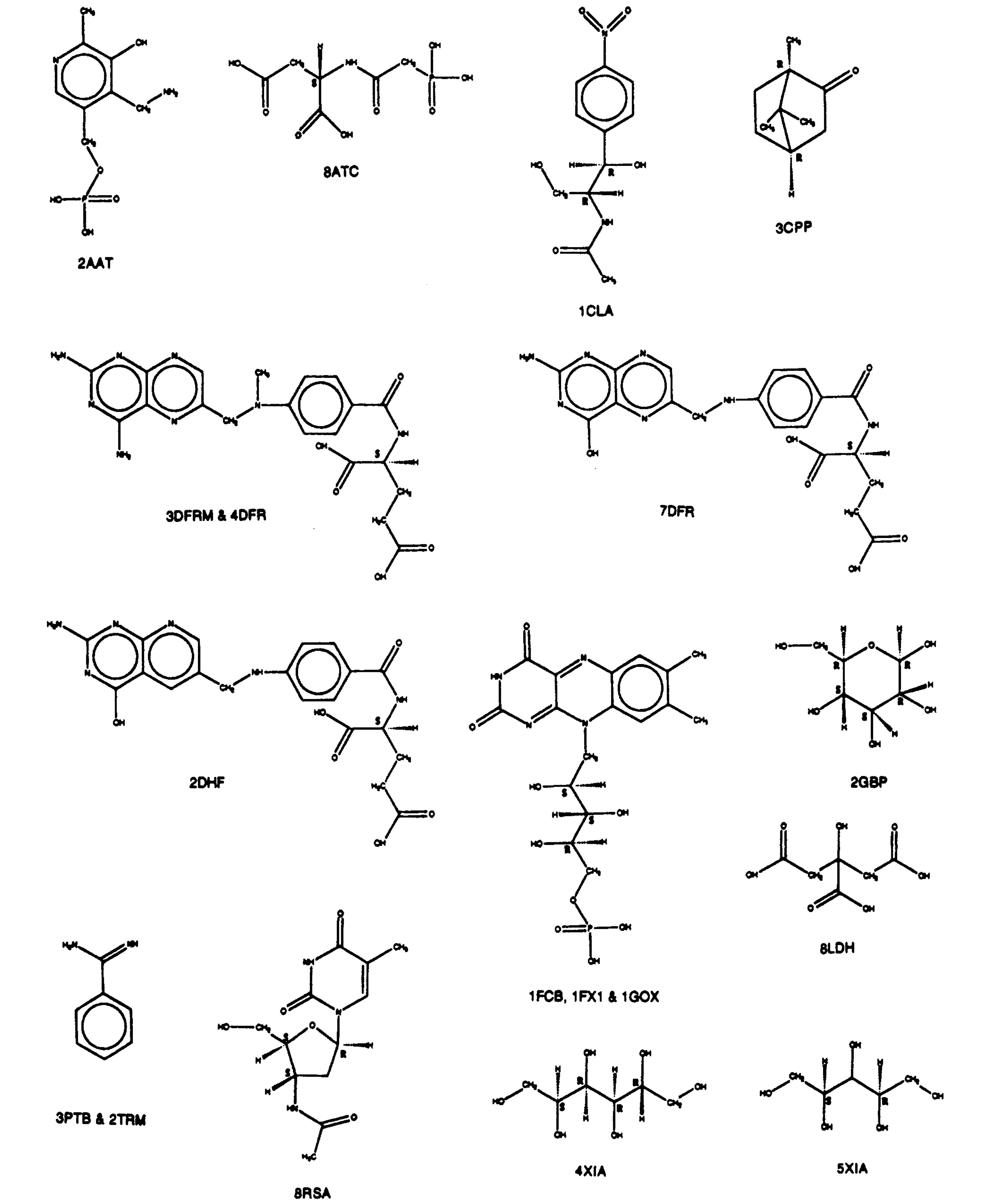


Figure 1. PDB small molecule structures in the first test set.

which does not seem to correspond closely to any of the above methods. It would therefore appear to be unfeasible to compare the small molecule structure to an ideal structure.

As a result it was decided that it would not be possible to take account of the accuracy of the atomic positions of the small molecule crystal structure and each case would have to be considered individually with no reference to an outside ideal or to other entries in the test set. It does serve to show that the atomic data from the PDB entries will not be exact, so a wide tolerance must be employed when built conformations are compared with the PDB bound conformation.

Comparison of Conformations. To compare the conformations produced by the different methods, it was decided that two measures would be used: a root mean square fit of ring atoms and a comparison of torsion angles in the backbone of the structure. An overall RMS fit was not considered as a gross figure does not give any indication of which parts of the structure are better than others. All of the structure building programs used in this study, except Converter, build the cyclic and acyclic portions of the structure separately before fitting them back together. Therefore comparing the building of each of these portions would appear to be sensible. Torsion angles were chosen in preference to interatomic distances because they are less sensitive to errors in analysis of the crystallographic data[22] and they can indicate differences in the conformation of structure which may have a good RMS fit.[23]

Similarity based on torsion angles was determined using a fairly wide range due to the uncertainty in the atomic positions of the test structures. Torsion angles within 60° of the bound structure value were considered to show some degree of similarity. Similarity based on RMS fits of rings was less clear, and it was decided that RMS values of ≤ 0.1 Å would be considered acceptable matches. Structures with the highest proportion of torsion angles $\leq 60°$ and RMS ring fits ≤ 0.1 Å were considered the best match.

Molecular Modeling Software and Hardware. SYBYL[24] was chosen as the molecular modeling program to enable display and comparison of structures because it was readily available on the VAX-VMS systems used in most of this study. Converter comparisons were carried out using the Biosym software INSIGHT II.[25] SYBYL provides all the features required to model and compare the structures but unfortunately was found to have problems assigning ring and end group atom and bond types when reading in crystallographic structures with incomplete hydrogen counts. As specific atom types are important in the production of SMILES strings[26,27] used as input to the structure building programs Concord, Cobra, and ChemDBS-3D, each structure had to be examined and edited by hand to produce the correct structure.

SYBYL does not output SMILES strings, so to produce strings representing the small molecules, a SYBYL mol file was generated for each structure and then entered in the Daylight software program GEMINI.[28] This produces a SMILES string with @ and @@ chirality flags which can be used as input to Cobra. Concord and ChemDBS-3D both require SMILES strings with R and S chirality flags. It was found that although R and S chirality flags can be assigned by SYBYL, these were sometimes incorrect when checked manually. The program CHIRON[29] accepts a SYBYL mol file as input and produces R and S chirality flags which appear to correspond most closely to the manual assignment. The appropriate carbon atoms in the GEMINI-produced SMILES strings were edited by hand to assign the R and S chirality flags.

Concord and Cobra conversions were carried out on a VAX 6000-460 on a network, and the CPU times for building were recorded. ChemDBS-3D conversion, keying, and database searching was done on a stand-alone IBM RS6000 320H, and CPU times were recorded. Converter was tested on a networked Silicon Graphics Personal IRIS 4D-30 and only the elapsed time was available. Accordingly, although the times for building are quoted for each method, they are not directly comparable and are given merely for the interest of users with similar hardware platforms.

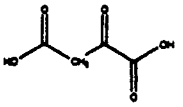
Description of the CSD and the Structure-Generation Packages. *Cambridge Structural Database, Version 4.5, July 1990.* The Cambridge Structural Database (CSD)[14,15] is the major database of experimentally determined 3D structures and is often used as the source of small molecule coordinate information for molecular modeling. Version 4.5 contained just over 90 000 organic and organometallic structures. This version provides 2D (QUEST) and 3D (GSTAT) search systems enabling structures to be located in the database. 2D searches can be carried out using a menu interface enabling input of structure diagrams. Searches can be very specific by using exact atom and bond types or more general by using groups of atoms or bond types. Hit lists of database structures are produced which can be examined further by eye or by using the 3D search system.

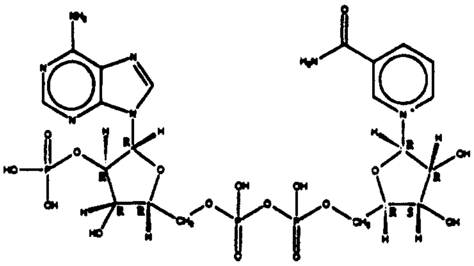
Concord, Version 2.9.1. Concord[2] generates one structure composed of "high quality approximate coordinates"[3] for almost all compounds comprised of organic elements with atomic connectivities ≤ 4. A combination of structure building rules and optimization procedures is used to produce the lowest energy configuration and conformation consistent with the input connectivity information. Problems are known to occur however with large structures of high potential flexibility which are not built in the lowest energy conformation, but rather in the fully extended conformation.[3]

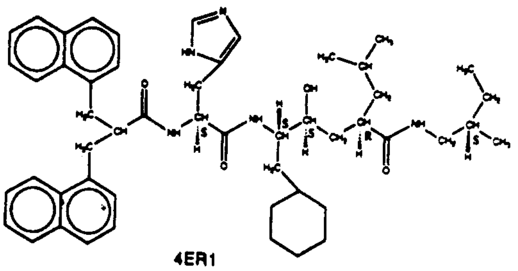
Concord builds the structures by analyzing a connection table set up from the unique version of an input SMILES string. The cyclic and acyclic portions are built separately, and the 3D pieces are then put back together to form the complete structure. Bond lengths in acyclic portions are assigned from a table of published values. The bond angles and torsion angles in simple, single cyclic portions are assigned by using a set of rules. Ring systems are built by prioritizing the gross conformations of each constituent ring and then placing them in order into the ring system using a minimizing strain function to produce the best bond and torsion angles. Acyclic bond angles are generally assigned on the basis of their hybridization from a small table of values; the torsion angles are allocated by using logical analysis to produce the best 1–4 interaction.

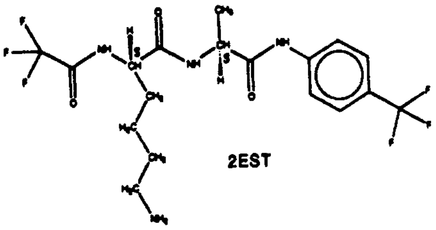
The main problem in building appears to be caused by close contacts which occur as the separately built portions of the structure are reassembled. Concord relieves many of these by adjusting the torsion angles, but some cannot be relieved and so the structure will not be built.

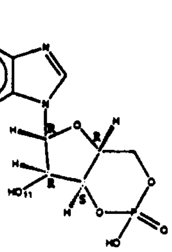
Cobra, January 1990, Version 1.1. Cobra aims to search the entire conformational space of a molecule to produce all of the low-energy conformations that the molecule can adopt.[6–8] This program was developed for use in conformational analysis studies, rather than in a database conversion system. The program adopts an artificial intelligence approach which uses a set of rules and a knowledge base to generate likely conformations from a unique SMILES string. Cobra determines the bond orders in the input structure on the basis of the atomic type and connectivity information and produces conformations for all possible stereoisomers if the stereochemistry is not explicitly defined. As in Concord, the input structure is separated into cyclic and acyclic portions, but in Cobra the structures of these portions are determined by a set of "conformational units" stored in the knowledge base. Both simple and complex units are stored, and a structure is represented by the most complex units possible. Ring systems are identified by assigning the smallest set of smallest rings possible to the molecule. If Cobra cannot match a part of the structure to any of its stored units, it tries in successive stages to match units which have more generalized atomic geometrical features. If part of a structure cannot be identified even with the most generalized units, the structure cannot be built. However the user can add a new unit to the knowledge base, provided that all the information about the possible conformations that it can adopt are known.

