Copyright © 2019 Tech Science Press CMC, vol.59, no.3, pp.995-1003, 2019

The Algorithm of Chemical Species Analysis for *Ab Intio* Molecular Dynamics Simulations and Its Application

Zhiyi Han[1], Yugai Huang[2, 3], Xiaoqiang Xie[1], Ying Mei[1] and Bin Gu[1, *]

Abstract: In *ab initio* molecular dynamics (AIMD) simulations of chemical reactions, it is important but difficult to identify the chemical species in the trajectory automatically and quickly. In this paper, based on the chemical graph theory, an algorithm for molecular species identification, according to the molecular coordinates and empirical bond length database, is presented. As an example, the chemical species in condensed glycine at room temperature are investigated with our algorithm in detail. The chemical species, including canonical and zwitterionic glycine, their protonated and de-protonated states, and the free protons, are all identified, counted and recorded correctly. Potential applications and further development of the algorithm are also discussed.

Keywords: Chemical species, graph theory, algorithm, AIMD simulations.

1 Introduction

In the last decade, with the lasting improvement of computational simulation capabilities, quantum *ab initio* molecular dynamics simulations (AIMD) have become more and more important for the multidisciplinary studies of physical, chemical, biological and material science [Chin, Rood, Lin et al. (2000); Kühne (2014); Kohanoff (2006)]. One of the great advantages of AIMD simulations is the accurate representation of dynamical changes of chemical bonds, which is crucial for the understanding of micro-dynamics of any chemical reaction [Kühne (2014); Kohanoff (2006)].

In AIMD simulations, the chemical species, e.g., clusters of atoms with chemical bonding, keep varying. The identification and statistics of specific chemical species emerging in AIMD simulations are routine procedures for most of the reaction studies. For example, in the study initial clustering and aggregation of air pollution particles such as PM2.5 and the cloud condensation nuclei (CCN) from the sources of SO_2, ammonia, water vapor, and even criegee intermediates, both the cluster size and isomer structure should be tracked step by step, for the understanding of the growth of the particles in the atmosphere with varying meteorological conditions [Wang, Huang, Gu et al. (2016);

[1] Jiangsu Key Laboratory for Optoelectronic Detection of Atmosphere and Ocean, Nanjing University of Information Science and Technology, Nanjing, 210044, China.

[2] Department of Chemistry, Jiangsu Second Normal University, Nanjing, 210013, China.

[3] Queen's University, Belfast, Northern Ireland, Belfast BT7 1NN, United Kingdom.

[*] Corresponding Author: Bin Gu. Email: gubit@163.com.

CMC. doi:10.32604/cmc.2019.02181 www.techscience.com/cmc

Herb, Nadykto and Yu (2011); Zhu, Kumar, Zhong et al. (2016); Myerson and Trout (2013)]. In a physiological environment, some bonding (or non-bonding) between specific atoms in protein and DNA are crucial for the survival of a cell [Kohanoff, McAllister, Tribello et al. (2017); Gu, Smyth and Kohanoff (2014)].

However, as far as we know, there is still no report on the algorithm of automatic identification, classification and statistics of the chemical species for AIMD simulations. In literatures, most of the related tasks are accomplished by artificial offline recognition snap by snap after the simulation [Wang, Huang, Gu et al. (2016); Gu, Smyth and Kohanoff (2014)]. AIMD is expected to be applied for large scale simulations with over million and more snaps recorded in the trajectory. There needs further guarantee on the unified criteria, accuracy and efficiency in these works. A standard and smart tool is needed for chemical clustering analysis of AIMD simulations [Kühne (2014); Kohanoff (2006)].

In this paper, based on the chemical graph theory [Pietrucci and Andreoni (2011); Dias and Milne (1992)], an identification algorithm for chemical species in AIMD simulation trajectory is presented. In Section 2, the algorithm, e.g., the manipulation of the connecting matrix (R_c) according to an empirical bonding criteria database, is given in detail. In Section 3, taking the condensed glycine at room temperature as an essential example, the molecular structure variations from complete canonical states to the mixture of zwitterionic and canonical states is studied. Further potential applications and discussions are given at the end of the paper.

