SATIS: Atom Typing from Chemical Connectivity

John B. O. Mitchell,*,† Alexander Alex,‡ and Michael Snarey‡

Department of Biochemistry & Molecular Biology, University College London, Gower Street, London WC1E 6BT, U.K., and Computational Chemistry, Pfizer Central Research, Ramsgate Road, Sandwich, Kent CT13 9NJ, U.K.

Received January 13, 1999

SATIS (simple atom type information system) is a protocol for the definition and automatic assignment of atom types and the classification of atoms according to their covalent connectivity. Its distinctive feature is that no bond type information is involved. Rather, the classification of each atom is based on a connectivity code describing the atom and its covalent partners. It is particularly useful when handling coordinate-based molecular representations with no bond order information, such as the PDB format. We survey the occurrence of the various connectivity codes in the 20 common amino acid residues in a sample of 304 different moieties from PDB protein−ligand complexes and also in a pseudo-random sample of 309 organic molecules from the CSD. We illustrate how connectivity codes can be grouped together to define atom types. We expect SATIS to be applicable to the derivation of atom types for statistical potentials, to the analysis of atomic interactions in structural databases, to studies of molecular similarity, and to the screening of virtual libraries in drug design.

1. INTRODUCTION

It is very useful in chemistry to assume that atoms of the same atomic number in similar chemical environments will usually give rise to local properties (such as local charge distribution[1]) that are approximately transferable between molecules. This underpins the concept of the functional group.

Numerous schemes have been proposed for the classification of atoms into various atom types, especially for biologically important molecules. The type assigned to an atom will generally reflect its atomic number, hybridization state, bonded partners, and bond orders. One of the most important uses of atom types is in the parametrization of potential energy functions. CHARMM,[2] AMBER,[3,4] and other molecular modeling packages have built-in atom typing schemes. These schemes are designed to incorporate all atom types necessary to describe common biological and organic molecules. Although these definitions are to some extent arbitrary, they reflect accepted notions of functional groups. Hendlich et al.[5] have recently published a scheme for automatically assigning atom and bond types to the ligands found in the Brookhaven Protein Data Bank (PDB),[6] a task which requires the ability to recognize both common and unusual functional groups across a wide range of organic compounds.

We seek a method of categorizing and indexing atoms that is different from those cited above by being independent of any subjective definitions, either of functional groups or of bond orders. Thus we define a set of connectivity codes that are dependent only on the atomic number of an atom and on the number and identity of its bonded partners. Thus there is no subjectivity in the assignment of the connectivity codes, except in those rare cases where the existence of covalent bonds is open to dispute.

When applying SATIS to atom typing, we can either use each connectivity code as an atom type in its own right or, as is often pragmatically useful, define atom types as sets of (one or more) connectivity codes describing chemically similar atoms. Our method was originally conceived as part of the BLEEP potential of mean force describing protein−ligand interactions.[7,8] In that context, it was appropriate to group a number of connectivity codes together in defining the atom types.

2. METHOD

(a) The Connectivity Code. SATIS is a concept rather than a program, and clearly the software requirements for converting different computational representations of chemical structure into connectivity codes would vary. Only the connectivity is required, and in principle any computational representation of chemical structure, or connectivity, such as SMILES[9] or the FDAT format of the Cambridge Structural Database (CSD),[10] could be used to generate connectivity codes automatically. Equally, connectivity can almost always be accurately derived from coordinates. Thus SATIS can be used with coordinate data alone, given a program like HBADD[12] to derive the list of covalently bonded partners for each atom in the molecule. The list of bonded partners is all the connectivity information required by our method; bond orders are not used. Second-nearest neighbors are used only in an extension of the method for C═O containing functional groups and in the expanded SuperSATIS codes. Both of these variations of SATIS are described below.

The connectivity information for each atom is formulated as a 10-digit connectivity code. The fixed length of the strings

* Corresponding author. Present address: Dr. John B.O. Mitchell, Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, U.K. E-mail: mitchell@biochemistry.ucl.ac.uk. Fax: (+44)-171-380-7463.

† University College London.

‡ Pfizer Central Research.

10.1021/ci9904214 CCC: $18.00 © 1999 American Chemical Society
Published on Web 05/25/1999

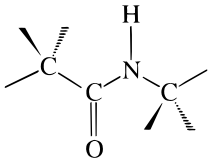

Figure 1. The peptide nitrogen is covalently bonded to two carbons and one hydrogen. The connectivity code begins with the atomic number of nitrogen (**07**), followed by those of the covalently bonded partners, in ascending numerical order (**01** for hydrogen, **06** for carbon, and **06** again). The lack of a fourth bonded partner is denoted by **99**. The full connectivity code is thus **0701060699**. The peptide oxygen has connectivity code **0806999995**; the final **95** is the extension code for a peptide or amide and replaces the usual **99**. The peptide carbon has the code **0606070895**.

is designed to facilitate sorting and data handling. The first two digits are the atom's atomic number (e.g., **06** for carbon or **16** for sulfur). The remainder of the code consists of four two-digit numbers, representing the atomic numbers of the atom's covalently bonded partners in ascending numerical order. If an atom has fewer than four bonded partners, the remaining positions in the connectivity code are filled with **99**. As examples, a peptide nitrogen (Figure 1) would have the connectivity code **0701060699** and a water oxygen atom **0801019999**.

