A Three-Pass Algorithm for Generation of BE-Matrices from IUPAC Names

Sourav Mandal, Department of Computer Science and Technology, Bengal Engineering and Science University West Bengal, India

Somnath Pal, Department of Computer Science and Technology, Bengal Engineering and Science University, Howrah, West Bengal, India

ABSTRACT

One of the basic needs of chemoinformatics is to represent chemical compounds graphically in computer. Among several matrix representations of chemical graphs, BE-Matrix is one of the popular choices to represent molecules. There are several line notations available to input a chemical graph. Although several algorithms exist to convert from different line notations to suitable computer representations but using IUPAC Name, a line notation, to give input to the computer is not a popular method, because of the lack of suitable algorithm from IUPAC names to BE-Matrix or its variants. However, each and every chemist is familiar with IUPAC names, and therefore it calls for development of a suitable algorithm for such purpose. In this paper a three-pass algorithm for generating BE-Matrix from IUPAC name have been proposed and illustrated with suitable examples. The third pass of the algorithm can independently be used to convert from symbolic chemical names of any compound to BE-Matrix, thus making task of a chemist much simpler.

Keywords: Bond-Electron Matrix, Functional Groups, IUPAC Name, Line Notations, Substituents

INTRODUCTION

Chemoinformatics is the application of computational methods to chemical problems, with particular emphasis on the manipulation of structural information. Compounds can be interconverted into other compounds by chemical reactions. The numbers of new chemical compounds are increasing day by day and becoming enormous. So it is not possible to memorize everything about all the available compounds for a human being. The solution of this problem is electronic processing of such compounds. But for this, the main challenge is to represent the chemical compounds in computer for efficient processing of input/output of chemical reactions.

The 2D graphical representation of chemical structures in structure diagrams can be considered to be the universal “natural language” of chemists. These structure diagrams are models and are designed to make the molecules more

DOI: 10.4018/ijcce.2013070103
conceivable. In such a model, the atoms are typified by their atomic symbols and the bonding electrons by lines. However, such chemical structure diagram cannot be given directly as input to the computer for subsequent processing. Line notations (Gasteiger & Engel, 2003) represent the structure of chemical compounds as a linear sequence of letters and numbers. Several Line Notation are available for representing chemical compounds, such as IUPAC name, Wiswesser Line Notation (WLN) (Wiswesser, 1985; Wiswesser, 1982; Wiswesser, 1954), ROSDAL (Barnard et al., 1989; Rohbeck, 1991; Heller, 1990), SMILES Coding (Weininger & Weininger, 1989; Hinze & Welz, 1991; Bone et al., 1999), Sybyl Line Notation (SLN) (Ash et al., 1997). It has been observed that among such Line Notations WLN is almost obsolete (Gasteiger & Engel, 2003), ROSDAL, SLN are popular, SMILES is quite an important representation. It was also observed in Gasteiger & Engel, 2003), it is difficult to obtain the structure of a compound directly from its IUPAC name. However, IUPAC nomenclature is more familiar to the chemists than ROSDAL, SMILES, or SLN. In this paper we attempt to generate Bond-Electron matrix (Dugundji & Ugi, 1973) (an useful matrix representation for processing chemical compounds by computer) directly from the IUPAC names.

In the following section we discuss the background of the paper, specifically graphical and matrix representations of chemical compounds, IUPAC names and rules governing IUPAC names. The three-pass algorithm for generation of Bond-Electron matrix from IUPAC names along with appropriate illustrations and discussions is described in the next section. The last section summarizes our conclusion.

BACKGROUND

Graphical Representation

Chemical structure can be described using molecular graph, where the nodes are the atoms and the edges are the bonds between the two atoms and the number of the parallel edges denotes the bond order (Balaban, 1995; Schultz, 1989; Ivanciuc, 2003). Sometime in the molecular graph the hydrogen atom can be omitted. We can describe some properties of the atom in the nodes of the molecular graph (Ivanciuc, 2010). For example atom number or atom type can be added with the nodes and bond order with the edge of the molecular graph. These properties are very important when performing any kind of operation with or upon molecular graph. The graphical representation mainly describes the ways the nodes are connected that is the topology of the molecule. A graphical representation of di-methyl ketone or propanone is shown in Figure 1.

Figure 1. Representation of di-methyl ketone or propanone as a labeled graph
Direct Numerical Simulation of Gas-Solids Flow Based on the Immersed Boundary Method
www.igi-global.com/chapter/direct-numerical-simulation-gas-solids/47496?camid=4v1a