

# Mining Spatial and Spatio-temporal Patterns in Scientific Data

Hui Yang  
Dept. of Comp. Sci. & Eng.  
The Ohio State University  
Columbus, OH, 43210  
yanghu@cse.ohio-state.edu

Srinivasan Parthasarathy  
Dept. of Comp. Sci. & Eng.  
The Ohio State University  
Columbus, OH, 43210  
srini@cse.ohio-state.edu

## Abstract

*Data mining is the process of discovering hidden and meaningful knowledge in a data set. It has been successfully applied to many real-life problems, for instance, web personalization, network intrusion detection, and customized marketing. Recent advances in computational sciences have led to the application of data mining to various scientific domains, such as astronomy and bioinformatics, to facilitate the understanding of different scientific processes in the underlying domain.*

*In this thesis work, we focus on designing and applying data mining techniques to analyze spatial and spatio-temporal data originated in scientific domains. Examples of spatial and spatio-temporal data in scientific domains include data describing protein structures and data produced from protein folding simulations, respectively. Specifically, we have proposed a generalized framework to effectively discover different types of spatial and spatio-temporal patterns in scientific data sets. Such patterns can be used to capture a variety of interactions among objects of interest and the evolutionary behavior of such interactions. We have applied the framework to analyze data originated in the following three application domains: bioinformatics, computational molecular dynamics, and computational fluid dynamics. Empirical results demonstrate that the discovered patterns are meaningful in the underlying domain and can provide important insights into various scientific phenomena.*

## 1. Introduction

Data mining is the process of discovering hidden and meaningful knowledge in a data set. As an area combining ideas from database systems, machine learning, and statistical learning, data mining has been successfully applied to many application domains. Web personalization, network intrusion detection, and customized marketing are

a few examples of successful applications. Recently, researchers have started to apply data mining techniques to various scientific domains, such as astronomy, bioinformatics and computational fluid dynamics, to facilitate the understanding of the underlying scientific phenomena.

This thesis work focuses on developing data mining techniques to analyze spatial and spatio-temporal data produced in different scientific domains, where *spatial data* (e.g., geographic data) is data pertaining to the location, shape, and relationships of features. When such data is time-varying in nature, it is said to be *spatio-temporal data*. Recent advances in computational sciences have led to the production of such data in many scientific domains. Examples of spatial and spatio-temporal data in such domains include data describing protein structures and data produced from protein folding simulations, respectively.

Mining spatial relationships in these data sets is an important and interesting problem. For instance, an important issue in bioinformatics is to identify structurally similar proteins. To address this issue, we and other researchers have shown that one can first discover spatial relationships that are frequently formed by non-local patterns in protein contact maps. Structurally similar proteins can then be identified based on the presence of a subset of such frequently occurring spatial relationships [9, 11, 15]. Furthermore, it is also important to capture the evolutionary behavior of specific spatial relationships over time (i.e., spatio-temporal relationships) as such behavior can be indicator or predictor of upcoming events (e.g., vortex (hurricane) dissipation, crack propagation and amalgamation in materials) [12, 13].

However, *mining spatial and spatio-temporal relationships in scientific data* is also very challenging. First, one needs to take into account the geometric properties (e.g., shape and size) of features (e.g., vortices in fluid flows). Such information is crucial to understand or explain important phenomena in many scientific applications (e.g., vortex amalgamation in fluid flows). Unfortunately, most of the related work consider features as single points in a multi-dimensional space [6, 7, 17]. Second, there is a strong need

to develop techniques to model spatial relationships among features. These interactions if captured properly can help domain experts to understand the underlying processes in very effective manner. Moreover, these techniques need to be cognizant of domain knowledge. Third, one needs to develop effective approaches to incorporate temporal information into the overall analysis. Fourth, effective reasoning methods are needed to make inferences on important events, such as defect amalgamation in materials, based on the extracted spatial and spatio-temporal relationships. Finally, recent technological advances in computational sciences have resulted in huge amounts of data. Therefore, such approaches must scale well to large data sets.

