BALI: Automatic Assignment of Bond and Atom Types for Protein Ligands in the Brookhaven Protein Databank

M. Hendlich,* F. Rippmann, and G. Barnickel

Preclinical Research, Merck KGaA, 64271 Darmstadt, Germany

Received July 30, 1996[⊗]

A method for assigning hybridization states and bond orders to protein ligands from the Brookhaven Protein Databank[1] is presented. The assignment procedure is based on the recognition of simple chemical groups and on an analysis of bond length and bond angle data. Special care was taken on the proper handling of aromatic heterocyclic ring systems which are frequently found in small molecule ligands. Hybridization states and bond orders are assigned with a success rate significantly higher than what has been reported for previously published algorithms. The main application of BALI is to identify protein ligands in the Brookhaven Protein Databank and to add information about bond and atom types with a minimum amount of manual intervention.

1. INTRODUCTION

With the recent explosion of high resolution protein structures deposited in the Brookhaven Protein Databank (PDB)[1] a comprehensive analysis of receptor/ligand complexes becomes possible. In order to provide an easy and selective access to the structural information about receptor/ligand complexes in the PDB we have developed a comprehensive database system for receptor/ligand complexes[2] which combines heterogenic information, e.g., crystallographic information from the PDB, biomolecular information like sequence alignments or information about mutations from various other data sources. This database allows complex queries (including substructure searches, similarity searches, and searches for specific interactions) regarding both small molecule and protein aspects.

A key issue in the development and for the maintainance of this receptor/ligand database is the determination of bonding orders and hybridization types of small molecule ligands as this information is not stored in PDB files. Knowledge of correct bond orders and atom types is essential for structure-oriented database queries and an accurate analysis of receptor ligand interactions.

During the processing of the protein ligands from the PDB it became apparent that for databases of thousands of chemical structures a manual assignment is impractical and an automatic typing procedure is needed. Several attempts have been made to address this problem in recent years. Meng and Lewis[3] have developed the IDATM program which calculates hybridization states for molecules from the Cambridge Structure Database[4] (CSD). The IDATM program uses bond lengths and valence angles to derive the correct types. Bond orders are not determined. A similar approach was used by Baber and Hodgkin[5] for calculating bond orders for molecules from the CSD. The authors reported that their program often mistypes conjugated ring systems which are frequently found in protein ligands. We also tested the BONDAGE program which was originally part of the DGEOM95[6] package but achieved rather unsatisfactory results (data not shown).

To facilitate the processing of the protein ligands for storage in our receptor/ligand database with a minimum of human intervention we decided to develop the program BALI which is able to identify protein ligands in PDB files in a flexible way and which derives bonding orders and hybridizations from the Cartesian coordinates. Our approach combines the basic strategy of the IDATM program with the recognition of a broader spectrum of functional groups and a ring perception algorithm for the detection of aromatic rings. Special care was taken to handle the assignment of alternating single/double bonds in conjugated ring systems properly. Using a sample of molecules from the CSD we show that BALI assigns bond orders and hybridizations with high precision and performs significantly better than the programs from Meng and Lewis[3] as well as Baber and Hodgkin.[5]

2. METHODS

Extraction of Protein Ligands from the PDB. Due to the unstructured format of the PDB data files the identification of protein ligands is not straightforward. To populate the receptor/ligand database we have developed a parser which scans the entire PDB for protein ligands such as inhibitors, substrates, co-factors, or prosthetic groups. Several strategies are used to identify ligands. Small molecules and metals are identified by searching PDBfiles for HETATM records. Molecules with less than six heavy atoms (e.g., water, sulfates, and phosphates) are ignored. Molecules which are covalently linked to a protein are only extracted if they are explicitly mentioned in the header of the PDB file to form a complex with the receptor. Nucleic acids are considered as ligands if they are in spatial contact with a protein.

The identification of peptidic ligands is more challenging. The discrimination between peptidic ligands and protein chains is difficult as peptidic ligands are not marked as such. Therefore, the parser extracts all chains with chain identifiers "I" and "J" and all short peptides up to 15 residues with a unique chain-identifier and which form a complex with a larger protein (>100 residues). As both criteria are rather arbitrary a manual check of the corresponding PDB-file is necessary for each extracted peptide.

