Postprint: Intelligent Simulation and Detection System for Molecular Cloud Cores in the Milky Way Imaging Scroll Painting Survey Data

Authors: Zhou Guangrong, Zeng Xiangyun, Zeng Shuguang, Huang Yao, Zheng Sheng, Luo Xiaoyu, Chen Zhiwei, Jiang Zhibo

Date: 2022-06-28T00:00:00+00:00

Abstract

Modern astronomy holds that molecular cloud cores are the birthplaces of stars. Comprehensive investigations into the detection of molecular cloud cores and their properties are essential for understanding star formation processes and the evolution of galaxies and the universe. With the implementation of the Milky Way Imaging Scroll Painting (MWISP) survey and the rapid accumulation of observational data, the development of an intelligent simulation and detection system for molecular cloud cores has become imperative. The system offers functionalities including intelligent detection of molecular cloud cores, simulation modeling, parameter retrieval, three-dimensional visualization, and data storage. By utilizing this system, researchers can conveniently and efficiently perform detection and three-dimensional visualization of molecular cloud cores in MWISP observational data, thereby facilitating a better study of their physical properties.

Full Text

An Intelligent Simulation and Detection System for Molecular Clumps in MWISP Observational Data

Zhou Guangrong[1], Zeng Xiangyun*[1], Zeng Shuguang[1], Huang Yao[1], Zheng Sheng[1], Luo Xiaoyu[1], Chen Zhiwei[2], Jiang Zhibo[2]

[1]Center for Astronomy and Space Science, College of Science, China Three Gorges University, Yichang 443002, People' s Republic of China
[2]Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210023, People' s Republic of China

Abstract

Modern astronomy holds that molecular clumps are the birthplaces of stars. Comprehensive detection and characterization of molecular clumps are essential for understanding star formation processes as well as the evolution of galaxies and the universe. With the ongoing Milky Way Imaging Scroll Painting (MWISP) survey and the rapid accumulation of observational data, developing an intelligent simulation and detection system for molecular clumps has become necessary. The system provides functions including intelligent detection of molecular clumps, simulation modeling, parameter reproduction, three-dimensional visualization, and data storage. Using this system, researchers can conveniently and efficiently perform molecular clump detection and 3D visualization on MWISP observational data, facilitating better investigation of their physical properties.

Keywords: Molecular clump; Three-dimensional visualization; Local Density Clustering; Parameter reproduction

1. Introduction

The detection of interstellar molecular hydrogen in the ultraviolet band by Carruthers [1] and CO at 2.6 mm wavelength by Wilson et al. [2] ushered in a new era of research on molecular interstellar medium, with the discovery of organic molecular media leading to the birth of molecular astrophysics. Molecular clouds constitute one of the fundamental components of interstellar medium, primarily consisting of molecular gas mixed with small amounts of atoms, ions, dust, and other components [3]. Molecular clouds in galaxies exhibit structures across a wide range of scales, with their densest regions referred to as molecular clumps [4,5]. Modern astronomy holds that stars form within molecular clumps [6,7]. Consequently, molecular clumps are crucial for establishing theoretical models of star formation observational characteristics in galaxies [8] and facilitate further studies on star formation and evolution [9].

The first phase of the Milky Way Imaging Scroll Painting (MWISP) survey plans large-scale observations of the Galactic plane from longitude -10° to +250° and latitude -5° to +5° using the 12CO (J=1-0), 13CO (J=1-0), and C18O (J=1-0) spectral lines. To date, the project has obtained 10,941 data cells, each measuring 30 $\times$30 with 16,384 velocity channels [10]. The second phase has already commenced, expanding the latitude coverage to -10.25° to +10.25°, which enriches the observational data across broad spatial scales of molecular clouds, different evolutionary stages, and diverse environments [11]. To better exploit the value of these data, detecting molecular clumps and analyzing their physical properties will provide scientific support for studies of early star formation stages.