2 The chemical species identification algorithm

2.1 The essential problem

In AIMD simulations, the elements of a system are atoms contained in the simulation cell. With the adiabatic approximation, the movements of the valence electrons are described with the time dependent many-body Schröedinger equations (SEs). In practice, the approximations such as Hatree-Fock (HF) and Density Functional Theory (DFT) of SE, are adopted to solve the SEs. The nucleus with core electrons are always regarded as classical mass points and described with the second Newton Law [Kühne (2014)].

In graph theory, a chemical species can be defined as a collection of bonded atoms. For a specific species, any atom is reachable from another one through chemical bond network. The chemical bonds mean stable and strong enough interactions between atoms, ions or molecules. Bonding interactions vary with atom species, local chemical arrangements and macro-circumstances such as temperature and pressure. To figure out the bond connections in the snaps of AIMD trajectory, the bonding criteria, e.g. the standard inter-atomic distances, should be observed or calculated theoretically. As a starting point, an empirical database of bonding criteria can be established from the experimental measurements with techniques such as X-ray diffraction and NMR methods [Tjandra and Bax (1997)] for typical chemical species.

For AIMD simulations, the coordinates of the atoms are available. The inter-atomic distances can be easily calculated from the atom coordinates. Therefor the connectivity matrix $R_c(N \times N)$ of the system, with N atoms, can be defined based on the bonding criteria [Dias and Milne (1992); Pietrucci and Andreoni (2011)]. To this end, the essential problem of chemical species identification is how to search and classify the local chemical bonding networks according to their topology information in the MD trajectories accurately and as quick as possible.

2.2 Algorithm and its realization

The flow diagram of our algorithm of chemical species recolonization is shown in Fig. 1. Before the configuration of the MD snap is loaded, the initial value of the connectivity matrix $R_c(N \times N)$ is set to be zero. Define the distance between the mass centers of a pair of atoms (the i-th and j-th atoms) as $r(i,j)$. If $r(i,j) \leq r_0(i,j)$, the relevant elements of R_c are set to be $R_c(i,j) = R_c(j,i) = 1$. Here, $r_0(i,j)$ is the bonding criteria of the i-th and j-th atoms in the database. For convenience, the diagonal elements $R_c(i,i)$ are all set to be 1. The none-zero elements of this original matrix $R_c(org)$ contains the topology information of current state of the system.

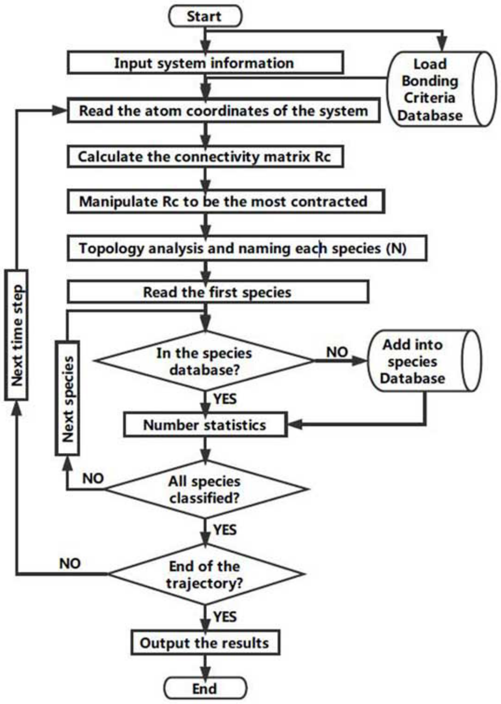
Start
Input system information
Load Bonding Criteria Database
Read the atom coordinates of the system
Calculate the connectivity matrix Rc
Manipulate Rc to be the most contracted
Topology analysis and naming each species (N)
Read the first species
Next time step
Next species
In the species database?
NO
Add into species Database
YES
Number statistics
NO
All species classified?
YES
NO
End of the trajectory?
YES
Output the results
End