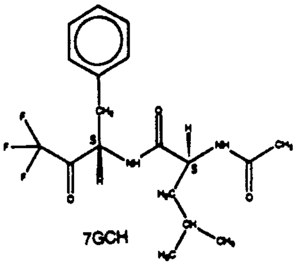

4CTS

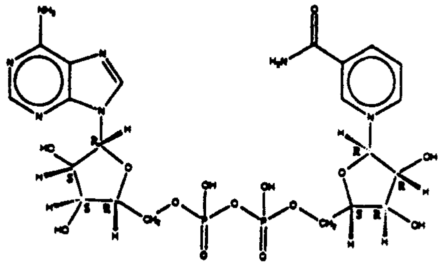

3DFR

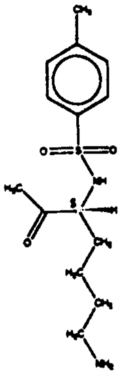

4ER1

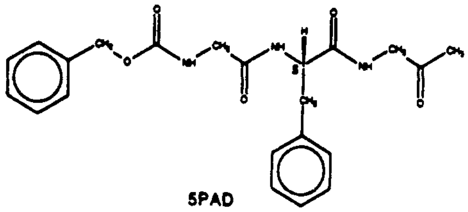

2EST

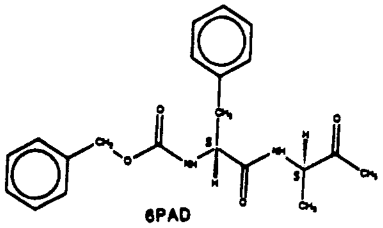

3GAP

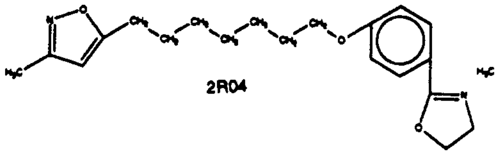

7GCH

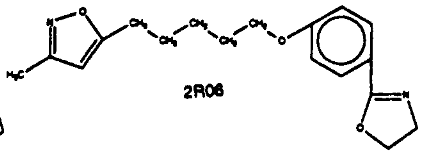

3GPD

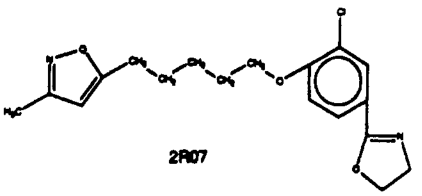

4PAD

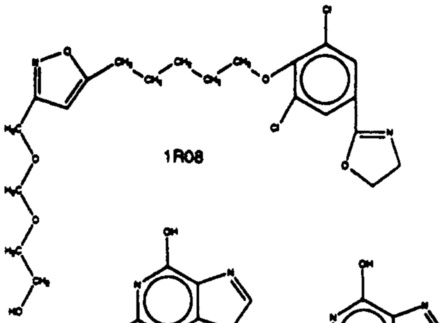

5PAD

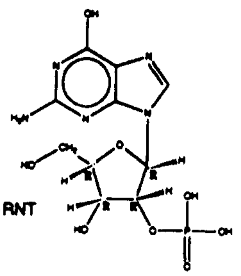

6PAD

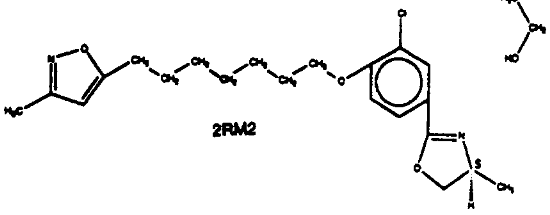

2R04

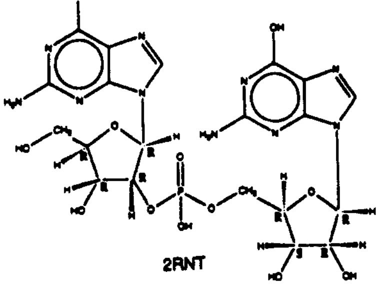

2R06

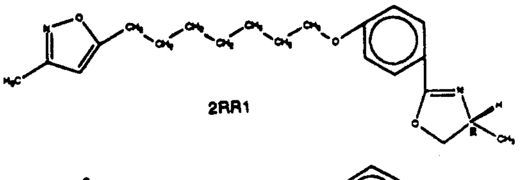

2R07

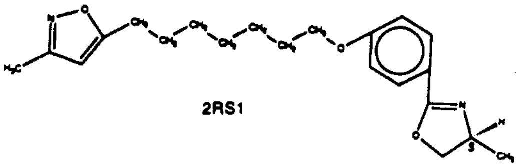

1R08

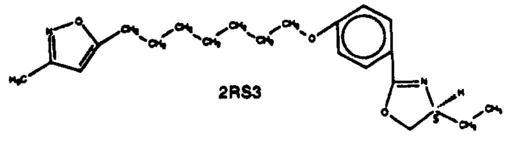

1RNT

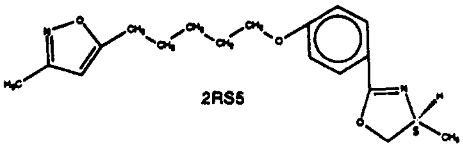

2RM2


2RNT


2RR1


2RS1


2RS3


2RS5

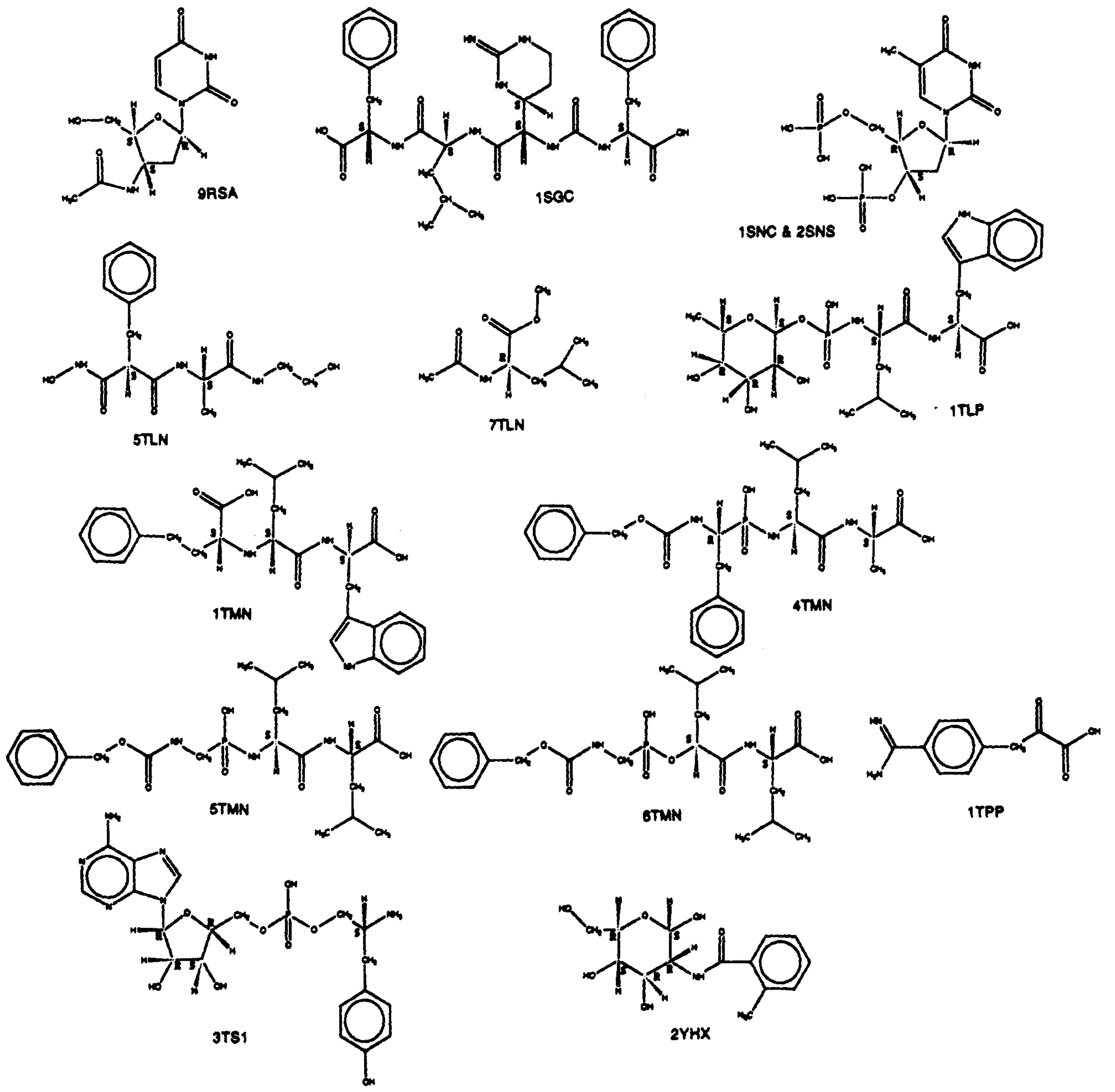


Figure 2. PDB small molecule structures in the second test set.

Once it has identified the units in a molecule, a unit graph is constructed which details the way in which the units are connected. Cobra then attempts to fit the units together following this graph and, if necessary, uses strain resolution algorithms to try to resolve problems in the emerging structure.

Each unit has associated with it all the different, low-energy "subconformations" that it can adopt, and these are now used to construct a 3D model of the structure commencing with the most highly connected unit in the graph. It is an underlying assumption that the conformation adopted by a unit in isolation is also that adopted by this unit when part of a larger molecule.[8] Every subconformation for each unit is tested in turn to build a structure; however, the search algorithm uses a series of rules to follow a minimal cost path and eliminate before testing those subconformations which will lead to a high-energy structure being formed. While each 3D model is being built, checks are carried out to see that no unacceptably close contacts are made and that user-defined restrictions (e.g. value of an interatomic distance between an atom pair) are not breached. The final result is a default or user-specified number of low-energy conformations constructed from the unit subconformations.

ChemDBS-3D, October 1991 Version. The ChemDBS-3D builder uses rule based searching techniques to generate a set of conformations for a structure.[9] SMILES strings are analysed and separated up into cyclic and acyclic portions for building. Ring fragments are stored in a knowledge base, and these are fused together to produce ring systems by using "sensible" rules; the Chem-X 3D sketch algorithm then adds the side chains. ChemDBS-3D cannot build a structure if it contains a cyclic fragment for which it does not have a specific building block stored in the knowledge base. Noncyclic portions are built by using bond lengths and angles from molecular mechanics parameter files with the torsion angles set to 180°.

This procedure is intended to generate a conformation of a structure with regular torsion angles, but this is not intended to be used as a representative of one of the possible conformations that the structure can adopt. Indeed, in some cases, especially for molecules with side chains, branching, etc., it may be necessary to perform conformational analysis after building in order to minimize steric crowding, and move apart atoms in contact after the initial build.[30] The structure is used as a starting point for a conformational generation procedure which explores the conformational space available to the structure.

Conformations are generated by systematically incrementing each rotatable bond in the structure by a predefined set of values. The time taken for any conformational analysis is dependent on the number of rotatable bonds and the step size used about the bonds. In a systematic search the increase in time taken to search around an extra bond is directly proportional to the number of points considered and is geometric.[11] To overcome this problem, a rule based search can be used which applies a set of torsion angle rules to the generated conformations and rejects those which are high-energy structures. Rule based searching results in a linear increase in time with the number of rotatable bonds which is preferable to the systematic search for structures with a large number of rotatable bonds.

Separate energy calculations on each conformation are not included because the torsion angle rules encode low-energy selection criteria. A fast van der Waals (vdW) bump check algorithm has been added to provide a more realistic method for including energy criteria as it should discard any conformation which has overlapping atoms. A full energy calculation would greatly increase the generation time.

These conformations may be used to generate a set of distance based search "keys" which can be stored in a database and used as a screen to filter out large numbers of structures from the database during a search. The keys are set up on the basis of the presence of "centers" and the distances between them in each of the conformations generated during the conformational analysis. The centers that can be used are heteroatom hydrogen bonds, heteroatom hydrogen acceptors, charge centers, and ring centroids, as these are the most common features which occur in a pharmacophoric pattern.

A ChemDBS-3D database can be used to search for pharmacophoric patterns or, as in this study, complete 3D structures. The searching technique follows a series of steps to eliminate the large majority of structures in the database whose conformations cannot match those of the query structure. First, a set of 3D keys are generated for the query structure by using the same technique as that outlined above for the database structures. These 3D keys are compared with those stored for each of the database structures, and those structures that have matching keys are then subjected to a substructure search which imposes atom-by-atom checks of atom type, connectivity, and geometry to see if they match those of the query. Finally the conformations of the database structures are regenerated by using (preferably) the same set of incremental values used to generate the search keys. These conformations are compared with the conformation of the query by using energy calculations and a subgraph isomorphism algorithm, and those database conformations that match to within a given tolerance are then fitted to the query pattern by using all pattern atoms as equally weighted fitting restraints to orient them correctly for display.