There are some similarities between SATIS and the augmented atoms described by Adamson et al.[11] Their scheme, unlike ours, took explicit account of bond types and also excluded hydrogen. One should also note that there is, in general, no simple one-to-one mapping between SATIS connectivity codes and the atom types of CHARMM[2] or similar molecular mechanics programs. There are many instances where a single CHARMM atom type contributes to a plurality of SATIS codes and also where a single SATIS code corresponds to more than one CHARMM atom type.

(b) Extension of the Method for C═O Containing Functional Groups. An optional extension to the method allows a distinction to be made between different double-bonded C and O atoms (the double bond being inferred from the oxygen's having only a single covalent partner, as indicated by the raw connectivity code). The final **99** of these atoms' connectivity codes is replaced by an extension code in aldehydes (where the extension code is **93**), ketones (**94**), amides or peptides (**95**), esters (**96**), carboxylates (**97**), and carboxylic acids (**98**); in all other functional groups, the **99** is retained. This assignment is carried out by means of a search of more distant neighbors. Thus the oxygen in a peptide (or amide) −CONH− group would be represented as **0806999995** and the carbon as **0606070895**. The other atoms in the functional group retain their connectivity codes unaltered, such as **0701060699** for the peptide nitrogen in Figure 1. This extension is used in the work described in this paper.

(c) SuperSATIS. When information about second-nearest neighbors is required, the SATIS code of each atom can be catenated together with those of its four bonded neighbors (in ascending arithmetic order of SATIS code) to form a 50-digit SuperSATIS code. Where there are fewer than four bonded neighbors, **9999999999** is used to fill up the code. An atom's SuperSATIS code depends on the atomic numbers of up to 17 atoms: the central atom itself, up to four nearest neighbors, and up to 12 second-nearest neighbors. The 50-digit code contains some redundancy: where an atom has four bonded partners, the central atom's atomic number occurs five times and those of its four partners occur twice each (34 digits are nonredundant). SuperSATIS provides a very high degree of discrimination between atoms in even slightly different chemical environments. Although SuperSATIS has not been used in the applications described in this paper, we have successfully implemented it for PDB format files.

(d) Implementation for PDB Format Files. The covalent connectivities for ligands from the PDB[6] are assigned using a slightly modified version of the program HBADD,[12] which is part of the LIGPLOT[13] package. HBADD assigns covalent bonds to all heavy atom (non-hydrogen) pairs separated by less than 2.0 Å and to pairs involving hydrogen where the interatomic distance is less than 1.42 Å. Bonds are also assigned if indicated by the CONECT records in the PDB file (except in rare cases where these records are clearly inconsistent with the coordinates). Nonpolar hydrogens, usually absent from X-ray crystal structures, are generated by reference to the PDB's "Het Group Dictionary".[14] The output from this program is, for each PDB file, a list of all the atoms in each ligand, together with their covalently bonded partners. The assignment of a connectivity code to an atom is a trivial procedure, given the HBADD[12] output and knowledge of the PDB naming conventions, which allow the atomic number to be inferred from the four-character atom name string.

3. APPLICATIONS AND RESULTS

The following work has been performed using the basic SATIS method together with extension codes.

(a) Non-Hydrogen Atoms in Amino Acid Residues. As an example of the SATIS method, we look at the heavy atoms found in the 20 common amino acid residues. The reasonably conceivable definitions would range from four atom types (based on atomic number only) to 167 types (every atom in every residue different). Warme and Morgan[15] divided these atoms into 19 types for their study of side chain interactions, while Laskowski et al.[16] used 26 atom types based on the earlier work of Engh and Huber.[17] Melo and Feytmans,[18] however, used as many as 40 distinct atom types for their potential of mean force describing interactions within proteins. Our method assigns these atoms 28 distinct connectivity codes (27 if extension codes are ignored). These are shown in Table 1.

(b) Connectivity Codes for 304 Moieties from the PDB. We have investigated 351 protein−ligand complexes[19] from the PDB,[6] which comprised the original data set we used to generate our potential of mean force.[7,8] We searched these PDB entries for different organic and inorganic moieties, by which we mean entities such as amino acid residues, ligand molecules, and metal ions. We found a total of 304 such moieties (defined by unique residue names, though we ensure that water only occurs once) of which a list is available,[19] obviously including the 20 standard amino acid residues and less obviously also 17 different nonstandard residues. Together, they comprised these 37 amino acid residues, 15 metal ions, and 252 ligand residues. A ligand (a term which here includes cofactors, inhibitors, and indeed any molecule complexed with the protein in the crystal structure) may consist of one or more residues in the PDB nomenclature;