With the steady progress of the MWISP project, molecular cloud data are accumulating rapidly, making manual detection and verification a time-consuming and labor-intensive task. To enable more efficient scientific analysis of molec-

ular cloud data, this paper designs and develops an intelligent simulation and detection system for molecular clumps specifically tailored to MWISP observational data. The system integrates molecular clump simulation, detection, catalog matching, parameter reproduction, 3D visualization, and storage into a unified platform with a user-friendly interface, facilitating convenient use by researchers. The system employs a three-dimensional Gaussian mathematical model [12] for generating simulated data to validate detection algorithm effectiveness. The molecular clump detection algorithm utilizes the Local Density Clustering (LDC) method proposed by Luo et al. [13]. Parameter reproduction employs the Multi-Gaussian Model (MGM) [13] to further refine clump parameters. Three-dimensional visualization intuitively displays the positions, shapes, and sizes of molecular clumps. Finally, MySQL database archives and stores molecular clump data and results, providing scientific data support for related research and accelerating scientific output.

Funding

This work is supported by the National Natural Science Foundation of China (U2031202, 11903083, 11873093).

Received date: ; **Revised date:**

Author biography: Zhou Guangrong, male, master' s student, research direction: astronomical technology. Email: 1971987925@qq.com, xyzeng2018@163.com

1. System Design and Implementation

1.1 Functional Design

All basic functions of the system modules have been implemented, from simulated data generation to final data storage, including molecular clump detection, catalog matching, parameter reproduction, and 3D visualization, making molecular clump research more intuitive. The system comprises five main modules, each containing several sub-modules. The system functional architecture is illustrated in Figure 1 [Figure 1: see original paper].

1.2 System Implementation

The system is implemented in Python, with the interface designed using the PyQt5 framework and object-oriented programming principles to realize all system functions. PyQt5 inherits the advantages of Qt, reducing coupling between modules and facilitating future system expansion and maintenance, while its integration with Python significantly enhances development efficiency.

2. System Functions

2.1 Data Generation

Data generation includes two distinct modes: simulated data generation and synthetic data generation. MWISP data are three-dimensional, comprising Galactic longitude, latitude, and velocity. Based on the study of M17 SW by Stutzki and Guesten [14], molecular clumps exhibit Gaussian distributions in both spatial and velocity coordinates, with their column densities also following a Gaussian distribution. The three-dimensional Gaussian mathematical model facilitates the reproduction of molecular clump parameters. Therefore, simulated data are generated using a 3D Gaussian model that employs mathematical calculations and interpolation to produce synthetic molecular clumps from given parameters. Synthetic data are created by randomly adding simulated clumps to real observational data.

2.1.1 Simulated Data Generation For given molecular clump parameters, the three-dimensional Gaussian mathematical model generates specific molecular clump data. This enables verification and optimization of detection algorithms for particular experimental requirements. Generating large amounts of simulated data also reduces reliance on real data for experiments. The complementary use of simulated and real data allows for cross-validation and verification of detection algorithms, enabling more comprehensive evaluation of algorithm performance. By analyzing real data, constraints are placed on relevant physical parameters of molecular clumps to establish a 3D Gaussian model generation framework.

When generating simulated data, if two molecular clumps satisfy either equation (1) or (2), they are considered non-overlapping:

$$|v_i - v_j| \geq \sigma_{v_i} + \sigma_{v_j}$$

$$\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \geq \sqrt{\sigma_i^2 + \sigma_j^2}, \text{where } \sigma_j = \sqrt{\sigma_{x_j}^2 + \sigma_{y_j}^2}$$

where (x_i, y_i, v_i) and (x_j, y_j, v_j) represent the centroid coordinates of the i-th and j-th clumps, respectively, and $(\sigma_{x_i}, \sigma_{y_i}, \sigma_{v_i})$ and $(\sigma_{x_j}, \sigma_{y_j}, \sigma_{v_j})$ denote the axis lengths of the i-th and j-th clumps along the principal, secondary, and velocity axes.