In order to compare the conformations produced by ChemDBS-3D with those produced by the other methods, it was necessary to build, key, and search databases created for each of the test sets. Therefore, it is important to appreciate that the performance of ChemDBS-3D may not be directly comparable with those of the other methods which were not allowed to search the entire conformational space and were not used in a database searching environment.

Converter, October 1992, Version 2.0 Beta. Converter uses distance geometry techniques to produce a set of 3D coordinates close to an energy minimum from a 2D chemical sketch.[10] Converter requires either a 2D chemical sketch input using the INSIGHT II Sketch interface, or 2D MOL and SD files from MDL Information Systems Inc. software.[31] In this study, all structures were sketched individually in 2D and then written to a MOL file to facilitate rapid conversion.

During conversion, the atom pairs of the 2D sketch are assigned "distance bounds", upper and lower interatomic distance separation values calculated from published tables of bond lengths. These distance bounds, together with information on the planarity of any sp^2 hybridized centers and chiral centers, are used to produce a "distance geometry description" of the sketch. The distance geometry description of the sketch is input into the distance geometry program DG-II[32] which produces the 3D coordinates. Full details of the algorithm are available,[32] but briefly DG-II takes the complete set of distance bounds and estimates the values of the interatomic distances in the structure by using the given bounds and a random number seed. The atomic coordinates are assigned to the proposed 3D structure such that the distances from each of the atoms are a "best fit" to the randomly generated interatomic distances. The deviation of the coordinates from the distance bounds as well as the chirality constraints are then minimized.

Converter can also be directed in the type of 3D structure it produces by specifying three options for the conversion. Torsion angles in chain regions can be allowed to be either all trans or a mixture of trans and gauche. Nonplanar six-membered rings can be required to adopt any sterically acceptable conformation or restricted to chain conformations only. Finally if there are any chiral centers with undefined chirality, Converter can be asked to produce either one configuration of random stereochemistry or to produce all possible configurations for these chiral centers.

The random assignment of the interatomic distances results in many different conformations of the structure being produced by different runs of the program. Converter can be restricted to producing the same conformation for every run, if required, by presetting the random number seed.

The output from Converter can be written via an SD file to a MACCS-3D database which will enable the structure to be used in 3D searching. This option was not used in this study because the MACCS-3D database system does not utilize the multiple conformations produced by Converter. The multiple conformations produced by Converter could each be written to the MACCS-3D database as separate entries, but the space requirement would make this prohibitive. It was decided that as the performance of the conversion packages is being examined in this study, examination of the multiple conformations produced by Converter is of prime importance.

RESULTS

Searching and Building. *CSD Searching.* The graphical interface to Version 4 of the CSD was used to construct the 2D structures of each of the bound small molecules for those 53 structures which passed the selection criteria discussed earlier. Initially, exact 2D representations of the complete structures were used to search the database, but many did not result in any hits being found. In these cases, the criteria were relaxed to enable substructural hits or structures with slightly different atomic characteristics. The relaxations included generalizing any ring bond types to allow single, double, or aromatic bonding, removing the terminal hydrogens and removing terminal atoms and/or generalizing terminal bond types. Obviously the degree of generalization resulted in some hit structures being different to the required structure.

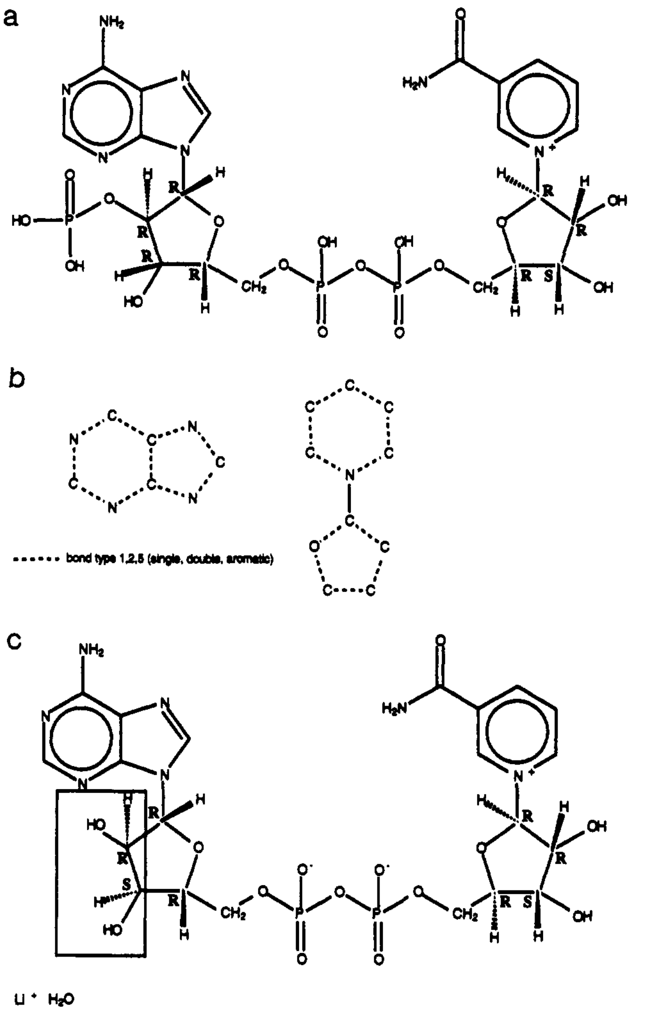


Figure 3. CSD search for NADPH (3DFR) (a) carried out by using two generalized fragments that were required to be in the same hit structure residue (b). In this case the only hit with atomic coordinates was for the structure NADH (c) which proved to have chirality different from NADPH.

Hits were rejected where the query structure was found as a substructure in a much larger structure and also if CHIRON showed that they had different chiral centers to the query structure. Hits could not be used when entries found during the search did not contain any atomic coordinates.

As an example of this procedure, consider the search for NADPH (3DFR) (Figure 3a). To get any hits in a CSD search for NADPH, the query 2D structure used two small fragments from NADPH which had generalized ring bonds and no terminal hydrogens and stated that the fragments must both occur in the same residue in the hit structure (Figure 3b). The search produced four hits, only one of which had atomic coordinates (NADLIH10) (Figure 3c). The CSD hit structure was NADH, and when chirality was examined, it was found that there was one chiral center different from NADPH. The structure was therefore placed in a second test set together with the other 34 structures, which either did not result in any hits despite the degree of generalization or only produced hits that were rejected due to the reasons given above.

Eighteen of the PDB small molecules produced hit entries in the CSD with 2D structures acceptably similar to the bound small molecule and with the same stereochemistry and were placed in the first test set. The conformations of the CSD hit structures were compared with the bound small molecule conformation, and the torsion angle values were used to determine the best match. Some searches produced one CSD hit structure that had a conformation that matched the PDB conformation better than the others in the hit list, but in other searches the choice was arbitrary as the conformations of all the CSD hit structures were very similar. For example there were seven hits for *n*-acetyldeoxythymidine (8RSA) (Figure 4a). None of the hits had exactly the same 2D structure as *n*-acetyldeoxythymidine. Six of the hits were from different studies of the same structure which was more like that of the bound structure than the seventh hit. FIXGAU03 was chosen as the representative structure from these six hits as its conformation matched slightly better than the others (Figure 4b). The CSD entry for FIXGAU03 was found to have two residues in its asymmetric unit, and the torsion angles in the conformation of the residue randomly assigned suffix "B" were found tomatch those in the bound structure better than those in "A".

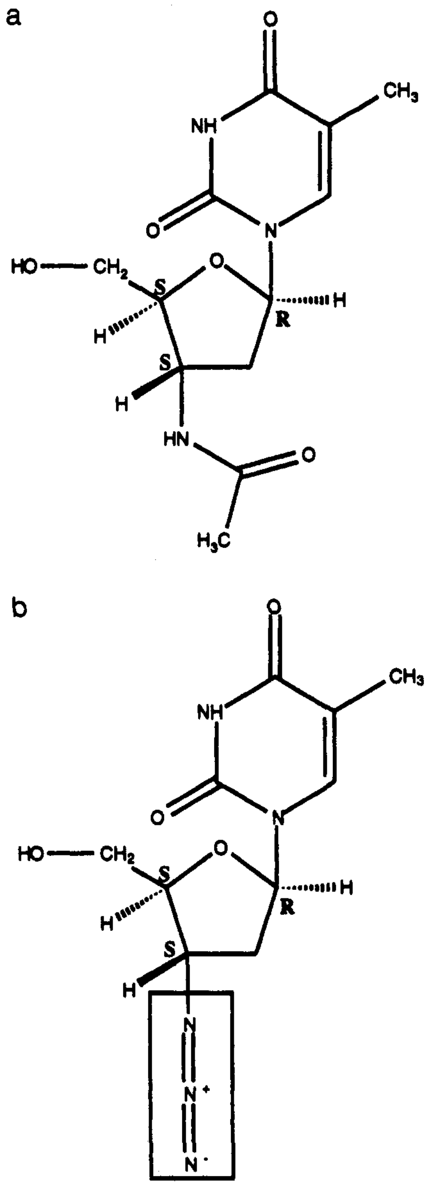


Figure 4. CSD search for acetyldeoxythymidine (8RSA) (a), producing six hits for the same structure from different crystal studies. The conformation that matched best was found to be one of the residues in the asymmetric unit of FIXGAU03 (b).

Table II lists the statistics of all the successful CSD searches and the reasons for rejecting structures. Searches for folate (7DFR) and deazafolate (2DHF) are not listed because, after generalization of their structures, they were found to hit the same structure (DOJZAD01) as methotrexate (3DFRM and 4DFR).

Table II. Results and Analysis for the CSD Searches on the First Test Set

			reasons for rejecting structures				
	query structure generalization	total no. of hits	substructures	no atomic coords	different chirality	acceptable matches in 2D	structure most similar to bonded conformn
2AAT	yes	3	2			1	PYRPOC
8ATC	yes	1				1	COYDUP
1CLA	yes	6	2	1	1	2	CLMPCL02
3CPP	yes	24	12	2	8	2	DMCMPH
3DFRM	yes	3	–	1		2	DOJZAD01
1FCB	yes	6	2		2	2	RIBBAD
2GBP	yes	20	2	2	14	2	GLUCSE01
8LDH	yes	2				2	CITARC and CITRAC10
3PTB	yes	2	1			1	CONYAF10
8RSA	yes	7				7	FIXGAU03
4XIA	no	11			7	4	GLUCIT
5XIA	no	3			1	2	XYLTOL

Concord Building. Input to the program was a SMILES string generated by Daylight GEMINI software via a SYBYL mol file of the PDB bound structure. SMILES strings produced by GEMINI have @ and @ @ chirality information; these were changed to R and S chirality flags by hand, based on chirality information produced by CHIRON from the mol file structure. Concord building was carried out on a VAX 6000-460. Table III shows that Concord built all of the unique structures in the first test set and 30 out of 34 structures in the second test set. Its failure to build four structures was caused by "much too close" contacts in the generated structure. No explanation was given as to how to proceed with alternative building strategies, and there is no opportunity to alter the rules used or input extra information. Close contacts were successfully relieved in four other structures in the first test set and seven in the second test set. The average CPU time taken to build a structure in each test set is shown in Table III.

Cobra Building. Input to the program was the SMILES string generated by Daylight GEMINI. Cobra uses @ and @ @ chirality flags, so the GEMINI produced SMILES string can be used without alteration. The program was asked to produce five conformations for each input structure. Cobra building was carried out on a VAX 6000-460.

Table IV shows that Cobra built all of the unique structures in the first test set but only 9 out of 34 in the second test set. Two structures required intervention to build correctly. Chloramphenicol (1CLA) initially would not build due to difficulties assigning bond orders; this was traced to a problem with the interpretation of the SMILES string. When the order of the elements in the SMILES string was changed and the chirality removed, the structure built correctly. A further problem was found when Cobra built β-D-glucose (2GBP) with chirality different from that of the input SMILES string. The string had to be edited by hand in a trial-and-error manner to produce the correct structure.