Table 1. Connectivity Codes for Non-Hydrogen Atoms in the 20 Amino Acid Residues[a–f]

			atom types			
connect. code	no. seen	description	M&F	W&M	X-Site	BLEEP
0806999995	22	20 × **O**; ASN **Oδ1**; GLN **Oε2**	5, 34	19, 12	3	*0803*
0606070895	22	20 × **C**; ASN **Cγ**; GLN **Cδ**	4, 33	18, 8	1	*0605*
0701060699	22	19 × **N**; HIS **Nδ1**; ARG **Nε**; TRP **Nε1**	3, 25, 36, 39	16, 9	6	*0702*
0601010606	20	13 × **Cβ**; 5 × **Cγ**; ILE **CG1**; LYS **Cδ**	8	2, 3	4, 11	*0601*
0601060607	19	19 × **Cα**	1	17	2	*0603*
0601060699	13	13 arom carbons in PHE, TYR, TRP	12	6	7, 14	*0602*
0601010106	8	8 methyl carbons	6	5	5	*0601*
0806999997	4	ASP **Oδ1&2**; GLU **Oε1&2**	28	13	9	*0803*
0601010607	4	GLY **Cα**; PRO **Cδ;** LYS **Cε**; ARG **Cδ**	2, 32, 35, 37	17, 2, 4	10, 11, 4	*0603*
0606060699	4	PHE **Cγ**; TYR **Cγ**; TRP **Cδ2** & **Cγ**	11, 13	7	17, 19, 20, 26	*0602*
0701010699	4	ASN **Nδ2**; GLN **Nε2**; ARG **Nη1&2**	18, 22	10, 11	13, 12	*0701*
0601060606	3	ILE **Cβ**; LEU **Cγ**; VAL **Cβ**	7	1	2	*0601*
0801069999	3	SER **Oγ**; THR **Oγ1**; TYR **Oη**	16, 40	14	8	*0802*
0601010616	2	CYS **Cβ**; MET **Cγ**	29	2	4	*0610*
0601060799	2	HIS **Cδ2**; TRP **Cδ1**	24	6	23, 7	*0607*
0606060799	2	TRP **Cε2**; HIS **Cγ**	14, 23	7	20, 24	*0607*
0606080897	2	ASP **Cγ**; GLU **Cδ**	27	8	1	*0606*
0601010116	1	MET **Cε**	30	2	5	*0610*
0601010608	1	SER **Cβ**	15	3	4	*0604*
0601060608	1	THR **Cβ**	17	1	2	*0604*
0601070799	1	HIS **Cε1**	26	6	21	*0613*
0606060899	1	TYR **Cζ**	31	7	15	*0608*
0607070799	1	ARG **Cζ**	21	?	1	*0613*
0701010106	1	LYS **Nζ**	20	11	22	*0704*
0706060699	1	PRO **N**	10	16	25	*0701*
0706069999	1	HIS **Nε2**	38	9	6	*0703*
1601069999	1	CYS **Sγ**	19	15	16	*1601*
1606069999	1	MET **Sδ**	9	15	18	*1601*
28	**167**	**totals**	**40**	**19**	**26**	**17**

[a] These results incorporate extension codes for C═O containing functional groups. [b] "M&F" refers to the atom types defined by Melo and Feytmans;[18] "W&M" refers to the atom types defined by Warme and Morgan;[15] "X-Site" refers to the types used by Laskowski et al.;[16] "BLEEP" refers to the types we define by grouping together connectivity codes for our BLEEP protein−ligand potential of mean force.[7] [c] Peptide N and C atoms are assumed to be bonded to the next residue in the sequence. [d] Histidine is assumed to be neutral. [e] ASP and GLU side chains are assumed to be deprotonated. [f] We were unable to type ARG Cζ using the Warme and Morgan scheme.

multiple-residue ligands are usually polymeric species such as polysaccharides. These 304 moieties comprise the data set for our survey of the occurrence of the connectivity codes. Each molecule or residue is considered only once, regardless of how many times it occurs in the 351 PDB entries.

The 304 moieties contain 8496 atoms, 4618 heavy atoms, and 3878 hydrogens, represented by 191 different connectivity codes. The most common connectivity codes, **0106999999** (3051 occurrences) and **0108999999** (491), correspond to hydrogens bonded to carbon and oxygen, respectively. The most common heavy atom connectivity code is **0601060699** (457 occurrences), which usually denotes a carbon in an aromatic ring, although it will less often denote an unsaturated aliphatic carbon. Among those types found less commonly are the isolated metal ions, which are by definition found no more than once each.

The 45 most frequently occurring codes are listed and described in Table 2. The four most common connectivity codes account for more than 50% of all atoms, 30 codes for 90%, and 114 of the 191 codes account for 99% of all atoms. If we exclude hydrogens, then 10 codes account for more than 50%, 41 codes for 90%, and 141 of the 187 codes account for 99% of all heavy atoms. The proportions of the elements among the nonmetallic heavy atoms are as follows: C, 59.3%; N, 10.0%; O, 26.6%; F, 0.5%; P, 1.9%; S, 1.3%; Cl, 0.2%.