[⊗] Abstract published in *Advance ACS Abstracts*, February 1, 1997.

S0095-2338(96)00348-4 CCC: $14.00

A total of 3014 protein ligands has been extracted from current release of the PDB (February 1996). Due to many duplicates (e.g., more than 300 heme groups) only 866 different compounds have been found. A detailed presentation of the receptor/ligand database and the types of protein ligands in the PDB will be presented elsewhere.

Determination of Atomic Connectivity. To determine the connectivity of protein ligands the CONECT-records from the PDB-files are used. The reason for this approach is the relatively low resolution of many structures in the PDB resulting in steric collisions of not covalently bonded atoms. Furthermore for some receptors different cocrystallized ligands have been stored in a single PDB-file. In these cases the calculation of the connectivity from interatomic distances would result in a large number of incorrect bonds.

In a number of PDB-files errors in the CONECT-records have been detected after visual inspection of the protein ligands. This visual inspection was done by producing 2D depictions of all ligands with the PRADO program of Daylight.[7] All errors have been corrected manually by changing the corresponding CONECT-records.

If CONECT-records for nonmetal atoms are missing, pairs of atoms, whose interatomic distance is less than the sum of their covalent radii plus a tolerance of 0.45 Å[3] are considered to be bonded. The atomic connectivity of the 20 standard amino acids is derived from the atom names. Covalently connected monomers are combined to polymers. All bonds to metals are ignored.

Assignment of Bonding Orders and Atom Types. The computation of bonding orders and hybridization types from the atomic coordinates is complicated by the fact that in most PDB-files hydrogen atoms are missing. This makes it difficult to discriminate between, e.g., hydroxy-oxygens and carbonyl oxygens just by analyzing bond length data. Even in high resolution X-ray structures of small molecules the distance distribution of single and double bonded oxygens overlap.[3] The low resolution of many structures in the PDB and the fact that many protein ligands in the PDB are highly distorted further complicates the calculation of bond orders and hybridizations from the coordinates. Another critical problem is to handle aromaticity properly. Several different approaches are currently in use:

·Aromaticity is defined by Hückel's $4n + 2$ π electrons rule (e.g., in the SMILES[8] system). All bonds of a ring system are typed as aromatic.

·Aromaticity is limited to six-membered rings (e.g., as in the CSD or in the MOL2-format of the SYBYL Program from Tripos[9]).

·In the alternating single/double bond concept of the Chemical Abstracts Service or the Beilstein Database aromaticity is more or less ignored. In terms of computation this is the most demanding definition because it requires the exact assignment of alternating single/double bonds which can be difficult in conjugated heterocyclic ring systems.

One aim in the development of the receptor ligand database is to cross-reference the ligands in the PDB with other small molecule databases like the Beilstein Database. For this reason we decided to use the alternating single/double bond concept. Furthermore it is relatively easy to regenerate the other two definitions of aromaticity from the single/double bond concept by simply setting ring atoms and bonds to the appropriate aromatic types. The assignment of bond and atom types is done in several steps starting with the most reliable features of a given structure:

·**Typing of atoms with filled valencies:** In the first phase the program types atoms which are fully characterized by their number of connections to other atoms. Bonds to, e.g., hydrogens or halogens must be single bonds, carbons with four covalent bonds and oxygens with two covalent bonds must be sp^3 hybridized, and all bonds must be single bonds.

·**Recognition of functional groups:** The second assignment step is based on the recognition of simple functional groups and is therefore also independent of geometrical data. This has the advantage that many wrong assignments caused by erroneous and ambiguous geometrical data can be avoided. Typical examples are amides where in many cases the nitrogen−carbon bond is more clearly a double bond than the oxygen−carbon bond. The groups recognized by this procedure comprise carboxyl groups, guanidinium groups, nitro groups, imidines, amides, phosphates, and sulfoxides. The appropriate bond and hybridization types are assigned automatically.

·**Assignments derived from geometrical information:** For all atoms and bonds which have not been typed during the previous stages, bond orders are derived from bond lengths and hybridization states from mean valence angles. Ideal values for valence angles, bond lengths, and the threshold values which separate, e.g., single from double bonds were taken from ref 5.