Five structures would not build initially, but the program prompted the generation to be repeated with a different fit criterion whose value was suggested, and this time the building was successful. Cobra could not build any unstrained conformations for five structures and so tried to produce strained conformations that would suffice. This procedure uses a lot of CPU time and disk space: in two cases the build exceeded the allowed CPU time, and in a third case the disk space was exceeded.

Other structures failed to build for a variety of reasons. With three structures, the program proceeded as far as template joining and then exited with no error message given. The program crashed when trying to build chymostatin A (1SGC). In two cases the program appeared to "hang" after notifying that it had found a zero distance and finally ran out of CPU time (the maximum CPU time allowed was 6000 CPU s). By far the most common cause of structures not being built correctly in the second test set was the incorrect interpretation of amide nitrogen in the SMILES strings. In a total of 15 structures the amide nitrogen was interpreted as being tetrahedral and chiral. These structures are obviously incorrect and were removed from this study. This problem was referred to the software author, and it is hoped that it will be fixed in future releases.

For those structures that it successfully built, Cobra produced five conformations for all but two structures. It incorrectly produced two configurations for benzamidine (3PTB) and for each configuration produced only two conformations due to the limited flexibility in this structure. Cobra produced four configurations for chloramphenicol (1CLA) as the chirality information was not included in the SMILES string (see above); the five conformations from the configuration with the correct chirality were used in the conformational comparison. The average CPU time taken to build *one* conformation for each structure in the test sets is shown in Table IV.

ChemDBS-3D Building. As described above, database production in ChemDBS-3D is separated into three stages: building, conformational keying, and searching. Keying must be carried out so that the entire conformational space can be sampled and representative conformations identified. A database must be produced to enable searches for the bound structures to be carried out. The 3D structure is built from the same SMILES string as the Concord structure, except that both the square bracket information ([) denoting additional information about an atom and the information denoting cis/trans isomerism (\,/) had to be excluded as ChemDBS-3D does not recognize them. Removal of this information from 15 structures was not found to affect the building. All ChemDBS-3D procedures were carried out on an IBM RS6000 320H.

ChemDBS-3D successfully built all structures in both test sets, but as shown in Table V, two structures required intervention to enable building. Camphor (3CPP) would not build due to a "missing fragment or unobtainable stereochemistry". The problem was traced to the carbonyl group, and the structure built successfully if this group was removed and then added back after the 3D structure was generated. Facilities exist by which a new fragment could be generated for this structure and added to the knowledge base to allow this structure to be recognized and built. An incorrect assignment of the chirality in the sugar ring in cyclic AMP

Table III. CONCORD Building

	successful generation	no. of relieved close contacts	CPU time (s)
		(a) Test Set 1	
2AAT	yes	1	0.22
8ATC	yes	0	0.23
1CLA	yes	0	0.38
3CPP	yes	0	0.47
3DFRM	yes	0	0.42
7DFR	yes	0	0.44
2DHF	yes	0	0.39
1FCB	yes	1	0.42
2GBP	yes	0	0.50
8LDH	yes	3	0.17
3PTB	yes	0	0.11
8RSA	yes	1	0.57
4XIA	yes	0	0.13
5XIA	yes	0	0.14
			av: 0.33
		(b) Test Set 2	
4CTS	yes	0	0.21
3DFR	no	3 and 1 unrelieved	
4ER1	yes	7	1.01
2EST	yes	0	0.29
3GAP	yes	0	0.80
7GCH	yes	0	0.25
3GPD	no	3 and 1 unrelieved	
4PAD	yes	0	0.22
5PAD	yes	0	0.31
6PAD	yes	0	0.30
2R04	yes	0	0.54
2R06	yes	0	0.54
2R07	yes	0	0.57
1R08	yes	0	0.65
2RM2	yes	0	0.67
1RNT	yes	11	0.63
2RNT	no	35 and 1 unrelieved	
2RR1	yes	0	0.65
2RS1	yes	0	0.63
2RS3	yes	0	0.60
2RS5	yes	0	0.59
9RSA	yes	0	0.43
1SGC	yes	0	0.76
1SNC	yes	1	0.47
5TLN	yes	0	0.24
7TLN	yes	0	0.17
1TLP	yes	32	0.90
1TMN	yes	1	0.48
4TMN	yes	2	0.41
5TMN	yes	0	0.33
6TMN	no	8 and 1 unrelieved	
1TPP	yes	0	0.19
3TS1	yes	3	0.63
2YHX	yes	0	0.48
			av: 0.50

(3GAP) was made despite the correct definition in the SMILES string. The chirality was corrected by using Chem-X. Errors were generated during building about the valency of phosphorus and sulfur in all structures that contained phosphate and sulfate groups, but all structures proved to be built successfully. The average CPU time taken to build the structures in the test sets is shown in Table V. The conformations of the initial structures that are built are not intended to be good representations of the conformations that the structures can adopt; the keying stage must be carried out to generate these conformations.

ChemDBS-3D Conformational Keying. The numbers of conformations generated during the keying stage are shown in Table V. ChemDBS-3D will only key structures with 10 or less rotatable bonds and which will result in less than 1 million conformations being generated.

Camphor (3CPP) does not contain any rotatable bonds and only has one center, so it cannot be keyed. However, as it built successfully, the structure that was built initially from the SMILES string remains in the database. Intervention in the keying stage was required for phosphonacetyl-L-aspartate (8ATC) as no conformations were produced by using the default bond increments. When these were relaxed, conformations were generated for the structure. Fifteen structures in the second test set could not be keyed as they exceed either the total number of rotatable bonds allowed by the system or the number of conformations which will be generated exceeds 1 million. The number of conformations that would have to be generated to allow the structures to key are shown in most cases in Table V.

On closer examination of the keying process for the first test set, there was found to be a problem with the assignment of rotatable bonds to the built 3D structures used in the keying and conformational regeneration stages. In 10 out of the 14 unique structures that were keyed, there were missing rotatable bonds (noted in Table V) and this affected the performance of some of the structure searches in the first test set. The problem was found to be due to the bonds around a hydroxyl group being incorrectly rejected as being nonrotatable. The problem was fixed in the version of ChemDBS-3D used with the second test set. The average CPU time taken to key each structure is shown in Table V.

ChemDBS-3D Searching. To find out whether ChemDBS-3D can produce conformations similar to those of the bound structures, databases containing conformational information (as distance keys) of the generated structures must be set up and searched by using the conformations of the bound small molecules as queries. Two ChemDBS-3D databases were set up—the first comprised the 14 unique structures from the first test set and the second the 19 unique structures from the second test set that were successfully keyed. The complete 3D PDB small molecule structures (including hydrogens although these are not used in the search) were used as query structures to search the appropriate database. These searches will be very exact as all centers and center to center distances will need to be matched for a hit to be found.

In all cases a "successful search" was one in which hits were found after the conformation comparison stage and in which the number of hits did not exceed 100—the limit used in this study for further analysis. The default distance tolerances used in the conformation comparison stage of 0.1 Å for bonded and 1:3 distances and 0.5 Å for nonbonded distances were found not to be broad enough for most searches. In the majority of cases the default tolerances had to be relaxed to enable any hits, and a maximum value of 4.5 Å was imposed.

Unfortunately, one effect of increasing the search tolerances is to diminish the effectiveness of the key search, resulting in more structures reaching the substructure search stage and thus increasing the search times; e.g. 10 out of 15 structures in the database passed the key search for Chain 2 of citrate (8LDH). The substructure search, however, was very effective and resulted in all but two of the searches in identification of the correct database structure. The structures for compound I R (2RR1) and compound I(S) (2RS1), as noted earlier, are very similar, and ChemDBS-3D searches could not differentiate between these structures. Table VI shows the results of the individual searches and the tolerances and search criteria used to produce the hits. Only seven structures had hits at the default distance tolerances, and in all other cases the distance tolerances were increased. In several searches, the nonbonded distance tolerance had to be increased to the maximum value to enable any hits to be found, but even with this wide tolerance, hits could not be found for guanylylgua-

Table IV. COBRA Building

	successful generation	comments on building	configurations produced	CPU time (s)
		(a) Test Set 1		
2AAT	yes		1	2.6
8ATC	yes		1	5.2
1CLA	yes	order of elements in SMILES string changed	4	1.95
3CPP	yes	second attempt using suggested fit criterion	1	1.2
3DFRM	yes		1	12.8
7DFR	yes		1	66.8
2DHF	yes		1	62.6
1FCB	yes		1	79.4
2GBP	yes	chirality of SMILES string changed	1	2.2
8LDH	yes		1	6.6
3PTB	yes	only two conformations found for each config	2	2.5
8RSA	yes		1	3.4
4XIA	yes		1	2.4
5XIA	yes		1	2.2
				av: 17.99
		(b) Test Set 2		
4CTS	yes		1	1.8
3DFR	no	exited without reason after processing string		
4ER1	no	odd degree of unsaturation		
2EST	no	amide nitrogen incorrectly assigned	8	
3GAP	yes		1	8.6
7GCH	no	amide nitrogen incorrectly assigned	4	
3GPD	no	exited without reason after processing string		
4PAD	no	amide nitrogen incorrectly assigned	2	
5PAD	no	amide nitrogen incorrectly assigned	8	
6PAD	no	amide nitrogen incorrectly assigned	4	
2R04	yes		1	72.6
2R06	no	exceeded CPU time after zero distance		
2R07	yes	strained conformations produced	1	55.6
1R08	no	strained conformations—exceeded disk space		
2RM2	yes	strained conformations using suggested fit criterion	1	412
1RNT	no	exceeded CPU time		
2RNT	no	exceeded CPU time		
2RR1	yes	second attempt using suggested fit criterion	1	88.4
2RS1	yes	second attempt using suggested fit criterion	1	81.4
2RS3	yes	second attempt using suggested fit criterion	1	89.2
2RS5	no	exceeded CPU time after zero distance		
9RSA	no	amide nitrogen incorrectly assigned	8	
1SGC	no	program crashed		
1SNC	no	amide nitrogen incorrectly assigned	2	
5TLN	no	amide nitrogen incorrectly assigned	8	
7TLN	no	amide nitrogen incorrectly assigned	2	
1TLP	no	exited without reason after processing string	8	
1TMN	no	amide nitrogen incorrectly assigned	8	
4TMN	no	amide nitrogen incorrectly assigned	8	
5TMN	no	amide nitrogen incorrectly assigned	8	
6TMN	no	amide nitrogen incorrectly assigned	4	
1TPP	no	amide nitrogen incorrectly assigned	2	
3TS1	yes		1	38
2YHX	no	amide nitrogen incorrectly assigned	2	
				av: 94.18

nosine (2RNT). In the majority of cases, the increments around a single bond were allowed to relax to enable 60° steps around a bond, as this is the value used to measure a match for a torsion angle. Finally, in many searches a systematic search was used in place of a rule based conformational regeneration; this enables more bonds to be rotated, and conformations will not be rejected due to energy constraints. The search for *n*-acetyl-L-leucyl-L-phenylalanyl trifluoromethyl ketone (7GCH) was abandoned because the length of time taken to try to determine the conditions which resulted in any hits became unacceptably long. The average CPU time taken to search for each PDB structure on an IBM RS6000 320H is shown in Table VI.

Converter Building. Input to the program was a 2D chemical structure sketch of, on the whole, non-hydrogen atoms only. As the chirality of the structures is important in this study, hydrogens were added to the chiral center carbons and the bonds around the center were given explicit directionality by wedge up/wedge down bonds. Explicit hydrogens were also required when sketching noncharged phosphate groups.

Table VII shows that Converter was successful in converting all of the input sketches in both test sets. The program was run five times for each test set, using the different conversion options available. Run 1 restricted chains to trans conformation and nonplanar six-membered rings to chair conformations only. Run 2 allowed the inclusion of gauche angles in chains and other conformations for six-membered rings. Runs 3 and 5 allowed gauche angles but restricted to chain conformations, and run 4 restricted chains to trans angles but allowed any six-membered ring conformation. The average elapsed time taken to build *one* conformation for each structure in the test sets is shown in Table VII.