(c) Connectivity Codes for 309 Entries from the CSD. We have also searched the CSD[10] (April 1998 release) for occurrences of the connectivity codes which were found in our PDB survey. For our sample, we required high-quality ($R \leq 0.075$), error-free organic crystal structures. Seeking a data set of similar size to the previous one, we initially extracted 340 entries, arbitrarily starting at the entry TOPQUK,[20] proceeding alphabetically and restricting ourselves to entries published in 1996 and 1997. We eliminated three entries which were duplicates, 26 where no hydrogen coordinates were present, and two with some missing hydrogens, leaving a set of 309 CSD entries.[19] This is designed to be a pseudo-random sample of organic molecules. All molecules contained in these entries were included, except that we limited ourselves to one occurrence of water. We have not implemented the derivation of SATIS connectivity codes directly from the FDAT files; rather we simply defined fragments corresponding to the various codes and searched for these fragments. Since we excluded 13 codes corresponding to isolated monatomic species, we were left to search for occurrences of 178 distinct connectivity codes. We did not search for any codes other than those found in the previous survey.

We found 15 718 atoms, 8095 heavy atoms and 7623 hydrogens, represented by 109 connectivity codes. Again, the most common connectivity code is **0106999999** (7349 occurrences; this includes only hydrogens bonded to carbons whose own codes are explicitly searched for), and the most common heavy atom code is **0601060699** (2347). The 30 most frequently found codes are all listed in Table 2. The

Table 2. Descriptions and Occurrences of Common Connectivity Codes

ranks			no. seen		
PDB	CSD	conn. code	PDB	CSD	description of connectivity
1	1	**0106999999**	3051	7349	hydrogen bonded to C
2	19=	**0108999999**	491	77	hydrogen bonded to O
3	2	**0601060699**	457	2347	carbon (with H) in aromatic ring or unsat. aliphatic
4	22=	**0801069999**	358	74	oxygen in C-**O**H group (or protonated -CO**O**H)
5	5	**0601010606**	340	541	carbon in saturated >**C**H_2
6	10	**0107999999**	333	197	hydrogen bonded to N
7	13	**0601060608**	311	143	carbon in >**C**H-O-
8	3	**0601010106**	201	694	carbon in H_3**C**-C methyl group
9	4	**0606060699**	174	683	carbon in arom ring, bonded to substituent/other ring
10	36=	**0701060699**	147	44	nitrogen in H**N**< (incl. peptide and amide -C(O)**N**H-C)
11	21	**0806999995**	138	75	oxygen in amide or peptide -C**O**NH-
12=	22	**0606070895**	135	74	carbon in peptide or amide C-**C**(O)NH-
12=	6	**0806069999**	135	324	oxygen in ether group C-**O**-C including sugars
14	7	**0601010608**	129	229	carbon in C-**C**H_2-OR (incl. primary alcohols)
15	9	**0606060799**	119	221	carbon, usually aromatic, bonded to N
16	110=	**0801159999**	114	0	oxygen bonded to H, P (as in phosphate)
17	29=	**0706069999**	109	55	nitrogen (H-bond acceptor), often aromatic
18	55	**0815999999**	102	16	oxygen bonded only to phosphorus (as in phosphate)
19	19=	**0601060607**	96	77	carbon in >**C**H-N (incl. peptide Ca)
20	12	**0601010607**	85	144	carbon in C-**C**H_2N
21	60=	**0806999998**	81	12	oxygen in carboxylic acid -C**O**OH
22=	60=	**0606080898**	80	12	carbon in carboxylic acid -**C**OOH
22=	29=	**0701010699**	80	55	nitrogen in C-**N**H_2
24	84=	**1508080808**	72	3	phosphorus in phosphate
25	35	**0806159999**	70	45	oxygen in C-**O**-P
26	17	**0706060699**	65	96	nitrogen bonded to 3C
27	11	**0601060606**	60	171	carbon in saturated aliphatic >**C**H-
28	45	**0606070799**	57	33	carbon bonded to 2 N, C arom/unsat. (adenine C6)
29	43=	**0816999999**	47	34	oxygen in sulfate type **O**═S
30	78	**0601070799**	46	4	carbon bonded to 2 N, H, arom/unsat. (adenine C8)
31	26	**0601060799**	37	70	carbon in aromatic/unsat. C-**C**H-N (cytosine C6)
32	66=	**0601060708**	36	9	carbon, saturated, bonded to H, C, N (ATP Cl′)
33	31	**0806999999**	35	53	oxygen in various **O**═C
34	8	**0606060899**	34	228	carbon in aromatic or unsat. >**C**-OH
35	78=	**0601060808**	33	4	carbon, saturated, bonds to H, C, 2O (usually in sugars)
36	60=	**0601010616**	28	12	carbon in C-**C**H_2-S
37=	28	**0601010107**	23	60	carbon in N-methyl H_3**C**-N
37=	36=	**0807999999**	23	44	oxygen in **O**═N (usually in -NO_2)
39	39=	**0806999997**	22	38	oxygen in carboxylate -C**O**O^-
40	96	**0815159999**	21	1	oxgygen bridging phosphates (as in ATP)
41=	46	**0607070899**	18	28	carbon, arom/unsat., bonds to C, N, O (uracil C4)
41=	34	**0906999999**	18	46	fluorine bonded to carbon **F**-C
41=	24=	**0606060894**	18	73	carbon in ketone >**C**═O
41=	24=	**0806999994**	18	73	oxygen in ketone >C═**O**
45	78=	**0607070799**	17	4	carbon, unsat., bonds to 3N (guan. C2, ARG Cζ)
46=	14	**0606060606**	13	141	carbon, saturated with bonds to 4C; >**C**<
49	18	**0601010108**	12	88	carbon in H_3**C**-O-
52=	15=	**0606080896**	10	98	carbon in ester -**C**(O)-O-
52=	15=	**0806999996**	10	98	oxygen in ester -C(**O**)-O-
52	27	**1706999999**	10	69	chlorine bonded to carbon **Cl**-C
191	178		8496	14441	