Figure 1: The flow diagram of the chemical species classification and statistics algorithm

To extract the inter-molecule networks and classify the molecule clusters in each MD snap from $R_c(org)$, a serial of manipulations of $R_c(N \times N)$ are designed. From the column view of $R_c(org)$, setting the number of the non-zero elements in each (for example, the i-th) column of R_c as m, the i-th atom will directly bond with other $(m-1)$ atoms. If any two columns (for example the i-th and the j-th) have any non-zero matrix element in the same row (for example the k-th row), then these two atoms (the i-th and j-

th) are indirectly connected through the atom of the non-zero row (the k-th atom). All of the atoms, directly or indirectly, connected with the i-th and the j-th atoms belongs to a local network. The collection of these atoms can be defined as a chemical species. With a serial of operations as: if $R_{c-ic} \cap R_{c-jc} \neq \{0\}$, then $R_{c-ic} = R_{c-ic} \cup R_{c-jc}$ and $R_{c-ic} = 0$. The networking information can be incorporated into the first column of these atoms, while other following columns will be set as zero.

Through the above column integrations, no crossing link is left between any two non-zero columns. As a result, all of the elements of the upper right corner of the matrix R_c changes into zero. Each column with non-zero elements represents an isolated chemical species. The number of the non-zero elements in the column equals to the atoms of the specific molecular cluster. The distinct topology of the chemical species can be identified with the value of each element.

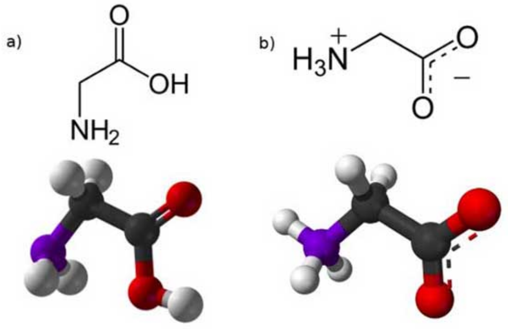

Figure 2: Isomers of glycine monomer: a) canonical, b) zwitterion (color online)

In practice, the chemical formula of each species can be represented by the serial of atomic species, each followed by the ordered atom numbers directly connected to that kind of atoms. For more clarity, we take glycine, the simplest amino acid, as example. As shown in Fig. 2, the glycine monomer has two states at room temperature with the chemical formula of NH_2CH_2COOH (canonical) and NH_3CH_2COO (zwitterion). In our algorithm and program, the topologies of these two isomers are represented by the sequence of atomic species (C-O-N-H). After the name of each atomic species, the numbers of directly connected atoms to these atoms are listed with increasing order. Hence, the canonical and zwitterion are represented as: C34$\underline{\text{O12N3}}$H11111 and C34$\underline{\text{O11N4}}$H11111, individually. They topology differences are highlighted by underlined sections. The number of the nearest neighbors of oxygens are {1, 2} and {1, 1}. For the nitrogen, they are {3} and {4} individually. The time-dependent evolution of all chemical species of glycine in condensed state will be studied in the next section.

In practice, the program of the algorithm has been written in Fortran 90. The trajectory of AIMD simulations in xyz format can be analyzed online and offline, snap by snap. The chemical species are classified automatically. In addition, some optional functions such as coordinates classification and statistics, as shown in Fig. 1 are also realized. At present, the code is available via email to the corresponding author.

3 Application: chemical species in amorphous glycine at room temperature

3.1 Basic properties of glycine

Glycine is an excellent example for species analysis because it has great chance to be both locally protonated and de-protonated with proton transferring. In the gas phase, glycine is found in its canonical (neutral) form as shown in Fig. 2(left). In aqueous solution, glycine transforms to its zwitterionic form, where a proton transfers from the acid to the amino group as shown in Fig. 2(right). Theoretically, the zwitterionic form is more stable than the canonical state by about 7 kcal/mol, with an interconversion barrier around 12 kcal/mol [Leung and Rempe (2005)].