To address these challenges, we have proposed a *generalized framework*. We begin by representing features as geometric objects instead of points. Multiple representation schemes are developed to meet the varying requirements of different scientific applications. Fig. 1 depicts the main schemes that the framework currently supports. We have also introduced multiple object-based distance measurements that take into account the shape and extent of the features, thus are robust in capturing the influence of an object on other objects in spatial vicinity. Furthermore, by consulting with domain experts, we have identified four types of spatial object association patterns (SOAP) to characterize different spatial relationships among features, namely, *clique*, *star*, *sequence*, and *minLink* (Fig. 2). In addition, we have proposed a simple yet effective approach to accommodate the temporal dimension into the mining process. This approach captures the evolution of different spatial relationships. We have also implemented several reasoning strategies to reason about domain-specific events based on the discovered patterns. Finally, we have implemented efficient and scalable algorithms to discover these patterns in large out-of-core data sets. This framework can also be easily extended to integrate new requirements arising from new applications.

In summary, we make the following contributions:

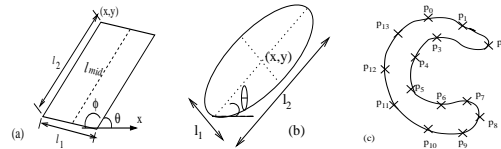
1. We present robust techniques for modeling the shape and extent of features (objects).
2. We have developed fast algorithms for extracting frequent spatial object interactions through the design of appropriate distance functions and interaction types.
3. We have developed a simple yet effective approach for mining spatio-temporal episodes of SOAP patterns. We further demonstrate that an approach that combines information from multiple SOAP models is capable of reasoning about critical events.
4. We have empirically evaluated our approaches on real case study applications and show that the algorithms scale well and are capable of processing large data sets.

We validate our framework on three case study applications drawn from the scientific and engineering community. Two of the applications analyze scientific simulation data originated in Computational Fluid Dynamics (CFD) and Computational Molecular Dynamics (CMD), and the other application is drawn from protein contact map analysis. The main challenges for these applications include feature detection, classification, and then extracting and modeling spatio-temporal or spatial interactions. Many techniques have been proposed to detect, extract and classify features from such data in the past [3, 4, 9, 14]. In this work we focus on the last aspect, namely, discovery of spatial or spatio-temporal patterns.

## 2 Basic Concepts

**Spatial Feature Representation** We have proposed three main different representation schemes: parallelepiped (or parallelogram in 2D), ellipsoid (or ellipse in 2D), and landmarks based representation, where landmarks are sampled boundary points [8]. These schemes can be used to model features from a variety of scientific domains. Parallelepipeds (or parallelograms) subsume MBBs, thus are applicable to model features in relatively regular shape. Whereas ellipsoids or ellipses are appropriate for vortices. Finally, landmarks are effective to model highly irregular-shaped features such as defect structures in materials. The number of landmarks needed to represent a feature is domain dependent. The framework also supports elemental shapes such as lines and splines.

**Dataset Representation** The data set  $\mathbb{D}$  consists of  $n$  features extracted from  $r > 1$  snapshots, taken at time steps  $t_1, \dots, t_r$  ( $t_1 < \dots < t_r$ ). (For spatial data that involves multiple maps, one can arbitrarily assign a unique ID to each map.) The  $n$  features are further categorized into  $l$  types, where the categorization is governed by the underlying domain. A feature’s geometric properties are captured by one of the three representation schemes described earlier. Note that the terms *features* and *spatial objects* are used interchangeably in the context.



**Figure 1.** Shape Representations: (a)Parallelogram (b)Ellipse (c)Irregular

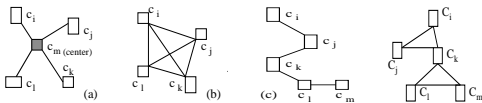
**Shape-based Distance Measurements** The framework supports three main distance measurements for two objects  $o_i$  and  $o_j$  in the same snapshot: (1) *Point-Point*: This is

simply the Euclidian distance between object centroids; (2) *Line-Line*: It is the shortest distance between two line segments identified for the two involved objects respectively. For instance, if an object is represented as an ellipse, one can take the major axis as the identified line segment; and (3) *Boundary-Boundary*: This is the shortest pairwise distance between the landmarks of  $o_i$  and  $o_j$ . Notice that the last two measurements take objects' geometric properties into account. The framework also supports Hausdorff distance [2].