To make simulated data more realistic, the system can add Gaussian noise at the same level as the background noise in real data. The simulated clump catalog reflects basic information about molecular clumps. Figure 2 [Figure 2: see original paper] shows integrated intensity maps along three axes for simulated molecular clouds with peak flux values ranging from 0.46 to 3, principal and secondary axis lengths from 2 to 4, velocity axis lengths from 1 to 7, rotation angles from 0° to 180°, and a signal-to-noise ratio of 0.23.

Table 1 presents the clump catalog for simulated molecular clouds (only the first five rows are shown), where Size1, Size2, and Size3 represent the full width at half maximum (FWHM) along the respective axes; Peak1–3 and Cen1–3 denote the peak and centroid coordinates; θ indicates the rotation angle of the molecular clump in the Galactic longitude-latitude plane; and Peak, Sum, and Volume represent the peak flux, total flux, and volume of the clump. Since clump central coordinates are calculated by the detection algorithm, the central coordinate values in the simulated catalog are identical to the centroid coordinate values.

2.1.2 Synthetic Data Generation To evaluate the detection rate of clump detection algorithms in specific sky regions, synthetic data are required for experimental testing. Synthetic data are generated by randomly adding several simulated clumps to real data, thereby expanding the experimental dataset. During synthetic data generation, to avoid altering the overall distribution of real clump data, the added simulated clumps should approximate real data in terms of peak flux and total flux. Therefore, we first perform statistical analysis on the peak and total flux distributions of clumps in real data to determine their distribution intervals and patterns. Based on these statistical results, when adding simulated data to real data, the peak and total flux of the added molecular clumps should collectively follow the same distribution. Simulated data are generated following the method described in Section 2.1.1, and the resulting peak and total flux distributions are combined with real data to constitute synthetic data, as shown in Figure 3 [Figure 3: see original paper]. The corresponding clump catalog is presented in Table 2 .

2.2 Clump Detection and Matching

Detecting clumps in molecular cloud data aims to generate molecular clump catalogs for subsequent scientific research. The LDC algorithm is employed to detect clumps in molecular cloud data, with results displayed through the system interface. To evaluate the stability of the molecular clump detection algorithm, catalog matching is applied to detection results from simulated or synthetic data to calculate recall and precision rates, thereby assessing the algorithm's stability and reliability.

2.2.1 Molecular Clump Detection The primary function of molecular clump detection is to identify clumps in simulated, synthetic, and real data using the Local Density Clustering (LDC) algorithm. Detection results are displayed in two text boxes on the system interface: "The number of clump" and "Detection time," indicating the number of detected molecular clumps and total detection time, respectively. As shown in Figure 4 [Figure 4: see original paper], detection of the simulated data generated in Section 2.1.1 yields 45 clumps with a processing time of 15.87 seconds. The system simultaneously displays the original data, the detection mask, and integrated intensity maps of clumps extracted from the original data using the mask. The integration direction can be switched using the Aix0, Aix1, and Aix2 buttons in the

lower-right corner. Each detection generates a detection catalog file where each column corresponds to the parameters in the simulated catalog, as shown in Table 3 . The detected values for principal axis, secondary axis, velocity axis, and volume in the results catalog are slightly underestimated because background truncation, employed to reduce noise effects, leads to smaller shape parameters, while total flux is overestimated due to additive noise. Missing rotation angle values and corrections to the principal, secondary, and velocity axes are addressed in the parameter reproduction module.

2.2.2 Catalog Matching Catalog matching evaluates the performance of molecular clump detection algorithms using three normalized metrics: F1 score, Recall, and Precision [15], with algorithm performance proportional to all three indicators. The calculation formulas are as follows:

$$P = \frac{C_N}{D_N}$$

$$R = \frac{C_N}{E_N}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

where C_N in equation (3) represents the number of correctly detected clumps and D_N represents the total number of detected clumps; E_N in equation (4) represents the number of simulated clumps.