In micro-solvation state, neutral glycine will transform into a zwitterion only after seven water molecules were placed around it [Bachrach (2008)]. At room temperature, pure glycine forms a hydrogen-bonded solid, of which three polymorphic crystal structures, α-, β- and γ-, are known [Iitaka (1960)]. The polymorphism of glycine makes the molecular structure and degree of crystallinity in condensed state dependent on environmental conditions, presence of additives, etc.

As an application of our algorithm, we performed a long enough AIMD simulation of glycine in a periodic box, which allows us to gain an understanding of its detailed chemical structures at room temperature. In this work the condensed state glycine is deemed as an amorphous to represent one possible situation under physiological conditions [Shu, Rani, Suryanarayanan et al. (2004); Gu, Smyth and Kohanoff (2014)].

3.2 Simulation details

The initial amorphous sample of the condensed glycine, which contains 32 canonical isomers in a cubic box of 15:05 Å with periodic boundary conditions, was prepared with the molecule editor ATEN [Youngs (2010)]. The structure was equilibrated for 1 ns via classical MD simulation using the CHARMM force field [MacKerell, Banavali and Foloppe (2000)] with DL-POLY simulation tool [Smith, Yong and Rodger (2002)].

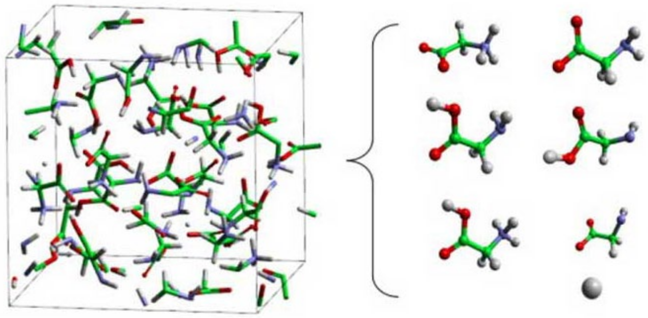

Figure 3: The simulation box which contains 32 glycine molecules (left). The chemical species identified in the AIMD trajectory (right) (color online)

After the classical equilibration, the AIMD simulations were carried out with the quantum module Quickstep (QS) of the open source code CP2K [VandeVondele, Krack, Mohamed et al. (2005)]. The AIMD simulations were performed at the PBE [Perdew, Burke and Ernzerhof (1996)] pure DFT functional level theory. Periodic boundary conditions was applied to the Poisson solver. The Goedecker-Teter-Hutter pseudopotentials [Goedecker and Teter (1996)], the TZVP-GTH basis sets, and the PBE [Perdew, Burke and Ernzerhof (1996)] exchange-correlation functional were utilized. The time step of MD simulations is 1 fs. After an *ab initio* optimization of 500 steps, a product simulation with NVT dynamics at 300 K was carried out for over 15 ps.

3.3 Results and discussions

In the chemical species analysis of the condensed glycine, the bonding criteria are listed as $r_0(CO) = 1.43\text{Å}$; $r_0(CN) = 1.47\text{Å}$; $r_0(CH) = 1.09$; $r_0(NH) = 1.01\text{Å}$; $r_0(OH) = 0.96\text{Å}$. If the distance between two atoms is less than 1.25 times the equilibrium bond length r_0, the atoms are bonded. As shown in Fig. 3, all of the five possible chemical species in the condensed glycine at room temperature are successfully identified by our program. The four glycine-based species in the sample are: canonical (Gly$_C$), zwitterionic (Gly$_Z$), deprotonated (Gly-H), and protonated (Gly+H). In addition, there exist some free protons (P), which are transferring among the glycine-based molecules.

