Atom subtypes present in protein structures: introducing dynamic atom similarity measure.

Grigoriy Mavropulo-Stoliarenko
Saint-Petersburg State University, 7-9 Universitetskaya nab., St.Petersburg, 199034, Russia,
gm2124@mail.ru

Keywords: Atom subtypes, covalent bond lengths, protein structure.

Motivation:

One of the important issues often encountered in studies that involve analysis of properties of protein structures (properties like atom packing, interatomic contacts etc) is the problem of assigning atoms of the same chemical type to a set of subgroups based on their microenvironment (covalent bonding pattern). The cause of this issue is purely combinatorial, because if we treat each of the 20 aminoacids' heavy atoms as being of its own subtype, that leaves us with 167 different atom types, which in turn gives thousands of parameters to estimate even in a simple case of two atom interactions. Literature review shows that each time this problem was faced it has been solved distinctly, often in an incompatible to the previous research way. We decided to address the question directly, in order to provide some stable framework for any future research in the field.

Although principles guiding generation of sets of atom subtypes are identifiable and quite reasonable, the detailed information of the procedures involved usually doesn't make it to the final manuscript, thus making proper inspection of authors' reasoning impossible.

When studied in detail, existing atom subtype classifications, often implicitly, address some general principles that we will outline below. For instance, Tsai el. all in their unifying work [1] isolated 13 types of atomic groups, implementing **XnHm** notation, where X - indicates the chemical nature of the non-hydrogen atoms; n - their valence; and Hm, the number (m) of hydrogen atoms attached to the non-hydrogen atom. Somewhat more complex classification is provided by Seeliger and de Groot, where they've adapted from OPLS-AA

forcefield 35 atom types, 6 of which being hydrogen atoms, present under diverse conditions. For instance, they have isolated 18 subtypes of carbon atom, which allows not only distinction of their valence, but also accounts for the “heavy” neighbors’ counts and types, and even accounts for outer electron orbital states of the neighboring atoms [2].

Method:

As we have shown above known atom subtype classifications capture general idea that atom’s properties are heavily influenced by covalent bonding pattern of the atom, and because of that, atoms of the same chemical type, but with different covalent bonding patterns should be processed separately. So the analysis of earlier works and some general sense reasoning can help us derive several principles that proper atom subtypes similarity measure should be based on:

- atom’s valence (types of its bonds), i.e. number of shared electron pairs in each atom’s bond, including possible “bond resonance” effects
- types of covalently bound atoms, which have their influence mostly due to differences in their electronegativity, which can cause partial atom polarization
- neighbors’ valence - i.e. outer electron orbital states, of the covalently bound atoms, since this can affect atom in question both thru joined electron density delocalization (aromatic rings) and “bond resonance” being main examples.

Following those principles we decided to base our definition of atom similarity on the covalent bonding driven approach, by implementing sum of minimum Hellinger distances between distributions of covalent bond lengths of atoms being matched. Covalent bond length distributions were collected from the set of high resolution representative (non redundant) X-Ray protein structures [3].

Results:

We provide sets of similarity lists for all heavy atoms present in protein structures, with corresponding numeric atom similarity estimation for each entry.

This will not only allow to dynamically group most similar atom subtypes, by varying similarity\sample size tradeoffs, but also can be used to assess and compare known atom subtype classifications.

1. J. Tsai et al. (1999) The packing density in proteins: standard radii and volumes, *J Mol Biol.*, 290(1):253-66.
2. D. Seeliger, B.L. de Groot (2007) Atomic contacts in protein structures. A detailed analysis of atomic radii, packing, and overlaps, *Proteins*, 68(3):595-601.
3. U. Hobohm et al. (1992) Selection of representative protein data sets, *Protein Sci*, 1(3):409-17.