·**Recognition of aromatic rings:** The correct typing of conjugated polycyclic ring systems which are frequently found in protein ligands has been a major problem in previously published algorithms.[5] For this reason we have implemented an algorithm for the perception of aromatic rings. A ring is classified as aromatic if it is planar and obeys Hueckel's rule of $4n + 2$ π-electrons. A ring is defined as planar if the mean ring torsion angle is lower than 7.5° for five-membered rings and 12.0° for larger rings. All bonds within the ring are typed as aromatic and all hybridization states as sp^2. The explicit assignment of alternating single/double bonds is done in a later stage (see below).

·**Resolution of conflicts:** Conflicting assignments occur because of the fact that bond orders and hybridization states are determined from bond lengths and mean bond angles independently. An example for such a conflict are atoms where the mean valence angle indicates a hybridization of sp^2, but all bond lengths are in a distance range which is typical for single bonds. Conflicting assignments are resolved using three criteria. (1) For atoms with three or more connections, mean valence angles are generally more reliable than bond length data. Therefore, bond orders with the highest deviation to the optimal values are changed to be in agreement with the information from the mean valence angles. (2) For atoms with one or two connections the hybridization states are simply retyped according to the bond orders. (3) If the number of bonds to an atom is larger than the valence of that atom the order of the double or triple bond with highest relative deviation to the optimal value is reduced. This process is repeated until all conflicts are removed or a predefined number of attempts has been exceeded. Furthermore all cases of conflicting assignments are reported for a subsequent manual inspection.

·**Assignment of alternating single/double bonds:** During the last stage of the assignment procedure alternating single/double bonds are assigned to networks of connected sp^2-hybridized atoms (e.g., aromatic ring systems detected during one of the previous stages). As the correct placement of

Table 1. Percentage of Compounds with Fully Correct Assignments for 91 Structures from the CSD and 120 Protein Ligands from the PDB[a]

	BALI	Meng and Lewis[3]	Baber and Hodgkin[5]
CSD (with hydrogens)	96.7%		83.5%
CSD (without hydrogens)	95.6%	80.2%	
PDB	90.8%		

[a] All assignments were checked manually. Whenever present counter ions and small solvent molecules were removed.

single/double bonds can be difficult for complex heterocycles (e.g., flavines or porphyrins) we generate all possible assignments of alternating single/double bonds. Each assignment is ranked using a cost function. The terms of this cost function evaluate the deviation from ideal bond length and angle values for each given single/double bond assignment. Other terms of the cost function penalize sp^2-hybridized carbons with no double bond assigned and bonds which connect two rings that have not been typed as double bonds. The highest ranking assignment is accepted.

Coordinates, hybridization states, connectivity, and bonding orders are stored in the MOL2-format of TRIPOS which is a quasi standard for the exchange of small molecule data and is readable by many molecular modeling programs. The conversion into the MOL2 format requires again a recognition of functional groups (carboxyl groups, guanidinium groups, sulfoxides, amides, and planar nitrogens) as many atom type definitions in the MOL2-specification depend on the neighboring atoms. A nitrogen in an amide group must be typed as "N.am" or a carbon in a guanidinium group as "C.cat". The appropriate MOL2 bond and hybridization types are assigned automatically. Furthermore a SMILES[8] string is produced for each ligand using the programming toolkit of Daylight. BALI was written in ANSI C and implemented on a Silicon Graphics Indigo II workstation.

3. RESULTS AND DISCUSSION

To assess the performance of BALI we used 120 randomly selected protein ligands from the PDB and 91 compounds from the CSD. The set of protein ligands comprises various types of compounds including peptidic ligands, small mol-