[a] These results incorporate extension codes for C═O containing functional groups. [b] This table lists the 45 most commonly found codes in the 304 PDB moieties. Codes outside this set which are among the top 30 for the 309 CSD entries are also shown. [c] The column "no. seen" is the number of occurrences of the given connectivity code in the 304 PDB moieties (column "PDB") and 309 CSD entries (column "CSD"). [d] The symbol > or < next to an atom implies bonds to two carbon atoms as well as any covalent partners explicitly indicated. Thus >**C**H-N implies a carbon atom bonded to one hydrogen, two carbons, and one nitrogen atom. [e] Only codes found in the PDB set were searched for in the CSD survey.

two most common connectivity codes account for more than 50% of all atoms, 23 codes for 90%, and 65 of the 109 codes account for 99% of atoms. If we exclude hydrogens, then 3 codes account for more than 50%, 32 codes for 90%, and 71 of the 106 codes account for 99% of the heavy atoms. The proportions of the elements among the nonmetallic heavy atoms are as follows: C, 81.3%; N, 5.2%; O, 11.1%; F, 0.6%; P, 0.1%; S, 0.9%; Cl, 0.9%. All these figures exclude atoms whose connectivity codes are not among the 178 investigated. The CSD set had a much smaller proportion of polar (non-carbon) heavy atoms than the PDB set (18.7% vs 40.7%). This reflects the special nature of biological molecules and their ligands. The paucity of phosphorus reflects the nonappearance of "biological" phosphorus connectivity codes in the general organic sample. The total number of atoms in our CSD set, including those whose connectivity codes were not searched for, was 18 990.

(d) Atom Types for the BLEEP Protein−Ligand Potential of Mean Force. For our BLEEP protein−ligand potential of mean force,[7] the sparseness of the data meant

Table 3. Grouping of Connectivity Codes into Atom Types for the BLEEP Protein−Ligand Potential of Mean Force

type	connectivity codes	no. seen	description
0101	**0107999999**	333	H bonded to N
0102	**0108999999**	491	H bonded to O
	0106999999, 0116999999, etc.	(3054)	not considered in potential[7]
0600	**0601050607**, etc.	1	C with unusual bonds/partners
0601	**0601010101, 0601010106, 0601010606, 0601060606, 0606060606**	614	C nonpolar, saturated
0602	**0601010699, 0601060699, 0601069999. 0606060699, 0606069999**	647	C nonpolar, unsaturated/aromatic
0603	**0601010107, 0601010607, 0601060607, 0606060607, 0601010115, 0601010615. 0601060615, 0606060615, 0601010133**, etc.	212	C bonded to N/P/As, saturated
0604	**0601010108, 0601010608, 0601060608, 0606060608, 0601010609, 0601060609, 0601010617, 0601060617, 0601010635**, etc.	466	C bonded to O/halogen, saturated
0605	**0601070895, 0606070895**	138	C amide/peptide
0606	**060108089x, 060608089x, 060808089x**	105	C carboxylate/acid/ester
0607	**0601010799, 0601060799, 0606060799, 0606079999, 0607999999, 0601011599, 0601061599, 0606061599, 0606159999, 0615999999**	160	C bonded to N/P and unsaturated/aromatic
0608	**0601010899. 060106089x, 060606089x, 0606089999, 0608999999, 0606060999, 0606061799, 0606063599, 0606065399**, etc.	70	C bonded to O/halogen and unsaturated/aromatic
0610	**0601010116, 0601010616, 0601060616, 0601060716, 0601060816, 0601061699, 0606060616, 0606061699, 0606071699**, etc.	58	C bonded to S/Se
0612	**0601010707, 0601060707, 0606060707, 0601070707, 0606070707, 0601010808, 0601060808, 0606060808, 0601080808**, etc.	90	C bonded to multiple polar atoms, saturated
0613	**0601070799, 0606070799, 0607070799, 0607079999, 0607070899, 0607080899, 0607089999**, etc.	146	C bonded to multiple polar atoms, unsat./aromatic **not** 0605, 0606
0617	**0607799999, 0601010180, 0601010182**, etc.	6	C bonded to metal
0701	**0701010199, 0701010699, 0706060699**	145	N with 3 bonds, nonpolar **not** 0702
0702	**0701060699**	147	N peptide or secondary amine
0703	**0706069999, 0706999999**	116	N acceptor, incl. cyanide
0704	**0701010101, 0701010106, 0701010606, 0701060606, 0706060606**	8	N with 4 bonds (charged)
0706	**0701011699, 0710060799, 0701060899, 0701061699, 0710070799, 0701151599, 0706060799, 0706079999, 0706080899**, etc.	33	N bonded to nonpolar atom(s)
0708	**0706062699, 0707269999**, etc.	9	N bonded to metal
0800	**0801999999, 0801059999, 0805999999, 0808999999**, etc.	5	O with unusual bonding/partners
0801	**0801019999**	1	O water
0802	**0801069999**	358	O hydroxyl
0803	**080699999x**	309	O in O═C
0804	**0806069999**	135	O ether
0805	**0801159999, 0806159999, 0815159999, 081599999**, etc.	309	O bonded to P/As
0806	**0801169999, 0806169999, 0816999999**, etc.	64	O bonded to S
0807	**0801079999, 0807999999**, etc.	25	O bonded to N
0808	**0823999999, 0826269999, 0829299999, 0829999999**, etc.	11	O bonded to metal
1601	**1601069999, 1606060699, 1606069999, 1606169999**, etc.	25	S reduced
1602	**1606060808, 1606060899, 1606070808, 1606080808**, etc.	22	S oxidized
1603	**1626262699, 1626269999**, etc.	14	S bonded to metal