Conformation Comparison. To compare the methods, further analysis was required to select the best conformation from the list of up to 100 hit conformations from the ChemDBS-3D database searches and from the five confor-

Table V. ChemDBS-3D Structure Generation and Keying

	successful generation	generation CPU, time (s)	no. of conformations accepted in keying	comments	keying CPU time (s)
			(a) Test Set 1		
2AAT	yes	0.53	378		4.32
8ATC	yes	0.20	2520	2 rotatable bonds missed during keying	68.83
1CLA	yes	0.25	12	1 rotatable bonds missed during keying	2.18
3CPP	yes—see text	0.58	0		1.48
3DFRM	yes	0.38	1280	1 rotatable bonds missed during keying	124.37
7DFR	yes	0.35	186624	1 rotatable bonds missed during keying	3317.30
2DHF	yes	0.35	1944	1 rotatable bonds missed during keying	132.20
1FCB	yes	0.42	27	3 rotatable bonds missed during keying	2.69
2GBP	yes	0.25	3		1.97
8LDH	yes	0.18	324	3 rotatable bonds missed during keying	3.16
3PTB	yes	0.16	1		1.90
8RSA	yes	0.42	6	1 rotatable bonds missed during keying	2.37
4XIA	yes	0.22	3	4 rotatable bonds missed during keying	1.78
5XIA	yes	0.18	3	3 rotatable bonds missed during keying	1.76
		av: 0.32			av: 244.42
			(b) Test Set 2		
4CTS	yes	0.27	18		1.76
3DFR	yes	0.82		>10 rotatable bonds	
4ER1	yes	1.16		$>42 \times 10^{12}$ conformations	
2EST	yes	0.33		$>5 \times 10^{6}$ conformations	
3GAP	yes—see text	0.75	1		2.03
7GCH	yes	0.32	5184		22.28
3GPD	yes	0.55		$>7 \times 10^{6}$ conformations	
4PAD	yes	0.29	888		18.58
5PAD	yes	0.36		$>7 \times 10^{8}$ conformations	
6PAD	yes	0.34		$>10 \times 10^{6}$ conformations	
2R04	yes	0.58	8340		60.51
2R06	yes	0.35	1620		9.59
2R07	yes	0.36	810		6.67
1R08	yes	0.50		$>25 \times 10^{6}$ conformations	
2RM2	yes	0.36	4170		43.98
1RNT	yes	0.33	12		2.19
2RNT	yes	0.55	540		35.10
2RR1	yes	0.38	8340		62.19
2RS1	yes	0.37	8340		59.47
2RS3	yes	0.37		>10 rotatable bonds	
2RS5	yes	0.35	1620		9.50
9RSA	yes	0.33	6		2.02
1SGC	yes	0.55		$>235 \times 10^{9}$ conformations	
1SNC	yes	0.85	108		4.81
5TLN	yes	0.29		$>3 \times 10^{6}$ conformations	
7TLN	yes	0.22	19		2.68
1TLP	yes	0.88		$>15 \times 10^{6}$ conformations	
1TMN	yes	0.84		$>204 \times 10^{6}$ conformations	
4TMN	yes	0.43		$>65 \times 10^{8}$ conformations	
5TMN	yes	0.39		$>32 \times 10^{8}$ conformations	
6TMN	yes	0.38		$>48 \times 10^{8}$ conformations	
1TPP	yes	0.21	12		21.99
3TS1	yes	0.44	157464		3784.75
2YHX	yes	0.34	2		2.21
		av: 0.42			av: 125.83

mations generated by Cobra and Converter. This analysis was carried out by eye on the basis of the values of the torsion angles; the conformation with the most torsion angles like those of the PDB conformation was used in the table of results. Although Cobra and ChemDBS-3D produce more than one conformation for each structure, they do not alter ring conformations when searching conformational space. Therefore the RMS fit of a ring is constant in each of the conformations produced for a structure by Cobra and ChemDBS-3D. Converter produces slightly different conformations for rings for each run due to the random element employed in the distance geometry approach. Also, by default different chair conformations will be built for six-membered rings, but other conformations (e.g., boat) can also be produced by setting a variable, as was the case in runs 2 and 4 of Converter. Therefore the RMS fit of each ring system had to be measured.

It was found that the majority of conformations produced by Cobra differed only in the position of the side chain atoms and shared the same backbone torsion angles. This was also true for the ChemDBS-3D hits. The torsion angles of the five conformations that Converter produced were found to be quite different in most cases due to the flexibility of the structures being built.

Complete results of RMS fits and torsion angle comparisons are contained in supplementary tables obtainable from the authors. Table VIII shows an example of one of these tables, the conformational comparisons for the structure of *n*-acetyldeoxythymidine bound to chain B of Ribonuclease A (CB8RSA). The 2D structure of *n*-acetyldeoxythymidine is numbered and has the chiral centers marked. The numbers are used to identify the two rings and the torsion angles that are used to compare the conformation of the bound structure with those from the CSD and structure generation packages.

Table VI. Results of ChemDBS-3D Search

	no. of structures passed key search	no. of structures passed substructure search	conditions of conformation comparison			no. of conformations passed conformation comparison	search CPU time (s)
			rotation (deg)	tolerance	type of search		
2AAT	1	1	60	1.0 + 1.0	rule based	60	19.12
CA8ATC	4	1	60	1.0 + 1.5	systematic	65	283.06
CC8ATC	4	1	60	1.0 + 1.5	systematic	47	281.54
1CLA	9	1	120	1.0 + 1.0	rule based	42	5.22
3CPP	11	1	60	0.1 + 0.5	systematic	1	5.77
3DFRM	6	1	120	1.0 + 4.5	rule based	22	670.27
CA4DFR	6	1	120	1.0 + 4.5	rule based	76	1157.42
CB4DFR	8	1	120	1.0 + 4.5	rule based	96	2197.83
7DFR	5	1	120	1.0 + 2.0	rule based	32	1407.79
CA2DHF	7	1	120	1.0 + 3.5	rule based	8	466.57
CB2DHF	8	1	120	1.0 + 3.5	rule based	4	513.65
CA1FCB	2	1	60	1.0 + 2.0	systematic	1	9.57
CB1FCB	2	1	60	1.0 + 2.0	systematic	1	9.34
1FX1	2	1	60	1.0 + 2.5	systematic	37	12.29
2GBP	9	1	60	0.1 + 0.5	systematic	2	2.41
1GOX	2	1	60	1.0 + 2.5	systematic	33	11.68
C18LDH	9	1	60	1.0 + 1.0	systematic	2	4.09
C28LDH	10	1	60	1.0 + 1.5	systematic	20	6.22
3PTB	9	1	60	0.1 + 0.5	systematic	2	2.26
CA8RSA	6	1	60	0.1 + 0.5	systematic	1	3.63
CB8RSA	8	1	60	1.0 + 1.0	systematic	4	4.02
2TRM	9	1	60	0.1 + 0.5	systematic	2	2.27
CA4XIA	13	1	60	1.0 + 4.5	systematic	2	2.69
CB4XIA	13	1	60	1.0 + 4.5	systematic	2	2.40
CA5XIA	13	1	60	1.0 + 2.0	systematic	2	2.60
CB5XIA	12	1	60	1.0 + 1.5	systematic	2	2.59
							av: 272.55
				(b) Test Set 2			
CA4CTS	16	1	60	0.1 + 0.5	systematic	2	3.36
CB4CTS	17	1	60	1.0 + 1.0	systematic	8	3.62
3DFR							
4ER1							
2EST							
CA3GAP	7	1	60	1.0 + 4.5	systematic	3	3.51
CB3GAP	7	1	60	1.0 + 4.5	systematic	3	3.21
7GCH	see text						
CR3GPD							
CG3GPD							
4PAD	1	1	120	1.0 + 1.0	rule based	100	57.13
5PAD							
6PAD							
2R04	12	1	120	1.0 + 2.0	rule based	100	33.91
2R06	12	1	120	1.0 + 1.0	rule based	100	14.95
2R07	3	1	120	1.0 + 1.0	rule based	100	18.75
1R08							
2RM2	3	1	120	1.0 + 2.5	rule based	100	124.37
1RNT	6	1	60	1.0 + 1.5	systematic	6	9.84
2RNT							
2RR1	12	2	120	1.0 + 2.5	rule based	200	290.61
2RS1	12	2	120	1.0 + 2.5	rule based	200	291.29
2RS3							
2RS5	12	1	120	1.0 + 1.0	rule based	100	28.02
9RSA	12	1	60	1.0 + 1.5	systematic	22	4.8
1SGC							
1SNC	6	1	60	1.0 + 1.0	systematic	4	334.11
2SNS	6	1	60	1.0 + 1.5	systematic	25	341.79
5TLN							
7TLN	19	1	60	1.0 + 1.0	rule based	63	20.2
1TLP							
1TMN							
4TMN							
5TMN							
6TMN							
1TPP	8	1	60	0.1 + 0.5	systematic	12	3.7
3TS1	1	1	120	1.0 + 3.5	rule based	100	117.25
2YHX	16	1	60	1.0 + 2.5	systematic	24	4.94
							av: 81.40

The table shows the individual RMS fits of each of the rings, the values of the torsion angles in the bound structure, and the torsion angle differences between these values and those from the CSD and built structures conformations. ChemDBS-3D and Converter both produce good conformations as the RMS fits of the two rings are ≤0.1 Å in both cases. ChemDBS-3D can be considered to produce the better conformation as all four of its torsion angle differences are <60°.

The conformational comparisons for each of the structures in the two test sets are summarized in Table IX. This table

Table VII. CONVERTER Building

	successful generation	av elapsed time (s)
	(a) Test Set 1	
2AAT	yes	14.2
8ATC	yes	7.8
1CLA	yes	7.4
3CPP	yes	5.4
3DFRM	yes	22.2
7DFR	yes	25.0
2DHF	yes	18.2
1FCB	yes	17.6
2GBP	yes	6.0
8LDH	yes	5.2
3PTB	yes	5.0
8RSA	yes	9.4
4XIA	yes	7.2
5XIA	yes	5.8
		av: 11.17
	(b) Test set 2	
4CTS	yes	9.0
3DFR	yes	45.8
4ER1	yes	111.0
2EST	yes	16.8
3GAP	yes	9.2
7GCH	yes	12.8
3GPD	yes	32.4
4PAD	yes	10.2
5PAD	yes	15.6
6PAD	yes	16.4
2R04	yes	15.0
2R06	yes	11.8
2R07	yes	11.0
1R08	yes	15.4
2RM2	yes	18.4
1RNT	yes	12.0
2RNT	yes	30.4
2RR1	yes	17.4
2RS1	yes	16.6
2RS3	yes	17.0
2RS5	yes	13.4
9RSA	yes	8.6
1SGC	yes	47.8
1SNC	yes	12.2
5TLN	yes	13.0
7TLN	yes	7.4
1TLP	yes	37.4
1TMN	yes	27.8
4TMN	yes	29.8
5TMN	yes	31.2
6TMN	yes	28.2
1TPP	yes	5.6
3TS1	yes	28.4
2YHX	yes	12.0
		av: 21.97

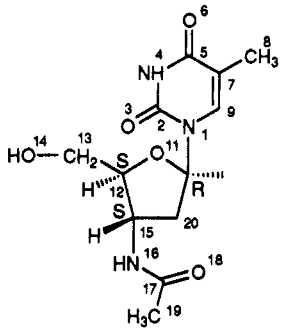
shows the number of rings which had RMS fits of ≤0.1 Å and the number of torsion angles in the conformation that are ≤60° from the torsion angle in the bound molecule. These figures are then used to determine the program which produces the conformation most like that of the bound conformation.

DISCUSSION

The results show that the 2D–3D structure conversion programs can produce conformations which are like the PDB bound conformations in some cases. The nonstandardization in the use of SMILES strings in these programs meant that manual intervention was required to edit the strings before entry into some of the programs. It would be extremely useful if all of these programs could interface directly to the Daylight version of SMILES.