**Spatial Relationships** Two objects  $o_i$  and  $o_j$  have a *closeTo* relationship if the distance between them is  $\leq \epsilon$ , a user-specified parameter. Two objects are *neighbors* if they have a *closeTo* relationship. We also consider the *isAbove* relationship, concerning the above/below spatial relationship, between two objects,

**Spatial Object Association Pattern (SOAP)** A SOAP characterizes the *closeTo* or *isAbove* relationships among multiple object types. The framework supports the discovery of four SOAP types: *Star*, *Clique*, *Sequence*, and *minLink* (Fig. 2). These SOAP types can be abstracted as undirected graphs. In such graphs, a node corresponds to an object-type and an edge indicates a fulfilled *closeTo* or *isAbove* relationship between two object types.

- *Star SOAPs* (Fig. 2a) have a *center* object-type, which is required to have a *closeTo* relationship with all the other object-types in the same SOAP.
- *Clique SOAPs* (Fig. 2b) require that a *closeTo* relationship holds between every pair of involved objects.
- *Sequence SOAPs* (Fig. 2c) identify sets of objects that are spatially arranged in the following manner: (1) all involved objects in a SOAP has a total order; and (2) two adjacent objects in a SOAP meet both the *closeTo* and *isAbove* criteria.
- *minLink SOAPs* (Fig. 2d) are a parameterized SOAP type, where the value of *minLink* is user-specified. Informally, a *minLink=l* SOAP requires that every involved feature has at least  $l$  neighbors in the same SOAP. Note that the set of *minLink=l* SOAPs subsumes all the other three SOAP types.



**Figure 2.** SOAP Types:(a)Star (b)Clique (c)Sequence (d)minLink=2

The above SOAP types capture three basic spatial relationships: distance-based, topological, and directional. The *closeTo* relationship is both distance and topology based.

As for directional relationships, we currently only consider the *isAbove* relationship, captured by *Sequence* SOAPs. Such relationships are important in understanding different interacting behaviors in many scientific applications. We plan to implement other types of spatial relationships and examine their usages for different applications in the future.

**Frequent and Prevalent SOAPs** We define two measures, *support* and *realization*, to characterize the importance of a SOAP. The support of a SOAP  $p$  is the number of snapshots in the data set where  $p$  occurs. Assume  $\text{support}(p)=s$ , let  $n_i$  be the number of  $p$ 's instances in the  $i^{\text{th}}$  snapshot where  $p$  appears,  $\text{realization}(p)=\min\{n_i\}$ . A pattern  $p$  is *frequent* if  $\text{support}(p) \geq \text{minSupp}$ , and *prevalent* if  $\text{realization}(p) \geq \text{minRealization}$ . Both *minSupp* and *minRealization* are user-specified parameters.

**Spatio-temporal Episodes** Spatial relationships among objects (or features) evolve over time. As a result, SOAPs of different types also evolve over time. We identify the following three evolutionary events for a SOAP  $p$  to characterize the stability of interactions among different features. (1) *Formation*: when the number of  $p$ 's instances changes from zero to non-zero; (2) *Dissipation*: when all  $p$ 's instances become invalid. The dissipation of a SOAP can occur due to many reasons. For example feature(s) involved in a SOAP may cease to exist or merge into a new one; and (3) *Continuation from time  $t_i$  to time  $t_{i+1}$* : if there exists at least one instance of  $p$  in each snapshot taken in  $[t_i, t_{i+1}]$ .

Formation and dissipation can occur to a SOAP many times. Thus a SOAP can exist in multiple disjoint temporal intervals, where each interval starts at a formation event and ends at a dissipation event. We refer to a SOAP's continuation in each of such intervals as a *spatio-temporal episode*.