[a] These atom types are those used in our BLEEP protein−ligand potential of mean force.[7] [b] For historical reasons, some type designations (e.g., *0609* or *0705*) are missing. The types originally corresponding to these have been merged with others. [c] The column "no. seen" is the number of occurrences of the given atom type in the 304 PDB moieties. [d] The use of "etc." indicates that other connectivity codes not listed are also assigned to this type. These can be inferred from the "description" column. [e] For other elements (atomic number *xx*), there is a single atom type *xx01* comprising all codes starting with **xx**. [f] The symbol **x** is used to indicate a plurality of possible extension codes for C═O groups. Thus **080699999x** covers codes **0806999993** to **0806999999**, etc. In fact, the definitions of atom types given here are unaffected by the use or otherwise of extension codes.

that we required many fewer atom types than connectivity codes. For the common elements (carbon, nitrogen, oxygen, and sulfur), we grouped several sets of connectivity codes together into individual atom types. This was carried out on the basis of chemical similarity. For instance, codes **0601010101**, **0601010106**, **0601010606**, **0601060606**, and **0606060606** all describe saturated nonpolar carbons and are grouped together into a single atom type (*0601*). All other non-hydrogen elements are described by one atom type each (*xx01*, where *xx* is the atomic number). For polar hydrogens, **0107999999** was defined as atom type *0101* and **0108999999** as type *0102*. These groupings (shown in Table 3) are appropriate in the context of the protein−ligand potential of mean force,[7] but it is anticipated that other groupings might be more useful in different circumstances. The definitions of the atom types described here are not affected by the use or otherwise of extension codes.

It is important to understand that the groupings of codes into atom types are arbitrary, but that the connectivity codes themselves are objective and uniquely determined by the covalent structure of a molecule. For the atom types shown in Table 3, we often assume that atoms differing by the substitution of a carbon neighbor for a nonpolar hydrogen should belong to the same atom type—hence the grouping described above for type *0601*. There are some circumstances in which would we wish to maintain the distinction between atoms differing in having carbon versus hydrogen as neighbors. This may be the case when the hydrogen is polar (and the central atom is therefore nitrogen or oxygen) and/or we wish to identify the atom types with specific functional groups, as is the case with peptide nitrogen (**0701060699**, type *0702*) and water oxygen (**0801019999**, type *0801*).

DISCUSSION

Our method provides a simple scheme for categorizing the atoms found in any covalently bonded molecule. It can be extended to the definition of atom types for either potential energy functions or analysis of spatial distributions.[16] The

analysis of the atoms found in the 20 common amino acid residues (Table 1) shows that SATIS gives sensible results and can automatically implement a classification scheme comparable with others devised for these atoms. One may choose to refine the scheme by grouping some connectivity codes together into atom types. SATIS is applicable to all covalently bonded atoms, so the possible problem of having no relevant atom type defined for unusual covalent connectivities is avoided. A particular advantage is that our connectivity codes do not depend on a subjective assessment of bond orders. Hydrogens are also particularly well-classified.

There are, however, some aspects of our method which are open to criticism. One is that the basic SATIS method is not always adequate to describe the local chemical environment of an atom completely. For O═C oxygens and carbons, we use the extension codes, described above, to identify aldehydes, ketones, amides, esters, carboxylates, and carboxylic acids. In the future, it may be desirable to introduce extension codes for other classes of molecules. For instance, we have already noted that the code **0601060699** cannot distinguish between an aromatic or olefinic carbon; similarly, a nitrogen in either a peptide or a secondary amine might be described by the code **0701060699**.