Table 2. Performance for 91 Compounds from the CSD[4] and Results for the Programs from Meng and Lewis[3] (M&L), Barber and Hodgkin[4] (B&H), and for BALI[a]

refcode	*R*-factor	atoms	M&L	B&H	BALI	refcode	*R*-factor	atoms	M&L	B&H	BALI
AAGAGG10	0.0373	27				NAHACA	0.0440	8			
AAGGAG10	0.0445	27				NAHACB	0.0610	8			
ABHPTB	0.0850	23				NAPMYC10	0.0700	30			(*)
ABHPTB	0.0850	23	*	*		NBPENC	0.0414	22			
ACANOB	0.0520	24				NDMSCN	0.0360	30			
ACFUCN	0.0430	14				NEBULR	0.0730	18			
ACIGRA	0.1090	45				NETRSN	0.0670	31		*	
ACMBPN	0.0400	24				NIGERI	0.0500	51			
ACMPXC	0.0510	24				NIVBIO	0.1400	44	*		
ACTBOL		21				NONACS	0.1720	52			
ACTBOL		21				NONACT	0.1030	52			
ACTBOL		21				NONACU	0.0460	52			
ACTBOL		21				NONAMT	0.1080	52			
ACTDGU10	0.0940	19				NONKCS	0.1251	52			
ACTDGU10	0.0940	19	*			NOSHEP10	0.1800	82	*	*	
AERMYC1O	0.1090	53	*		*	OTETCB	0.1170	33	*	*	
AGNGEC11	0.0650	51				OXERTH	0.0770	58	*		
AMCILL	0.1060	24				OXOFMB	0.0430	20			
AMDMCN	0.0620	14				OXTETD	0.0720	33	*		
AMICET10	0.0600	44				OXTETK	0.0800	33	*		
AMOXCT	0.0610	25				OXYTET	0.0600	33			
AMPIAB10	0.1370	69		*		OXYTET01	0.0540	33			
ANFLCN	0.0870	59				PEANAG	0.1500	66			
ANSMYC10	0.1050	23		*		PEANNA	0.1400	66			
ANTBPE	0.1070	34				PENTBH10	0.0770	18			
ANTBRN	0.1070	75	*	*		PILLBL10	0.1330	29			
ANTETC	0.0620	30	*	*		PILLMA	0.0380	39		*	
ANTINA	0.0930	75				PIPBCX	0.0500	33			
ANTMYC01	0.0480	24				PIPCIL	0.0500	36			
ANTMYC03	0.0500	24				PMEPEN	0.1300	24	*		
ANTROS01	0.0650	60				PODACE	0.0560	24			
ANTSUL10	0.0866	22				PRMARI	0.1800	62	*	*	*
AOTETC	0.0760	39	*			PRMESA	0.0760	17			
APLASM	0.0840	55				PROMYC10	0.1300	16	*	*	
APOMRC	0.0450	20		*		PROMYC10	0.1300	7			
APOMRC	0.0450	20		*		PROMYC10	0.1300	84			
APYMPR	0.0480	17	*	*		PROMYC10	0.1300	84	*		
AZPCOH	0.0720	8				PRPENG	0.0730	17	*		
BAMLIK	0.0530	38				PRPENG	0.0730	23			
BEVJER10	0.0480	24			*	PRTYLD	0.1310	28			
CIMMUG	0.0610	22				PURMYC10	0.0550	34		*	
CIMNAN	0.0620	21				PXMPEN	0.0540	23			
HAZMOR	0.1340	24				TYBUCBD01	0.0550	20			
MORPHC	0.0460	21				TBUCBD02	0.0480	20			
MORPHI	0.1300	21				TBUCBD10	0.0460	20			
MORPHM	0.0450	21									

[a] Compounds with at least one wrong assignment are marked with an "*". The error in NAPMYC10 occurred only when all hydrogens have been removed.

Table 3. Performance for 120 Protein Ligands from the PDB[a]