**Main Objectives** Given a scientific data set, our main objectives include: (1) efficiently discovering different types of frequent and prevalent SOAPs; (2) generating spatio-temporal episodes; and (3) validating and analyzing the mining results by integrating domain knowledge. To meet these objectives, other challenges include feature detection, classification, and representation. Such tasks are generally governed by the underlying application domain.

### 3 Framework and Implementation

An overview of the framework is given in Fig. 3. This framework consists of six main tasks and is mainly motivated by the following three scientific applications: defects simulation in Molecular Dynamics, protein structure analysis in Bioinformatics, and vortex simulation in Computational Fluid Dynamics. We have implemented techniques to detect, extract, and classify features for the first two applications [5, 9]. For the third application, we use existing algorithms to detect vortices and subsequently classify the

detected vortices [4]. In this article, we focus on the four tasks enclosed by the dashed rectangle in the figure.

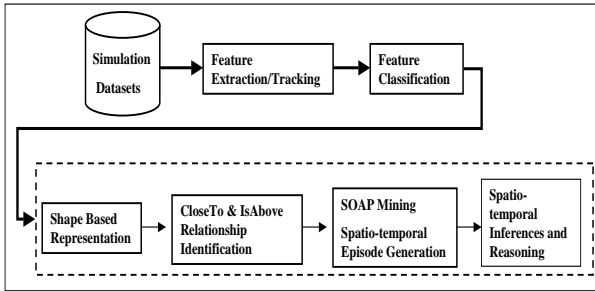


Figure 3. Overview of the framework

### 3.1 SOAP Mining

Due to lack of space, here we only present the main strategies to realize such a framework. Such strategies are employed to ensure an efficient and effective implementation for large out-of-core data sets.

**Data Organization** We organize the data in a format analogous to the vertical format used for association rule mining [16]. Such a format groups features of the same type together and further orders features within a group by time and location. This format allows efficient elimination of infrequent SOAPs at an early stage, thus greatly reduces the search space for the following computational tasks.

**Equivalence Classes** To improve spatial and temporal locality while generating SOAPS, we adopt equivalence classes to organize the discovered SOAPS. An equivalence class consists of SOAPS of the same type, same size, and different at the last object. As a result, our algorithms can efficiently discover different types of SOAPS from large amounts of data.

**Anti-monotone Property** Similar to traditional association mining for transactional data [1], we explore the anti-monotone property, i.e., a set of objects is frequent only if all of its subsets are frequent, to further reduce the search space while generating frequent SOAPS.

**Optimizations** We also introduce several optimization strategies to quickly identify valid neighbors of an object, eliminate infrequent SOAPS at an early stage, and to realize fast spatial join operations.

### 3.2 SOAP Episodes Generation and Analysis

For each frequent SOAP, we construct its spatio-temporal episodes by identifying the associated formation and dissipation events. We then use these episodes to address two important issues: (1) to reason about critical events such as the merging of multiple features; and (2)

to model how interactions among a certain set of features evolve over time. Please refer to our previous work [12] for solutions to these issues.

## 4 Empirical Evaluation

To evaluate the efficacy of this framework, we have applied it to the following scientific applications: (1) Protein structural analysis in *bioinformatics*. By discovering spatial patterns in protein structural data, we have identified a set of structural “fingerprints” of a protein class. For instance, we have identified spatial patterns that can distinguish  $\alpha$ -proteins from other proteins such as  $\beta$ -proteins; (2) Analysis of vortex evolution in *computational fluid dynamics* (CFD). By applying the framework to CFD simulation data, we have shown that the extracted spatio-temporal patterns can effectively model the evolution of certain interactions among vortices. For instance, we have identified patterns to capture the interactions that lead to the amalgamation as well as dissipation of vortices; and (3) Analysis of defect evolution in *molecular dynamics*. By applying the framework to simulation data on defects in materials, we have demonstrated that the proposed spatio-temporal patterns can capture defect-defect interactions that lead to several important events, for instance, the interactions that lead to the creation of large-extent defects. The results have also been presented to and validated by domain experts.

Due to the space constraint, we will not report detailed empirical results here. Please refer to our previous work [10, 11, 12, 13] for details.