The numbers of the five species in the simulation box during the AIMD simulation are shown in Fig. 4. It can be seen that, during the initial stage of about 4 ps, the system is out of equilibration. The number of canonical glycine keeps decreasing, while the ratios of all other species keep increasing. After the transition period, a dynamical equilibrium is reached in the following simulation. All the fluctuations of the number of each species are no more than 2. In the simulation box, the average number of each glycine-based state is around 8, individually. It means that, the original canonical molecules have changed equally to be the four glycine-based species. Meanwhile, the average number of free protons is around 1.6 in the simulation sample. These dynamically transferring protons promote the transformation among canonical and zwitterionic states of glycine.

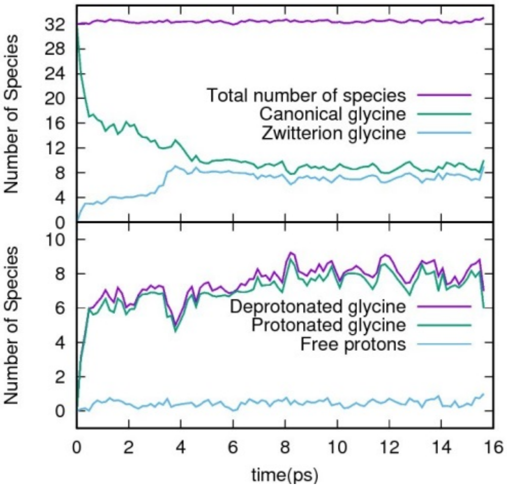

Figure 4: Time evolution of the number of molecules species in the sample of amorphous glycine. The bonding criteria is set as $R < 1.25R_0$ (color online)

It is clear that our algorithm of chemical identification and classification for AIMD simulations is quite effective. It should be noticed that the number distribution of molecules of the species depends on the bonding criteria database. As a reference, we recommend the bond length database (https://cccbdb.nist.gov/expbondlengths1.asp) provided by National Institute of Standards and Technology (NIST) for general chemical species identification for AIMD simulations.

4 Conclusions

In this work, we present an algorithm of automatic and quick identification of all chemical species in the trajectory of *ab initio* molecular dynamics simulations based on graph theory [Pietrucci and Andreoni (2011); Dias and Milne (1992)]. The input data are mainly the atomic coordinates and the general bonding criteria database. The chemical species can be recognized and classified in the dynamics simulations. Our program can be used for both online and offline analysis. It can be widely used for any long-time *ab initio* molecular dynamics simulations of molecular clustering and chemical reactions [Myerson and Trout (2013); Wang, Huang, Gu et al. (2016)].

At present, the bonding order parameters have not been taken into consideration in the program. To develop the methods for more complex chemical environments with fast electron transferring and for those atoms with multiple bonding orders, the dynamic bonding criteria should be designed, according to the online calculations of electron density distribution along AIMD simulations [Lu and Chen (2013)]. It is the next aim of our project. In addition, some object-orientated functions and interfaces, for more versatile analysis along with chemical species analysis, can also be introduced into the program.

Acknowledgement: The authors thank Dr. Gareth Tribello of Queen’s University Belfast for helpful discussions on the algorithm. This work was partially supported by the program of the Key Laboratory for Aerosol-Cloud-Precipitation of CMA-NUIST (No. KDW1304), the National Natural Science Foundation of China (Grant No. 11105075).

References

Bachrach, S. M. (2008): Microsolvation of glycine: a DFT study. *Journal of Physical Chemistry A*, vol. 112, no. 16, pp. 3722-3730.

Chin, M.; Rood, R. B.; Lin, S.; Müller, J. F.; Thompson, A. M. (2000): Atmospheric sulfur cycle simulated in the global model gocart: model description and global properties. *Journal of Geophysical Research: Atmospheres*, vol. 105, no. D20, pp. 24671-24687.

Dias, J. R.; Milne, G. W. A. (1992): Chemical applications of graph theory. *Journal of Chemical Information and Computer Sciences*, vol. 32, no. 1, pp. 1.