The structures in this study covered a fairly wide range of structural types. In the first set all of the programs built all of the structures, although with intervention in some cases for Cobra and ChemDBS-3D. In the second test set, 4 structures could only be built by Converter and 12 structures could only be built by Concord and Converter. For the majority of these last 16 cases, ChemDBS-3D could build an initial structure but could not proceed through the keying stage. These were considered unsuccessful builds as the conformations of the initially built structures are not intended as good representations of the possible conformations adopted by the structures.

Table VIII. Structure of *n*-acetyldeoxythymidine from Chain B of 8RSA and Table of Results for the Conformational Comparisons with Conformations from the CSD and the Four Structure Generation Packages[a]



RMS1: N1-C2-N4-C5-C7-C9
RMS2: C10-O11-C12-C15-C20
T1: C9-N1-C10-C20
T2: C15-C12-C13-O14
T3: C12-C15-N16-C17
T4: C15-N16-C17-C19

	RMS1 diff	RMS2 diff	T1 diff (deg)	T2 diff (deg)	T3 diff (deg)	T4 diff (deg)
8RSA Chain B			−109	162	−149	173
CSD FIXGAU03 (B)	0.035	0.288	−8	+11	+77	+18
Concord	0.057	0.189	−129	+77	+27	+7
Cobra	0.062	0.251	+50	+24	−89	+7
ChemDBS-3D	0.060	0.044	−7	+15	−1	+7
Converter	0.043	0.040	+11	+134	−26	+7

[a] The top portion of the table shows the main chain torsion angles of the PDB bound structure, identified in the structure diagram as T1–T4. The results give the RMS fits for the two rings and the number of degrees by which the torsion angle values from the CSD and built structures differ from the bound value.

Concord successfully built 92% of the structures. Failures to build were due to the occurrence of "too close contacts" which the program could not rectify, although it relieved close contacts in many structures. It was found that the conformations of structures that had close contacts relieved were sometimes very similar to those of the bound conformations (e.g. 4ER1 and 1RNT) and other times were not similar (e.g. 1TLP and 3TS1). Concord version 3.0 uses energy minimization to relieve close contacts, so this version may be able to build the few structures that could not be built by the version used in this study.

Cobra built approximately 48% of the structures and suffered from a number of different problems. These ranged from program crashed to incorrect interpretation of SMILES strings. The program and manual describe a number of alternative strategies to enable rebuilding to be attempted, but in some cases this did not provide the solution to the problem. Cobra was not designed specifically for use in the database environment and a system would have to be developed which could use the multiple conformations that are generated by the program but which would not result in large storage overheads.

ChemDBS-3D was successful in the interpretation of all of the SMILES strings to build initial structures after intervention in two cases; however, when keying is considered, the performance of ChemDBS-3D reduces to 62% because of the limit on the size of the structure that can be keyed. Default settings do not permit keying of structures with more than 10

Table IX. Summary of Conformational Comparison[a]

PDB small molecule			CSD		CONCORD		COBRA		ChemDBS-3D		CONVERTER		
	no. of rings	no. of torsions	RMS ≤ 0.1 Å	torsion ≤ 60°	RMS ≤ 0.1 Å	torsion ≤ 60°	RMS ≤ 0.1 Å	torsion ≤ 60°	RMS ≤ 0.1 Å	torsion ≤ 60°	RMS ≤ 0.1 Å	torsion ≤ 60°	most similar conformation
(a) Test Set 1													
2AAT	1	4	1	3	1	3	1	2	1	4	1	4	ChemDBS-3D/Converter
CA8ATC	0	7		4		4		4		5		5	ChemDBS-3D/Converter
CC8ATC	0	7		3		5		4		5		5	Concord/ChemDBS-3D/Converter
1CLA	1	6	1	5	1	4	1	4	1	6	1	5	ChemDBS-3D
3CPP	2	2	2	2	2	2	0	2	0	2	2	2	CSD/Concord/Converter
3DFRM	2	10	2	6	1	9	1	6	2	8	2	6	Concord/ChemDBS-3D
CA4DFR	2	10	1	8	1	7	1	8	2	8	1	5	ChemDBS-3D
CB4DFR	2	10	2	8	2	8	1	8	2	8	2	6	CSD/Concord/ChemDBS-3D
7DFR	2	10	2	7	2	7	2	8	2	8	2	8	Cobra/ChemDBS-3D/Converter
CA2DHF	2	10	2	8	2	6	2	7	2	7	2	7	CSD
CB2DHF	2	10	2	8	2	6	2	7	2	6	2	8	CSD/Converter
CA1FCB	1	7	0	5	0	4	0	3	0	3	1	4	CSD/Converter
CB1FCB	1	7	0	5	1	4	1	3	0	3	1	4	CSD/Concord/Converter
1FX1	1	7	0	3	1	5	1	6	0	3	0	6	Cobra
1GOX	1	7	0	5	1	4	1	3	0	5	1	4	CSD/Concord/ChemDBS-3D/ Converter
2GBP	1	1	1	0	0	0	1	1	1	1	1	1	Cobra/ChemDBS-3D/Converter
C18LDH	0	5		4		1		5		3		4	Cobra
C28LDH	0	5		4		3		3		3		4	CSD/Converter
3PTB	1	1	1	1	1	1	1	1	1	1	1	1	all do well
CA8RSA	2	4	1	3	1	3	1	3	2	3	1	3	ChemDBS-3D
CB8RSA	2	4	1	3	1	2	1	3	2	4	2	3	ChemDBS-3D
2TRM	1	1	0	1	0	1	0	1	0	1	0	1	ChemDBS-3D
CA4XIA	0	5		3		4		2		1		3	Concord
CB4XIA	0	5		4		4		2		1		3	CSD/Concord
CA5XIA	0	4		2		3		2		2		3	Concord/Converter
CB5XIA	0	4		2		3		2		0		3	Concord/Converter
total	27	153	19	107	20	103	18	100	20	101	23	108	
(b) Test Set 2[b]													
CA4CTS	0	3				2		3		3		3	Cobra/ChemDBS-3D/Converter
CB4CTS	0	3				2		3		3		3	Cobra/ChemDBS-3D/Converter
3DFR	4	13									1	10	
4ER1	4	24			4	17					4	14	Concord
2EST	1	14			1	11					1	10	Concord
CA3GAP	3	1			1	0	0	1	2	1	2	1	ChemDBS-3D/Converter
CB3GAP	3	1			1	0	0	1	2	1	2	1	ChemDBS-3D/Converter
7GCH	1	11			1	7					1	11	Converter
CR3GPD	4	11									2	6	
CG3GPD	4	11									4	8	Converter[b]
4PAD	1	8			1	7			1	7	1	6	Concord/ChemDBS-3D
5PAD	2	14			2	9					2	8	Concord
6PAD	2	11			2	8					2	9	Converter
2R04	3	10			2	6	2	7	2	8	3	7	ChemDBS-3D/Converter
2R06	3	8			2	4			2	7	3	6	ChemDBS-3D/Converter
2R07	3	8			2	5	2	2	2	8	3	4	ChemDBS-3D
1R08	3	14			2	6					3	9	Converter
2RM2	3	10			2	5	2	4	2	8	3	5	ChemDBS-3D
1RNT	2	4			1	3			2	4	2	4	ChemDBS-3D/Converter
2RNT	4	8									2	5	
2RR1	3	10			2	5	2	6	2	7	3	5	ChemDBS-3D
2RS1	3	10			2	5	2	6	2	7	3	6	ChemDBS-3D/Converter
2RS3	3	11			2	5	2	6			3	6	Converter
2RS5	3	8			2	5			2	7	3	6	ChemDBS-3D/Converter
9RSA	2	4			1	2			2	4	2	4	ChemDBS-3D/Converter
1SGC	3	19			1	15					1	13	Concord
1SNC	2	6			1	6			2	6	2	5	ChemDBS-3D
2SNS	2	6			0	5			2	6	1	5	ChemDBS-3D
5TLN	1	11			1	8					1	8	Concord/Converter
7TLN	0	6				4				5		3	ChemDBS-3D
1TLP	2	12			1	7					1	9	Converter
1TMN	2	14			1	10					1	9	Concord
4TMN	2	16			2	13					2	14	Converter
5TMN	1	16			1	13					1	11	Concord
6TMN	1	16									1	10	
1TPP	1	4			1	3			1	3	1	2	Concord/ChemDBS-3D
3TS1	3	9			2	5	2	5	2	7	2	6	ChemDBS-3D
2YHX	2	4			1	3			1	3	1	3	Concord/ChemDBS-3D/Converter

[a] The number of RMS fits ≤ 0.1 Å and the number of torsion angles ≤ 60° are shown. The estimation of the conformation most similar to that of the bound conformation is based on a comparison of the actual values of the RMS fits and torsion angles. If conformations from different methods are judged to be equally good, they are all listed. [b] In the five conformational comparisons where only Converter produced conformations, the Converter conformation is only considered to be a good match if more than 70% of the torsion angles are ≤60° and the majority of RMS fits are ≤0.1 Å. This is only the case for CG3GPD.

rotatable bonds or that will result in more than 1 million conformations being considered. Keying a structure which exceeds these limitations could dramatically increase the time taken to build a database. For example, the keying stage for tyrosinyl adenylate (3TS1) which has 9 rotatable bonds and which results in the keying of 157 464 conformations takes 3784.75 CPU s (~1 CPU h) on an IBM RS6000 320H.

Converter was successful in building all of the structures in this study, although the method of input was a 2D structure diagram rather than a SMILES string. The Converter method produces random, sterically accessible conformations of a molecule for consideration, and the output can be read directly into the MACCS-3D database system. However this database system does not utilize the multiple conformations that must be produced by Converter to sample the entire conformational space available to a structure. There is no guarantee that the first conformation that Converter produces is necessarily the "best", whether this is considered in terms of the global energy minimum or the conformational geometry.

The comparison of conformations produced by these methods was based on the analysis of the ring and chain regions separately; therefore it is interesting to look at their success in building these regions. Generally the RMS fit of nonflexible rings from all methods was found to be ≤0.1 Å which would indicate that the structure builders successfully predict structures for nonflexible rings. The main exception to this was the fused ring system in flavin mononucleotide (1FCB, 1FX1, 1GOX). ChemDBS-3D could only produce an RMS fit of 0.24–0.29 Å, while the other methods produced RMS fits of <1.5 Å. Only ChemDBS-3D and Converter appear to successfully build flexible rings. For example, in the case of *n*-acetyldeoxythymidine (8RSA) shown in Table VIII, the CSD conformation of the nonflexible pyrimidine ring (RMS1) is most like that of the bound structure, but the RMS fit of the flexible furanose rings (RMS2) produced by ChemDBS-3D and Converter are the only ones to match to around 0.1 Å. This was also found when comparing the RMS fits of the furanose ring in guanylic acid (1RNT) (Figure 5a) and the furan ring in cyclic AMP (3GAP) (Figure 5b), although the chirality of this ring was altered after building in ChemDBS-3D. In the case of the cyclophosphate in cyclic AMP (3GAP), only Converter could build a ring similar to the bound conformation. Generally, the other methods were found to produce RMS fits of between 0.2 and 0.5 Å for flexible rings.

All structure generators were found to have some success in predicting parts of the acyclic regions of the structures. Concord produces fully extended acyclic chains with torsion angles of 180° in most cases, and as many of the chains in the bound structures were not fully extended, this proved to lower the performance of Concord. Cobra builds structures with less extended chains, but their torsion angle values were not found to be similar to those of the bound structure. ChemDBS-3D and Converter also typically build less extended chains; however these proved to be more similar to the bound structure. ChemDBS-3D did not perform well in the searches for sorbitol (4XIA) and xylitol (5XIA) due to the problem of the failure of the program to recognize the rotatable bonds around a hydroxyl group. This resulted in a very restricted conformational space being examined during keying and conformational regeneration. ChemDBS-3D could only produce two conformations for each of these structures, and most of the torsion angles differed from those of the bound structure by over 100°.