## 5 Conclusion and Ongoing Work

To conclude, we have proposed a general framework for mining spatial and spatio-temporal patterns in scientific data sets. The framework models features as geometric objects rather than points. It also supports multiple distance measurements that take into account objects’ shape and extent and thus are more effective in capturing interactions among objects in the vicinity. We have developed algorithms to discover four different types of spatial object association patterns. We have also accommodated temporal information in the overall analysis to characterize the evolutionary behavior of an interaction. Finally, we have identified effective approaches to reason about critical events.

We are currently extending the framework to address the following aspects: (1) capturing other types of spatial relationships such as topological and directional relationships; (2) discovering rare but important patterns; and (3) analyzing object-based trajectories. Such trajectories take into account properties such as shape and size of a feature and thus allow for the effective representation of evolving features. We are also in the process of applying this framework

to identify meaningful spatio-temporal patterns in protein folding simulation data.

**Acknowledgments:** Hui Yang would like to extend her gratitude to Dr. S. Parthasarathy for his guidance and encouragement through her PhD studies. She would also like to thank Dr. J. Wilkins, Dr. T. Lenosky and S. Mehta for providing and helping validate the results in Molecular Dynamics, and thank Dr. R. Machiraju, Dr. D. Thompson, M. Jankun-Kelly and K. Marsolo for valuable comments, discussion and validation of results pertaining to Fluid Dynamics and bioinformatics.

## References

- [1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases VLDB*, Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, pages 487–499. Morgan Kaufmann, 12–15, 1994.
- [2] M. J. Atallah. A linear time algorithm for the hausdorff distance between convex polygons. *Information Processing Letters*, 17:207–209, 1983.
- [3] C. Henze. Feature detection in linked derived spaces. In *IEEE Conference on Visualization*, 1998.
- [4] M. Jiang, T. Choy, S. Mehta, S. Parthasarathy, R. Machiraju, D. Thompson, J. Wilkins, and B. Gatlin. Feature mining paradigms for scientific data. In *SIAM Data Mining Conference*, 2003.
- [5] S. Mehta and S. Barr. Dynamic classification of defect structures in molecular dynamics simulation data. In *SIAM Data Mining Conference*, 2005.
- [6] Y. Morimoto. Mining frequent neighboring class sets in spatial databases. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 353–358, 2001.
- [7] R. Munro, S. Chawla, and P. Sun. Complex spatial relationships. In *The 3rd IEEE International Conference on Data Mining (ICDM2003)*, 2003.
- [8] C. R. Rao and S. Suryawanshi. Statistical analysis of shape of objects based on landmark data. *Proceedings of National Academy Science, USA*, 93(22):12132–12136, 1996.
- [9] H. Yang, K. Marsolo, S. Parthasarathy, and S. Mehta. Discovering spatial relationships between approximately equivalent patterns. In *The 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD)*, 2004.
- [10] H. Yang, S. Mehta, and S. Parthasarathy. A generalized framework for mining spatio-temporal patterns in scientific data. In *Technical Report OSU-CISRC-5/05-TR14, Ohio State University*, 2005.
- [11] H. Yang, S. Parthasarathy, and S. Mehta. Mining spatial object associations for scientific data. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.
- [12] H. Yang, S. Parthasarathy, and S. Mehta. Towards association based spatio-temporal reasoning. In *Proceedings of the 19th IJCAI Workshop on Spatio-temporal Reasoning*, 2005.
- [13] Hui Yang, Srinivasan Parthasarathy, and Sameep Mehta. A generalized framework for mining spatio-temporal patterns in scientific data. In *Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 716–721, 2005.
- [14] Kenneth Yip. Structural inferences from massive datasets. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI) (1)*, pages 534–541, 1997.
- [15] M.J. Zaki. Mining protein contact maps. In *The 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD)*, 2003.
- [16] M.J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. Technical Report TR651, Rensselaer Polytechnic Institute, 1997.
- [17] X. Zhang, N. Mamoulis, D. W. Cheung, and Y. Shou. Fast mining of spatial collocations. In *KDD '04: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 384–393, 2004.