Goedecker, S.; Teter, M. (1996): Separable dual-space gaussian pseudopotentials. *Physical Review B-Condensed Matter and Materials Physics*, vol. 54, no. 3, pp. 1703-1710.

Gu, B.; Smyth, M.; Kohanoff, J. (2014): Protection of DNA against low energy electrons by amino acids: a first-principles molecular dynamics study. *Physical Chemistry Chemical Physics*, vol. 16, no. 44, pp. 24350-24358.

Herb, J.; Nadykto, A. B.; Yu, F. (2011): Large ternary hydrogen-bonded prenucleation clusters in the earth's atmosphere. *Chemical Physics Letters*, vol. 518, pp. 7-14.

Iitaka, Y. (1959): Crystal structure of β-glycine. *Nature*, vol. 183, no. 4658, pp. 390-391.

Kohanoff, J. (2006): *Electronic Structure Calculations for Solids and Molecules*. Cambridge University Press, Cambridge.

Kohanoff, J.; McAllister, M.; Tribello, G. A.; Gu, B. (2017): Interactions between low energy electrons and DNA: a perspective from first-principles simulations. *Journal of Physics: Condensed Matter*, vol. 29, no. 38.

Kühne, T. D. (2014): Second generation car-parrinello molecular dynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 4, no. 4, pp. 391-406.

Leung, K.; Rempe, S. B. (2005): *Ab initio* molecular dynamics study of glycine intramolecular proton transfer in water. *Journal of Chemical Physics*, vol. 122, no. 18, pp. 4637-201.

Lu, T.; Chen, F. (2013): Bond order analysis based on the Laplacian of electron density in fuzzy overlap space. *Journal of Physical Chemistry A*, vol. 117, no. 14, pp. 3100-3108.

MacKerell, A. D.; Banavali, N.; Foloppe, N. (2000): Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, vol. 56, no. 4, pp. 257-265.

Myerson, A. S.; Trout, B. L. (2013): Nucleation from solution. *Science*, vol. 341, no. 6148, pp. 855-856.

Perdew, J. P.; Burke, K.; Ernzerhof, M. (1996): Generalized gradient approximation made simple. *Physical Review Letters*, vol. 77, no. 18, pp. 3865-3868.

Pietrucci, F.; Andreoni, W. (2011): Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale. *Physical Review Letters*, vol. 107, no. 8.

Shu, J. B.; Rani, M.; Suryanarayanan, R.; Carpenter, J. F.; Nayar, R. et al. (2004): Quantification of glycine crystallinity by near-infrared (NIR) spectroscopy. *Journal of Pharmaceutical Sciences*, vol. 93, no. 10, pp. 2439-2447.

Smith, W.; Yong, C. W.; Rodger, P. M. (2002): Dl poly: application to molecular simulation. *Molecular Simulation*, vol. 28, no. 5, pp. 385-471.

Tjandra, N.; Bax, A. (1997): Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*, vol. 278, no. 5340, pp. 1111-1114.

VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T. et al. (2005): Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach. *Computer Physics Communications*, vol. 167, no. 2, pp. 103-128.

Wang, Y.; Huang, Y.; Gu, B.; Xiao, X.; Liang, D. et al. (2016): Formation of the $H_2SO^-_4$ dimer in the atmosphere as a function of conditions: A simulation study. *Molecular Physics*, vol. 114, no. 23, pp. 3475-3482.

Youngs, T. G. A. (2010): Aten-An application for the creation, editing, and visualization of coordinates for glasses, liquids, crystals, and molecules. *Journal of Computational Chemistry*, vol. 31, no. 3, pp. 639-648.

Zhu, C.; Kumar, M.; Zhong, J.; Li, L.; Francisco, J. S. et al. (2016): New mechanistic pathways for criegee-water chemistry at the air/water interface. *Journal of the American Chemical Society*, vol. 138, no. 35, pp. 11164-11169.