We have a ready alternative, however, where a higher level of discrimination is required. SuperSATIS codes can be used to take account of second-nearest neighbors: this might be done to discriminate peptide nitrogens from secondary amine ones and, more generally, to distinguish basic and nonbasic nitrogens. SuperSATIS would also pick out those carbons in isolated olefinic double bonds as opposed to aromatic or large conjugated systems. General use of SuperSATIS would obviate the need for extension codes, but would also increase the length of the bit strings by a factor of 5. We have a working version of SuperSATIS for PDB format files.

In other chemical contexts, the basic method is more adequately discriminatory. Protonated and deprotonated forms of a molecule, such as carboxylic acids and carboxylates, will have different codes. Indeed, occasionally we may perhaps achieve too much discrimination. In protein structures, it is not easy to tell whether histidine's Nϵ2 is protonated (**0701060699**) or not (**0706069999**): the solution here is probably to group these two codes together as one atom type. Another minor problem is that the current implementation of SATIS does not handle five-valent atoms: it is questionable whether the extended string lengths would be justified given the rarity of finding an atom with five or more partners. In this situation, we recommend taking the four partners of highest atomic number to define the connectivity code. Also, there is occasionally a problem of ill-defined covalent bonds, such as in some organometallics. One might choose to use a version of SATIS with longer strings (perhaps 14 digits, corresponding to up to six partners) for studies of systems, such as transition metal complexes, where higher valencies are frequently encountered. SATIS can unambiguously handle all elements up to atomic number 92 (uranium), and the special context of extension codes should not cause confusion with elements up to atomic number 98 (californium). We do not believe that the inability to handle the exotic elements with larger atomic numbers is a serious difficulty.

Our survey of the occurrences of the connectivity codes in 304 moieties from the PDB illustrates how a large proportion of the total number of atoms is accounted for by relatively few codes (30 codes covering 90% of atoms). The list of the most frequently occurring codes is broadly as one might expect. The most unusual feature is probably the large number of aromatic or unsaturated carbons bonded to one carbon and two nitrogens, with 57 examples of code **0606070799**. This is largely due to the many derivatives of adenine, guanine, and cytosine in the data set, as are the 46 instances of the similar code **0601070799**. It is also slightly surprising to find as many as 28 methylene CH_2 carbons bonded to sulfur (**0601010616**), 23 *N*-methyl carbons (**0601010107**), and 17 unsaturated carbons with bonds to three nitrogens (**0607070799**, these may be in heteroaromatics, like guanine C2, or else in guanidinium, like arginine Cζ). In contrast, we see only five aldehyde carbons (**0601060893**).

Our second survey, of 309 high-quality CSD[10] entries, revealed some quite different trends. In particular, the CSD set displayed a much smaller fraction of polar atoms. The CSD sample had a much higher proportion of the main carbon codes: aromatic/unsaturated CH (**0601060699**), saturated CH_2 (**0601010606**), saturated CH_3 (**0601010106**), and aromatic/unsaturated C (**0606060699**). Codes such as OH oxygen (**0801069999**), all carboxylic acid codes, and especially, all codes involving phosphorus are relatively much less frequent among the pseudo-random sample of organic molecules than among those found in the PDB. This survey excluded those codes which had not been found in the PDB survey.

We have found the SATIS connectivity codes to be perfectly adequate for defining atom types for our protein−ligand potential of mean force based on atom−atom distances in the PDB.[7] This was a case where the number of atom types to be used was always going to be significantly limited by the paucity of data. We expect that SATIS will continue to prove useful for database potentials. We do not, however, claim that SATIS could be used for generating atom types for high-accuracy potential functions based on quantum chemical data.[1] Clearly there are instances where two atoms with the same code have significantly different local charge distributions.

SATIS can provide criteria for grouping together atoms for the analysis of spatial distributions of atom types over a number of crystal structures. Such analysis is carried out by programs such as X-Site[16] and IsoStar.[21] The method may also prove useful in studies of molecular similarity. By comparing bit strings, either from SATIS or SuperSATIS, we can measure the similarity between two molecules. By grouping atomic connectivity codes from two molecules into appropriate bins, it should be possible to calculate similarity measures for molecules of differing sizes. The codes can also be used to identify corresponding atoms in different molecules. Over a large sample of molecules, the method could be used to measure database diversity. When coupled with some bioactivity data, we can search for features in the distributions of SATIS codes in active, as opposed to inactive, molecules.

One potentially important application of SATIS, in the context of drug design, is in the screening of virtual libraries consisting of tens of thousands of compounds. SATIS codes

would be used to represent each atom with atomic properties, such as polarity, taken to be a function of the atom's connectivity code. Once suitable parameters had been derived, this would provide a very rapid method of estimating molecular properties, such as water:octanol partition coefficients, where the sheer size of the virtual library makes more accurate methods prohibitively expensive. SATIS provides a simple measure of the numbers of hydrogen bond donors and acceptors, which is likely to be useful in a number of contexts, such as predicting the absorption of candidate drug molecules through cell walls. The estimates of molecular properties would allow many molecules from the virtual library to be eliminated as potential pharmaceuticals.