ligand	error	ligand	error	ligand	error
2GP+108+pdb1fut		2GP+305-B+pdb1trp		3GP+105+pdg1rls	
478+200+pdb1hpv		4PB+600+pdb1tsi		A-C-A-C-G-T-G-T+1-A+pdb193d	
A-C-G-C-+1-D+pdb1hvo		A-G-A-T-A-A-A-C+106-B+pdb1gau		ADP+205+pdb1uky	
AGF+309+pdb8cpa		AHG+3-A+pdb1fpf		AHM+336-A+pdb1fbf	
ALA-ALA-PRO-BNO+4-P+pdb1p05		AMP+337-A+pdb1fbp		AMP+338-C+pdb4fbp	*
AMP+4-B+pdb1fpe		AMP+468-1+pdb1gph		AMP+471+pdb1lgr	
AMP+920-C+pdb7gpb		ARG-PRO-ASP-PHE-CYS-LEU-GLU-PRO-PRO-TYR-+1-l+pdb4tpi		ASP+1-S+pdb1wat	
ATP+1+pdb1atp		ATP+154-B+pdb4at1		AXP+500+pdb1inw	
BAR+201+pdb2cht		BAR+212+pdb2cht		BME+16+pdb1hsn	
BOC-PHE-HIS-CAL-LYS+1-l+pdb3er3		BOG+501+pdb1tcc		BOG+502+pdb1tcb	
BRB+1004+pdb1hld		BTN+150-B+pdb2avi		BTN+500+pdb1bib	*
CAM+2+pdb4cp4		CFM+496-B+pdb1mio		CIT+2+pdb2cts	
CMS+404-A+pdb1chm		CTP+999-B+pdb1rac		CTP+999-D+pdb1rah	
EDR+999+pdb1ack		ETR+176+pdblerb	*	FAD+395+pdb1pbf	
FAD+395+pdb1pxa		FBP+2+pdb2ldb		FBP+323-A+pdb1pfk	
FK5+108+pdb2fke		FMN+322+pdb2pia		G-A-A-A-G-C-C-A-T-TH-+1-A+pdb1ahd	
G-A-T-A-T-C+1-C+pdb2da8		G-G-U-A-G-G-G-G-G-G+42-T+pdb1ser		G3P+1+pdb1gle	
GAL+104-E+pdb1lta		GAL+104-F+pdb1lta		GCO+400-B+pdb1xlf	
GDN+218-D+pdb1hnc		GDP+1-A+pdb1efg		GDP+200+pdb1crp	
GDP+200+pdb1crr		GEL+935+pdb1poe		GLA+307+pdb5abp	
GLU+471-B+pdb2lgs		GOL+10+pdb1rtm		GPS+218-I+pdb2gst	
GSP+355+pdb1gia		HAP+280+pdb1hfc		HEM+1-B+pdb1hho	
HEM+1-C+pdb1hgc		HEM+142-A+pdb1cmy		HEM+154+pdb1hsy	
HEM+154+pdb1mln		HEM+154+pdb2mm1		HEM+154-B+pdb1myj	
HYA+960-A+pdb1gyi		HYA+970-B+pdb1gyi		ICT+1+pdb1ikb	
INH+256+pdb1srt		IVA-VAL-VAL-LTA+4-l+pdb1apt		KET+412+pdb1map	
LDA+616+pdb1prc		LPM+639+pdb1eab		LTR+81-N+pdb1wap	
MAL+312-A+pdb1at1		MAL+690+pdb1cgv		MAL+702+pdb1csc	
MES+434+pdb1dgd		MES+434+pdb1dge		MMA+4+pdb1lob	
MTX+361-B+pdb3drc		N1T+681-A+pdb1tkb		NBE+2+pdb1h7	*
NDP+2-A+pdb8cat	(*)	NOJ+480+pdb1dog		OLI+200-C+pdb1pmp	*
OTE+545+pdb3por		OXM+402-B+pdb9ldt		PGH+249-A+pdb7tim	
PGH+250-2+pdb1tpb		PGH+250-A+pdb1tpw		PIM+422+pdb1phf	
PLP+258+pdb1asb		PMP+411-A+pdb1aia	*	PMP+411-A+pdb1aic	*
PMP+413+pdb1amr		PQQ+101-C+pdb3aah		PRE+224+pdb1com	
REA+178+pdb1fem		RET+134-C+pdb1opb	*	SEO-SEO+901+pdb141l	
SEO-SEO+901+pdb1l91		SIA+1-E+pdb4hmg		T-C-A-A-C-A-G-C-T-G-+1-G+pdb1mdy	
T-G-T-C-A-G-T-T-A-G-+22-B+pdb1msf		TSA+1+pdb1fig		U10+10+pdb1yst	
U10+502+pdb1pcr	*	UDP+353+pdb1bgu		UFP+529+pdb2tdd	*
UVC+1pdb6rsa		XBP+476-A+pdb1rsc		XYL+390+pdb2xis	

[a] Compounds with at least one wrong assignment are marked with a "*". The assignments for NDP+2-A+pdb8cat (NADPH) corresponds to the oxidized state. The ligand names are a combination of the three letter code in the PDB: the residue number, the chain code, and the PDB entry name. For polymers only the PDB number of the first residue is given.