The structure generator which produced the best conformation in each case was determined by examining the performance in building both the cyclic and acyclic regions. In some cases the choice of the best structure generator was difficult as one region of the structure matched well but the other did not. For example, CSD and ChemDBS-3D searches produce conformations which have five out of seven of the torsion angles in the acyclic region in flavin mononucleotide (1GOX) within 60° of the bound structure, whereas Concord and Converter have only four out of seven. However, if the RMS fit of the fused ring is considered, Concord and Converter produce a better ring system (RMS fit ≤ 0.1 Å) than CSD and ChemDBS-3D (RMS fit > 0.1 Å). In cases like this, all of the methods are considered to have performed equally well. In the second test set there were four structures which only Converter could build due to program limitations in the other structure generators. One of these structures (3GPD) contained two examples of bound conformations, therefore there were five conformational comparisons for which only Converter output could be examined. In these five cases, Converter was only considered to have produced a good match if ≥70% of the torsion angles were matched and if the RMS fit of the rings were ≤0.1 Å. Table IXb shows that only one of the conformations met these criteria—Chain G of NAD (3GPD).

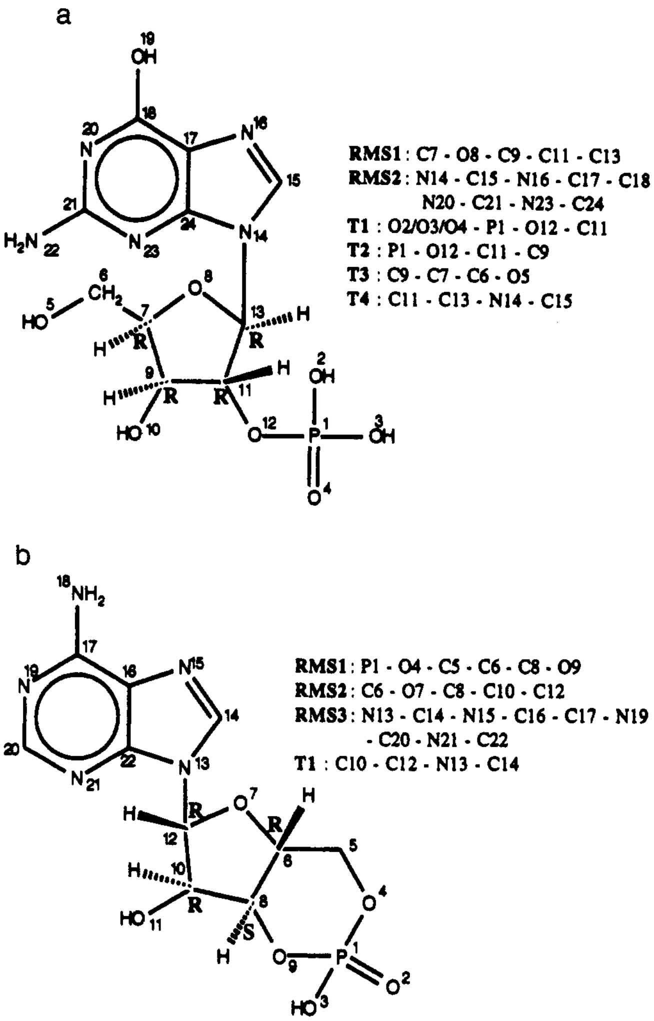

a
RMS1 : C7 - O8 - C9 - C11 - C13
RMS2 : N14 - C15 - N16 - C17 - C18 - N20 - C21 - N23 - C24
T1 : O2/O3/O4 - P1 - O12 - C11
T2 : P1 - O12 - C11 - C9
T3 : C9 - C7 - C6 - O5
T4 : C11 - C13 - N14 - C15
b
RMS1 : P1 - O4 - C5 - C6 - C8 - O9
RMS2 : C6 - O7 - C8 - C10 - C12
RMS3 : N13 - C14 - N15 - C16 - C17 - N19 - C20 - N21 - C22
T1 : C10 - C12 - N13 - C14

Figure 5. Flexible rings built successfully only by ChemDBS-3D and Converter. They produce RMS fits of ≤0.1 Å for the furanose ring (RMS1) in guanylic acid (1RNT) (a) and the furan ring (RMS2) in cyclic AMP (3GAP) (b). Concord and Cobra produce fits of between 0.22 and 0.35 Å. Only Converter produces an RMS fit of 0.1 Å for the cyclophosphate (RMS1) in cyclic AMP (b). The three other methods produce fits between 0.4 and 0.52 Å.

Only the first test set can be used to compare the performance of the structure generators because so many structures could not be generated by Cobra and ChemDBS-3D in the second test set. Looking at the individual conformational comparisons in the first test set, it can be seen that ChemDBS-3D appears to most often produce conformations which are like those of the bound structures (14 out of 26 cases, although in 8 of these cases it only produces a joint best conformation). Converter is the next most successful with 12 out of 26 cases (all 12 are joint best conformations), followed by Concord with 11 out of 26 (10 joint), CSD with 10 out of 26 (9 joint), and Cobra with 5 out of 26 (3 joint). The method used in ChemDBS-3D to produce a conformation searches the entire conformational space available to the structure; therefore it might be expected that it would produce a conformation more like that of the bound structure. Converter and Cobra are only asked to produce five conformations for a structure and therefore do not search the entire conformational space. Despite this it was found that conformations produced by Converter were often very similar to the bound structure. Converter could be considered to perform better than ChemDBS-3D in the second test set as ChemDBS-3D could not key 15 structures due to program limitations.

If the total number of torsion angles and RMS figures for each structure generator are compared, it can be seen that Converter produces the best figures with 23 out of 27 RMS figures ≤0.1 Å and 108 out of 153 torsion angle differences ≤60°. However, these figures are not very different from the other structure generators or CSD, so it could be said that all methods appear to do equally well at producing parts of structures that are similar to those of the bound structures (within the RMS and torsion angle tolerances).

The results for the CSD structure comparisons are interesting. Overall the CSD seems to provide fewer conformations that are similar to the bound conformation than ChemDBS-3D and Converter but almost as many as Concord. Originally it was thought that the main cause of discrepancy with the bound ligands might be due to the amount of generalization that had to be applied to the query structures to enable hits to be found, resulting in CSD hits of different chemical structures. However, this was only found to be true in three cases: camphor (3CPP) (Figure 6a) and DMCMPH (Figure 6b); flavin mononucleotide (1FCB, 1FX1, and GOX) (Figure 6c) and RIBBAD (Figure 6d); *n*-acetyldeoxythymidine (8RSA) (Figure 4a) and FIXGAU03 (Figure 4b). Despite the differences in chemical structure, the majority of the torsion angles in these structures were found to be very similar to those of the PDB structures.

All other CSD hit structures had the same chemical structure as the bound PDB ligand. Some of these had conformations that were similar to the bound ligand and others did not. For example, the CSD conformation for chloramphenicol CLMPCL02 is quite similar to that of the PDB bound one (1CLA), while the CSD structure for methotrexate DOJZAD01 is quite different to that in the PDB (3DFRM, CA4DFR, and CB4DFR). Also some entries had parts of the structure that matched well while other parts did not. For example, as might be expected, the nonflexible cyclic portion of the CSD structure for β-D-glucose GLUCSE01 matched well with that from 2GBP, but the small acyclic region did not. These differences are likely to be due to the conformational changes that occur in the small molecules during binding and to the fact that the molecules are conformationally quite flexible, or it may be a reflection of the accuracy of the coordinates for the ligands in the PDB. It would be interesting to compare the structures of the PDB bound ligands with their counterparts in the CSD, and indeed this comparison is being carried out.[33]

Finally it can be shown that the selection of the "best" conformation from a CSD hit list can be difficult too. For example, the PDB entry for citrate (8LDH) contains two

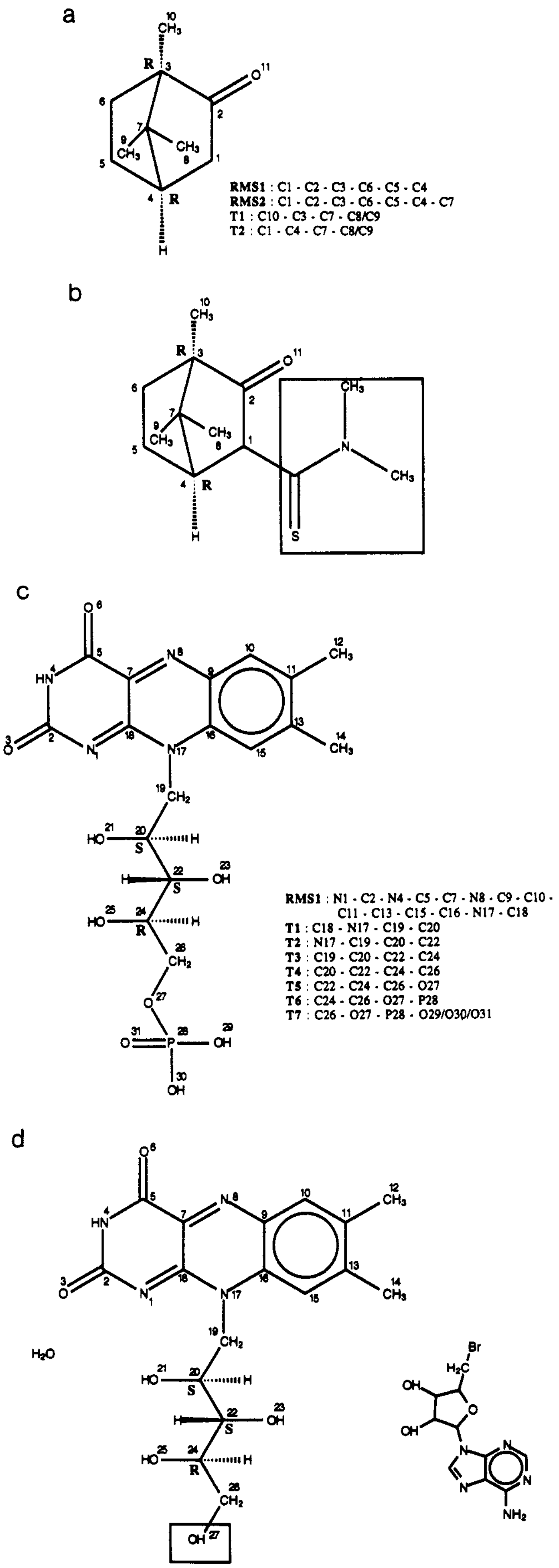

a
RMS1 : C1 - C2 - C3 - C6 - C5 - C4
RMS2 : C1 - C2 - C3 - C6 - C5 - C4 - C7
T1 : C10 - C3 - C7 - C8/C9
T2 : C1 - C4 - C7 - C8/C9
b
c
RMS1 : N1 - C2 - N4 - C5 - C7 - N8 - C9 - C10 - C11 - C13 - C15 - C16 - N17 - C18
T1 : C18 - N17 - C19 - C20
T2 : N17 - C19 - C20 - C22
T3 : C19 - C20 - C22 - C24
T4 : C20 - C22 - C24 - C26
T5 : C22 - C24 - C26 - O27
T6 : C24 - C26 - O27 - P28
T7 : C26 - O27 - P28 - O29/O30/O31
d
H₂O

Figure 6. Generalization of the structure of camphor (3CPP) (a) in the CSD search, resulted in the hit structure DMCMPH (b) which has an extra side chain. Also the search for flavin mononucleotide (1FCB, 1FX1, and 1GOX) (c) produced hit structure RIBBAD (d) which is missing the terminal phosphate group.

small molecules bound to different protein chains. The CSD search produced two hits. When the conformations of these small molecules were compared with the CSD hit structures, it was found that the conformation of chain 1 most closely fitted that of CSD entry CITARC, while chain 2 was more closely approximated by CITRAC10. It would appear therefore that if structures from the CSD are intended for use in molecular modeling studies, all of the hits should be carefully examined.

Although Concord only produces one, low-energy conformation per structure, unlike the other three structure building approaches, the conformations that it produced in this study were usually found to be a good approximation to the bound conformation. Concord is currently the most widely used structure generation package, and as it has been available for some time, it is to be assumed that it will have been tested more robustly than the other packages. These results show that Concord does not perform well with all types of structures but that its building strategies can produce approximate structures for a wider range of structures than Cobra or ChemDBS-3D. It is clear that these approximate structures are in some cases very similar to the bound conformations used in this study, but for other structures they are very different. It is difficult to determine structural commonality in those cases where Concord performs badly as both good and bad matches for acyclic structures and structures with both acyclic and cyclic portions are found. Concord structures were originally intended to be used in conjunction with energy minimization procedures to produce accurate structures,[10] and the use of these procedures would appear to be emphasized by the results of this study.