CONCLUSIONS

SATIS is a scheme for indexing atoms with a 10-digit connectivity code, which depends only on the identities of their covalently bonded partners and neither on bond orders nor, for the basic method, on second-nearest neighbors. An optional extension to the method provides more discriminating information for the special case of C═O containing functional groups. Where information on second-nearest neighbors is more generally required, 50-digit SuperSATIS codes can be used. The SATIS connectivity codes are not dependent on subjective assessments of bond or atom types, and the method classifies hydrogens usefully according to their bonded partner. SATIS is a concept rather than a computer program—it has so far only been properly implemented for PDB format files, but it is applicable to any format storing connectivity and/or coordinate data. The method has been used in a survey of the occurrence of the different codes in 304 moieties, mostly ligands, found in the PDB and in a similar survey of 309 organic entries in the CSD. We have also used it to generate atom types for the BLEEP protein−ligand potential of mean force.[7] We expect SATIS to be applicable to the derivation of atom types for statistical potentials, to the analysis of atomic interactions in structural databases, to studies of molecular similarity, and to the screening of virtual libraries in drug design.

ACKNOWLEDGMENT

J.B.O.M. thanks Pfizer Ltd. for financial support. We are grateful to Dr. Roman Laskowski for providing a modified version of the program HBADD. This is a publication from the BBSRC Structural Biology Centre in Birkbeck College and University College London.

REFERENCES AND NOTES

(1) Faerman, C. H.; Price, S. L. A Transferable Distributed Multipole Model for the Electrostatic Interactions of Peptides and Amides. *J. Am. Chem. Soc.* **1990**, *112*, 4915−4926.

(2) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D., States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(3) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.

(4) Weiner, S. J.; Kollman, P. A.; Nguyen, D.; Case, D. A. An All Atom Force Field for Simulations of Proteins and Nucleic Acids. *J. Comput. Chem.* **1986**, *7*, 230−252.

(5) Hendlich, M.; Rippmann, F.; Barnickel, G. BALI: Automatic Assignment of Bond and Atom Types for Protein Ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 774−778.

(6) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; 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.

(7) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP − a potential of mean force describing protein−ligand interactions: I. Generating the potential. *J. Comput. Chem.*, in press.

(8) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M. BLEEP − a potential of mean force describing protein−ligand interactions: II. Calculation of binding energies and comparison with experimental data. *J. Comput. Chem.*, in press.

(9) 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.

(10) Allen, F. H.; Kennard, O. 3D Search and Research Using the Cambridge Structural Database. *Chem. Design Autmn. News* **1993**, *8*, 1 & 31−37.

(11) Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of Structural Characteristics of Chemical Compounds in a Computer-based file. Part II. Atom-Centred Fragments. *J. Chem. Soc. C* **1971**, 3702−3706.

(12) Luscombe, N. M.; Laskowski, R. A. HBADD − A Computer Program for Assigning Molecular Connectivities, Department of Biochemistry, University College London, London WC1E 6BT, U.K.

(13) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a Program to Generate Schematic Diagrams of Protein−Ligand Interactions. *Prot. Eng.* **1995**, *8*, 127−134.

(14) Het Group Dictionary. Available from the Brookhaven National Laboratory World Wide Web site: ftp://pdb.pdb.bnl.gov/pub/resources/hetgroups/het_dictionary.txt.

(15) Warme, P. K.; Morgan, R. S. A Survey of Atomic Interactions in 21 Proteins. *J. Mol. Biol.* **1978**, *118*, 273−287.

(16) Laskowski, R. A.; Thornton, J. M.; Humblet, C.; Singh, J. X-Site: Use of Empirically Derived Atom Packing Preferences to Identify Favourable Interaction Regions in the Binding Sites of Proteins. *J. Mol. Biol.* **1996**, *259*, 175−201.

(17) Engh, R. A.; Huber, R. Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement. *Acta Crystallogr.* **1991**, *A47*, 392−400.

(18) Melo, F.; Feytmans, E. Novel Knowledge-Based Mean Force Potential at Atomic Level. *J. Mol. Biol.* **1997**, *267*, 207−222.

(19) Details of the data sets used in this work are available on the Internet at http://www.biochem.ucl.ac.uk/bsm/biocomp/mitchell_et al_list.html.

(20) Armstrong, D. R.; Davidson, M. G.; Davies, R. P.; Mitchell, H. J.; Oakley, R. M.; Raithby, P. R.; Snaith, R.; Warren, S. A Stable Methyl Phosphane Oxide/Lithium Amide Complex: a Structural and MO Calculational Investigation of the Mechanism of Proton Abstraction by Alkali Metal Reagents. *Angew. Chem., Intl. Ed. Engl.* **1996**, *35*, 1942−1944.

(21) Bruno, I. J.; Cole, J. C.; Lommerse, J. P. M.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. IsoStar: A Library of Information About Non-Bonded Interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525−537.

CI9904214