This module accepts either individual files or folders as input parameters. Individual files refer to one simulated catalog and one detection catalog; folders refer to directories containing multiple simulated and detection catalogs. Matching results are categorized into three types: correct matches, incorrect matches, and missed detections. Figure 5 [Figure 5: see original paper] shows the catalog matching results. For the simulated data from Section 2.1.1, the matching yields a precision of 1.0, recall of 0.9, and F1 score of 0.947.

2.3 Parameter Reproduction

To mitigate noise effects on detection results, the molecular clump detection algorithm employs background truncation, which introduces deviations between detected and true values for the principal axis, secondary axis, velocity axis, and peak flux. Additionally, the detection process does not calculate the rotation angle for each clump. Instead, multi-Gaussian fitting is applied to detected molecular clumps to reproduce parameters such as principal axis, secondary axis, velocity axis, and peak flux, while also calculating the corresponding rotation angle. The calculated rotation angle either matches or is complementary to the

rotation angle in the simulated catalog. Table 4 presents the corrected results from the detection catalog in Section 2.2.1 after parameter reproduction.

2.4 3D Visualization

As three-dimensional data, molecular clumps cannot be fully understood through integrated intensity maps alone. Three-dimensional visualization compensates for the lack of spatial information in 2D representations, enhancing understanding of molecular clumps. The system platform provides multi-source display of detected molecular clumps, where 3D rendering facilitates spatial identification of clumps with different morphologies, while integrated intensity maps and slice views along different directions help researchers examine detailed information about individual clumps. This reveals distinct external characteristics of different clumps and guides investigation of their intrinsic physical property differences. Figure 6 [Figure 6: see original paper] displays the 3D rendering, integrated intensity map, and slice views of a single molecular clump.

2.5 Data Storage

Simulated and synthetic molecular clump data are significant for research on molecular clump-related algorithms, while detection of clumps in real molecular cloud data provides reliable analysis materials for researchers. The MWISP molecular cloud survey contains massive amounts of molecular clump data, and digital archival storage provides reliable preservation of these valuable data, supporting related scientific research. Observational molecular clump data can be represented as three-dimensional matrices; however, directly storing 3D matrices in databases would lose internal data relationships. To securely store 3D molecular clump data and clump catalogs in databases, the molecular clump data are first converted to binary format, after which the catalogs and corresponding data are stored in the database. Figure 7 [Figure 7: see original paper] illustrates the relationship between the molecular clump data table and the molecular clump information table.

3. Detection of Molecular Clumps in the M16 Region

The M16 region is a small portion of the MWISP project, spanning Galactic longitudes from 15°15 to 18°15 and latitudes from 0° to 1°30 . The detection results and analysis for this region are shown in Figure 8 [Figure 8: see original paper], where red dots indicate detected molecular clump positions. A total of 658 clumps were detected, with the corresponding catalog presented in Table 5 . Statistical analysis of the M16 detection catalog reveals the distributions of peak flux and total flux for molecular clumps in this region, as shown in Figures 9(a) and 9(b). In both panels, the vertical axis represents the percentage of molecular clumps, while the horizontal axes represent peak flux and total flux, respectively. The figures show that the proportion of clumps peaks at a peak

flux of approximately 4 and reaches maximum total flux around 300.

4. Conclusion

The system has completed construction of all modules. Faced with increasingly abundant molecular cloud observational data, this system can significantly reduce data processing time. Through cross-validation using multi-source data including simulated and synthetic molecular clouds, the system achieves a molecular clump detection accuracy of 0.947, providing reliable and scientific data support for related research, accelerating scientific output, and strengthening the foundation of molecular clump observations in China. For the M16 region, 658 molecular clumps were detected from observational data, providing reliable data support for related scientific studies in this region. Future work will focus on improving molecular clump detection algorithms and simulated data generation models, refining existing module functions to provide robust technical support for molecular clump and related scientific research in China.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (U2031202, 11903083, 11873093). This paper utilizes data from the MWISP project, a multi-line survey along the northern Galactic plane using the PMO-13.7m telescope in 12CO/13CO/C18O. We thank all members of the MWISP working group, particularly the staff of the PMO-13.7m telescope, for their long-term support.