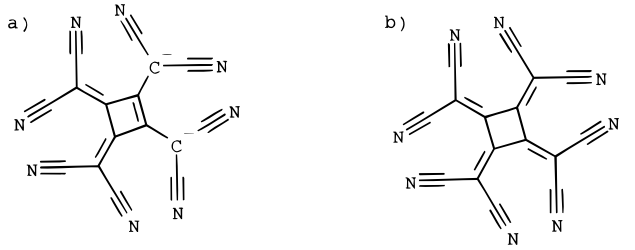

Figure 1. (a) Assignments for BEVJER10 as given in the CSD and the original publication.[10] (b) Calculated assignments.

ecule ligands, and nucleic acids. The structures from the CSD represent a test set which has been used for evaluating similar programs.[3,5] For each molecule in the test set we computed bond and atom types as described in the methods section. Each assignment was checked manually.

Of the 91 molecules from the CSD 88 (96.7%) had assignments of bond and atom types without any error (Table 1). If all the hydrogens are omitted the success rate drops only slightly to 87 (95.6%) molecules with completely correct assignments. This compares favorably with the results from Meng and Lewis with 73 (80%) correct assignments, and Baber and Hodgkin with 75 correct assignments (82%). A detailed report for all 91 molecules is shown in Table 2.

Most errors occur as a consequence of geometric distortions at terminal groups. Such an example is the terminal bond of an ethyl group in AERMYC10 which was typed as a double bond because of the short bond length of 1.339 Å and a valence angle of 123.344°—typical for a sp^2 hybridized carbon. A similar distortion of terminal ethyl groups causes a wrong assignment in PRMARI. The only wrong assignment which was not caused by a geometric distortion occurs in BEVJER10[10] where all ring bonds are typed as single bonds (Figure 1). BEVJER10 is highly symmetric with delocalized π bonds and a formal charge of -2. BALI

assigns a double bond to each sp^2 carbon which results in an assignment of bond types as shown in Figure 1b.

In three other cases (ANTETC, OXTETK, BAMLIK) BALI produces assignments which differ from the bond types given in the CSD. For ANTETC and OXTETK the assignments produced by BALI are in agreement with the original publications.[11,12] For BAMLIK the program typed the nitrogen−carbon bond of an amide group as a double bond because the carbonyl oxygen was missing in the 3D data extracted from the CSD. This was not counted as an error of the assignment procedure because the complete amide group would have been correctly recognized in the pattern recognition phase.

The execution time for typing all 91 compounds was 6.5 s on a SGI Indigo2 with a R4400 running at 150 MHz which shows that the program is sufficiently fast to process a large number of compounds in a reasonable time.

The success rate for protein ligands from the PDB is slightly lower than the success rate for structures from the CSD. One hundred nine (90%) of all 120 compounds were typed without any error (Table 1). This can be explained by the fact that the resolution of protein structures in the PDB is generally lower than the resolution of small molecules in the CSD. A detailed report for all 120 molecules is shown in Table 3. As observed for the structures from the CSD, most errors occur as a consequence of geometric distortions at terminal regions or at atoms with only one or two bonds to other atoms. A typical example is Biotin (BTN+500+pdb1bib) where a bond is typed as a triple bond because of an abnormal short bond length of 1.133 Å. A similar distortion causes a wrong assignment for ubiquinone where both double bonded oxygens of the quinone ring have bond lengths longer than the threshold value which separates single from double bonds. Therefore both oxygens are typed as sp^3 and the quinone ring as aromatic.

A principal hurdle for an approach which is based on geometrical criteria is that it is not possible to differentiate between electronic states which are not clearly reflected in differences in the geometry. BALI is not able to differentiate between, e.g., oxidized and reduced states of several cofactors by geometric criteria as hydrogen atoms are missing in most cases. This can been seen in the assignments for NADPH (NDP+2-A+pdb8cat) where BALI was not able to deduce the correct protonation from the coordinates. The calculated bond and atom types are in agreement with the oxidized state. A possible solution to this problem is to try to deduce the protonation states of ligands from the interacting groups in the protein. Work in this direction is currently in progress.