Cobra did not perform particularly well in this study, either at building the structures or at predicting the bound conformations. It was asked to produce the top five energetically feasible conformations for each structure, and in the majority of cases these conformations differed only in the position of their hydrogen or side group atoms. As the positions of these groups are not considered in this study, this meant that there was, in effect, only one conformation generated for these structures. The only way to produce conformations with different backbone torsion angles, the parts that are examined in this study, would be to remove the hydrogens and the side groups and input a (chemically sensible) structure composed of only the rings and the main backbone.[34] However, it was felt that adopting this procedure would invalidate the comparison of the results from Cobra with those from the other structure generators. An alternative approach would be to generate a large number of conformations with Cobra and group these into clusters by using user-specified features in a clustering algorithm. One member of each of these clusters could then be examined to determine the range of features in the results set. Cobra was not designed to enable any further analysis of results, but this feature may be present in a future release of the program.

For the structures that it successfully keyed, ChemDBS-3D most often produced the conformations that were most like those of the bound conformation. This may be due to its search strategy which allows a large amount of conformational freedom when regenerating conformations to compare with a query conformation. However, in the majority of cases, the search tolerances and criteria used to find the hits were increased from the default values used to generate the keys for the database structures. In this study, the database structures that should provide hits during each ChemDBS-3D search were known and the tolerances could be adjusted accordingly to find the expected hits. If the searches were being used to attempt to suggest novel leads, the tolerances required to find the "correct" hits would not be known. The searches would have to be carried out by using an exhaustive set of tolerances until the user was satisfied that any possible hits were found. Using this philosophy, the user would have to be prepared to handle a large volume of output which would require further analysis before the selection of the best conformation for use in a particular study. Also, Chemical Design recommends that if an in-house database was being built with these structures, it would be advisable to go back and regenerate the database by using these new distance tolerances and criteria, as the current keys do not accurately reflect the conformational freedom of the structures.[30]

These problems might be due in part to the error in atomic positions in crystallographic data which means that the default search tolerances are too stringent for these query structures, especially where there are large center to center distances involved. If crystallographic data are to be included in a database built with ChemDBS-3D, it would appear that broader tolerances than the default should be used in both keying and searching to compensate for experimental error in the coordinates of the structures. Also in this study the hit structures must satisfy all center to center distances in the input structure, which would not be a requirement in normal searches.

The use of a distance geometry approach by Converter would appear to produce conformations similar to those of the bound small molecule in a number of cases. The restriction of output to the first five conformations, in this study, could to some extent have hampered the performance as, unlike Cobra, the conformations produced by each run are very different in terms of the main chain angles and ring conformations. As well as the random number seed resulting in different conformations being produced, the user can also set various options to specify further features of the structure being built. These options should be used when building structures containing the relevant features as some of the best conformations chosen for inclusion in the tables of results in this study come from the runs in which these options were set.

CONCLUSIONS

Many 3D databases are now being constructed by using the structure builders described in this study (FCD-3D, MDDR-3D; the Lederle in-house system, Chemical Abstracts Service, CHDD—Chapman & Hall *Dictionary of Drugs*). This study suggests that no one 2D–3D structure building method consistently produces conformations that are like those of the crystallographically determined bound molecules, although in some cases some of the methods do perform well. Overall there also seems to be no more similarity between the PDB and CSD structures than between the generated and PDB structures. Therefore it appears that neither computer generated nor CSD crystal structures can be used to predict the bound conformation of a ligand without additional conformational investigations. This is not an unexpected conclusion since these methods do not take into account the influence of the protein structure in any way and it is known that the conformation of a ligand does depend on its environment.[35,36] It would be interesting to develop this study further to see if the structure building methods are predicting the unbound (i.e. CSD) conformations any more successfully, although this study and other studies[16,17] do suggest, again, that some of the methods do well, in some cases. Furthermore an initial inspection of the data suggests that there is no correlation between similarly in CSD and PDB structures and the ability of the structure converters to predict the ligand's bound conformation; i.e. the structure generators do not give good results only when the CSD conformation is similar to that of the PDB.

Converter was successful at building 100% of the 3D structures from input 2D structure diagrams. The structures were very similar to the conformations required in a large number of cases even though in this study it was not allowed to sample the entire conformational space of a structure. Presently the program can be used as a stand-alone conversion package outside a database environment or it can be used to produce one conformation that is entered into MACCS-3D, which has announced plans to develop a conformationally flexible database system to examine whether the database structures could match the query.[37] If the strengths of the program are to be utilized, a database would have to be produced which could store and manipulate the multiple conformations produced by the distance geometry approach with limited storage overheads.

ChemDBS-3D successfully interpreted the SMILES strings for all structures, but due to the keying restrictions it could not produce conformations for larger structures and could only be considered to have success in building and keying 62% of the structures. Conformational comparison of the built structures showed that, for the structures that it could key, ChemDBS-3D conformations were the most similar to those of the bound small molecule. To find the best conformer, however, a wide range of database search tolerances and variables had to be used and hit lists of up to 100 structures had to be compared by eye. If the databases set up in this study by ChemDBS-3D were to be used in a working environment, the keys for the structures would have to be regenerated by using the more relaxed variables that produced successful searches.

Concord built 92% of the structures which were often a good approximation of the bound structure, but further minimization of these structures would seem to be required in some cases. Cobra only built 48% of the structures correctly, and the conformations of those structures built successfully were not usually very similar to those of the bound conformation. Further development of this program would seem to be required before it is robust enough to be useful in database applications. Hit structures produced from CSD searches proved to have mixed success and would appear to indicate that structures from those searches resulting in multiple hits should be carefully examined and compared before use in modeling of bound structures.

Upgrades of much of the software used in this study are being produced at a feverish pace which indicates the level of activity and expectation that authors and users are placing in these tools for assistance with molecular modeling problems. Additionally there are new programs becoming available (e.g. Tripos Unity[24] and MACCS-3D flexible search[37]) which offer alternatives to the database programs discussed here.

ACKNOWLEDGMENT

Thanks to Harren Jhoti and Alan Wonacott in the Protein Structure Group at Glaxo, Greenford, U.K., for help with the crystallographic problems; Keith Davies, Dick McCulloch and Malcolm Baird at Chemical Design for the use of their IBM RS6000 machine and their help with ChemDBS-3D; Marvin Waldman for his help with Converter; and Biosym Technologies, Inc., for the use of their hardware.

REFERENCES AND NOTES

(1) Jakes, S. E.; Willett, P. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures. Selection of Interatomic Distance Screens, *J. Mol. Graphics* **1986**, *4*, 12–20.

(2) Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. *Concord, A Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures*; The University of Texas at Austin and Tripos Associates: St. Louis, MO, 1988.

(3) Pearlman, R. S. Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Aut. News* **1987**, *2*, 1–6.

(4) Rusinko, A., III; Sheridan, R. P.; Nilakantan, R.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. Using Concord to Construct a Large Database of Three-Dimensional Coordinates from Connection Tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251–255.

(5) Güner, O. F.; Henry, D. R.; Pearlman, R. S. Use of Flexible Queries for Searching Conformationally Flexible Molecules in Databases of Three-Dimensional Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 101–109.

(6) Leach, A. R.; Prout, K.; Dolata, D. P. Automated Conformational Analysis: Algorithms for the Efficient Construction of Low-energy Conformations. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 271–282.

(7) Leach, A. R.; Dolata, D. P.; Prout, K. Automated Conformational Analysis: Algorithms for Molecular Perception. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 316–324.

(8) Leach, A. R.; Prout, K. Automated Conformational Analysis: Directed Conformational Search Using the A* Algorithm. *J. Comput. Chem.* **1990**, *11*, 1193–1205.

(9) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312–316.

(10) *Converter. Version 2.0 beta*; Biosym Technologies Inc: San Diego, CA, 1992.

(11) Davies, K.; Upton, R. Experiences Building and Searching the Chapman & Hall Dictionary of Drugs. *Tetrahedron Computing Methodology* **1990**, *3*, 665–671.

(12) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. In *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; Data Commission of the International Union of Crystallography: Bonn/Cambridge/Chester, 1987; Chapter 2.4, pp 107–132.

(13) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535–542.

(14) 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. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187–204.

(15) Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. The Cambridge Crystal Data Centre: Computer-Based Search, Retrieval, Analysis, and Display of Information. *Acta Crystallogr. Sect B: Struct., Crystallogr. Cryst. Chem.* **1979**, *B35*, 2331–2339.

(16) Hendrickson, M. A.; Nicklaus, M. C.; Milne, G. W. A.; Zaharevitz, D. CONCORD and CAMBRIDGE: Comparison of Computer-Generated Chemical Structures with X-ray Crystallographic Data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 155–163.

(17) Nicklaus, M. C.; Milne, G. W. A.; Zaharevitz, D. Chem.-X and CAMBRIDGE: Comparison of Computer Generated Chemical Structures with X-ray Crystallographic Data. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 639–648.

(18) *Fine Chemicals Directory 1990–1991*; Aldrich Chemical Co.: Gillingham, U.K., 1990.

(19) Thornton, J. M.; McArthur, M. W.; Smith, D. K.; Gardner, S. P.; Hutchinson, E. G.; Morris, A. L.; Sibanda, B. L. In *Accurary and Reliability of Macromolecular Crystal Structures. Proceedings of the CCP4 Study Weekend*, 26–27 Jan 1990; Henrick, K., Moss, D. S., Tickle, I. J., Compilers; Science & Engineering Research Council, Daresbury Laboratory: Daresbury, U.K. 1990; pp 39–52.

(20) Creighton, T. E. *Proteins: Structure and Molecular Principles*; W. H. Freeman and Co.: New York, 1983.

(21) Glusker, J. P.; Trueblood, K. N. *Crystal Structure Analysis: A Primer*, 2nd ed.; Oxford University Press: Oxford, U.K., 1985.

(22) Schweizer, W. B. In *Crystallographic Computing: Data Collection Structure Determination, Proteins, Databases*, Papers of the International Summer School, 9th Meeting; Sheldrick, G. M.; Krueger, C., Goddard, R., Eds.; OUP: Oxford, U.K., 1984; pp 119–127.

(23) Martin, Y. C.; Bures, M. G.; Willett, P. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers Inc.: New York, 1990; Chapter 6, pp 213–263.

(24) *SYBYL Molecular Modelling Software, Version 5.4*, Tripos Associates: St. Louis, MO, 1991.

(25) *INSIGHT II Program, Version 2.2.0 Beta*; Biosym Technolgies Inc: San Diego, CA, 1992.

(26) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(27) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(28) *GEMINI. Daylight Software, Version 3.6.3*; Daylight Chemical Information Systems: New Orleans, LA, 1992.

(29) *CHIRON Program, Version 4.2*; Stephen Hanessian, University of Montreal: Montreal, Canada, 1992.

(30) *ChemDBS-3D Workshop. Course Notes*; Chemical Design Ltd.: Oxford, England, 1991.

(31) *Molecular Design Software*; MDL Information Systems Inc.: San Leandro, CA.

(32) Havel, T. F. An Evaluation of Computational Strategies for Use in the Determination of Protein Structure from Distance Constraints Obtained by Nuclear Magnetic Resonance. *Prog. Mol. Biol. Biophys.* **1991**, *56*, 43–78.

(33) Nicklaus, M. C. Personal communication.

(34) Leach, A. Personal communication.

(35) Behling, R. W.; Yamane, T.; Navon, G.; Jelinski, L. W. Conformation of Acetylcholine Bound to the Nicotinic Acetylcholine Receptor. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 6721–6725.

(36) Fisher, C. L.; Roberts, V. A.; Hagler, A. T. Influence of Environment on the Antifolate Drug Trimethoprim: Energy Minimization Studies. *Biochemistry* **1991**, *30*, 3518–3526.

(37) *Molecular Connection-the Molecular Design Ltd. Newsletter*; MDL Information Systems Inc.: San Leandro, CA, 1992; Vol. 11.