References

[1] Carruthers G R. Rocket Observation of Interstellar Molecular Hydrogen [J]. The Astrophysical Journal, 1970, 161: L81-85.
[2] Wilson R W, Jefferts K B, Penzias A A. Carbon Monoxide in the Orion Nebula [J]. The Astrophysical Journal, 1970, 161 (1): L43.
[3] Heyer M, Dame T M. Molecular Clouds in the Milky Way [J]. Annual Review of Astronomy & Astrophysics, 2015, 53 (1): 583-629.
[4] Williams J P, Blitz L, Mckee C F. The Structure and Evolution of Molecular Clouds: from Clumps to Cores to the IMF [J]. Physics, 2012, 97.
[5] Kauffmann J, Pillai T, Goldsmith P F. Low Virial Parameters in Molecular Clouds: Implications for High Mass Star Formation and Magnetic Fields [J]. The Astrophysical Journal, 2013, 779 (2): 185.
[6] Krumholz M R, Mckee C F, Tumlinson J. The Star Formation Law in Atomic and Molecular Gas [J]. Astrophysical Journal, 2009, 699 (1): 850-856.
[7] 李金增. 电离氢区　分子云复合体的红外发射研究 [J]. 天文研究与技术, 1996 (01): 85-86.
Li Jinzeng. Infrared Research on the H Region-Molecular Cloud Complexes [J]. Astronomical Research & Technology, 1996 (01): 85-86.
[8] Rivera-Ingraham A, Ristorcelli I, Juvela M, et al. Galactic Cold Cores. VIII. Filament formation and evolution: Filament properties in context with evolu-

tionary models [J]. Astronomy & Astrophysics, 2017, 601: A94.
[9] Baume G, Ramirez Alegria S, Borissova J. Studying young stellar populations in G345.5+1.5 molecular cloud [J]. New Astronomy, 2022, 93.
[10] Yang S, Ji Y, Zhang S, et al. The Milky Way Imaging Scroll Painting (MWISP): Project Details and Initial Results From the Galactic Longitude of +25°.8 TO +49°.7[J]. The Astrophysical Journal Supplement Series, 2019, 240(1): 9.
[11] Yuan L, Ji Y, et al. Molecular Gas Structures traced by 13CO Emission in the 18,190 12CO Molecular Clouds from the MWISP Survey. arXiv e-prints, 2022, arXiv:
[12] Pranav P. Topology and geometry of Gaussian random fields II: on critical points, excursion sets, and persistent homology. arXiv e-prints, 2021, arXiv:2109.08721.
[13] Xiaoyu L, Sheng Z, Yao H, et al. Molecular Clump Extraction Algorithm Based on Local Density Clustering [J]. Research in Astronomy and Astrophysics, 2022, 22(1):
[14] Stutzki J, Guesten R. High spatial resolution isotopic CO and CS observations of M17 SW - The clumpy structure of the molecular cloud core [J]. Astrophysical Journal, 1990, 356 (2): 513-533.
[15] 周飘, 罗骁域, 郑胜, 江治波, 曾曙光. 一种针对 MWISP 项目分子云团块的 3DCNN 证认方法（英文）[J]. 天文学报, 2020, 61(05): 32-45.
Zhou Piao, Luo Xiaoyu, Zheng Sheng, Jiang Zhibo, Zeng Shuguang. A 3D CNN Molecular Clump Verification Method for MWISP Project(English) [J]. Acta Astronomica Sinica, 2020, 61(05): 32-45.

Note: Figure translations are in progress. See original paper for figures.