4. CONCLUSION

In this paper we have presented the program BALI which calculates bond orders and atom types for molecules from atom coordinates with high quality. The main purpose of BALI is to extract and process the fastly growing data about protein ligands from the PDB where information about bond and atom types are generally not available but urgently needed for a detailed analysis of receptor/ligand interactions. The assignment is done in several stages which include a recognition of simple functional groups, a ring perception, and an optimization of the assignment of alternating single and double bonds to networks of sp^2 hybridized atoms. Special care was taken for the correct typing of conjugated ring systems which are frequently found in ligands in the PDB.

Compared with programs from other authors BALI performs significantly better. Despite the generally low resolution of protein structures and the fact that in most cases hydrogen atoms are missing, the program types more than 96% of structures from the CSD and 90% of the protein ligands from the PDB without any error at all. The most important result is that BALI produces virtually no errors for well resolved structures. This is an advandage compared to other similar published approaches. Furthermore structures with contradicting geometries are reported for a subsequent more detailed manual inspection.

A conversion of all protein ligands in the PDB is currently in progress. Because of errors in some PDB files and structures with distorted and sometimes contradicting geometries it is not possible to fully automate this process. A manual check of all assignments is still needed. But the high quality of the assignments in combination with an excellent visualization tool such as PRADO makes the conversion of ligands from the PDB possible with a minimum of human intervention.

A further improvement in the peformance of BALI might be achieved with the use of predefined templates for small molecule ligands as many molecules have multiple copies in the PDB (e.g., several hundred hemes and nucleotides). Work in this direction is currently in progress.

ACKNOWLEDGMENT

This work was supported by funds from the Bundesminister für Bildung und Forschung, Germany.

REFERENCES AND NOTES

(1) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. D.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanochi, T.; Tasumi, M. The Protein Data Bank: A Computer Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(2) Hemm, K.; Hendlich, M.; Aberer, K. Constituting a Receptor Ligand Information Base from Quality-Enriched Data. In Proceedings from the Third International Conference on Intelligent Systems for Molecular Biology. ISBN 0-929280-83-0. **1955**, 170−178.
(3) Meng, E. C.; Lewis, R. A. Determination of Molecular Topology and Atomic Hybridisation States from Heavy Atom Coordinates. *J. Chem. Comput.* **1991**, *12*, 891−898.
(4) Allen, F. A.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development of Version 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.
(5) Baber, J. C.; Hodkin, E. E. Automatic Assignment of Chemical Connectivity to Organic Molecules in the Cambridge Structure Database. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 401−406.
(6) Blaney, J. M.; Crippen, G. M.; Dearing, A.; Dixon, J. S. DGEOM, *QCPE Catalog* **1990**, *10*, #590.
(7) Daylight User Manual; Daylight Chemical Information Systems, Inc.: Irivine, CA.
(8) Weininger, D.; SMILES 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
(9) Tripos User Manual; Tripos Associates: St. Louis, MO.
(10) Gerecht, B.; Kämpchen, T.; Köhler, K.; Massa, W.; Offermann, G.; Schmid, R. E.; Seitz, G.; Sutrisno, R. Pseudooxokohlenstoff-Dianionen der C4-Reihe mit Dicyanmethylenfunktionen. *Chem. Ber.* **1984**, *117*, 2714−2729.
(11) Barton, D. H. R.; Ley, S. V.; Meguro, K.; Williams, D. J. Reaction of Teracycline Hydrochloride with N-Chlorosuccinimide: X-Ray Crystal Structure of the Major Product. *J. Chem. Soc., Chem. Commun.* **1977**, 790.
(12) Jogun, K. H.; Stezowsky, J. J.; Chemical-Structural Properties of Tetracycline Derivatives. 2. Coordination and Conformational Aspects of Oxyteracycline Metal Ion Complexation. *J. Am. Chem. Soc.* **1976**, *98*, 6016−6026. OXTETK.

CI9603487