

Defect Detection in Silicon and Alloys

M. Coatney, S. Mehta, A. Choy, S. Barr, S. Parthasarathy, R. Machiraju, and J. Wilkins.
The Ohio State University, Columbus, OH, USA
Contact: srini@cis.ohio-state.edu

Abstract

Understanding the structural stability and dynamics of material defects are fundamental to materials science. Molecular dynamics (MD) simulations offer the opportunity to study these dynamics. Understanding the underlying dynamics has typically involved the exploration of the simulation results in a post-processing retroactive phase. Analyzing this data is a complex problem, due to thermal noise present in the system and since the data produced by such simulations can easily lead to terabytes. In this work we present a data mining methodology that can detect the presence of defect structures in materials. Our work relies on a novel utilization of frequent pattern algorithms, i.e., frequent patterns are used to prune away non-defect elements within the material. Preliminary results with this approach are quite encouraging, and show that the proposed method is quite adept at extracting defects from silicon lattices and alloy lattices.

1 Introduction

Silicon is an important semiconductor given its use in chip fabrication. In order to speed up chips the size of transistors and powerful connections in each new generation of chips are designed to be smaller than the previous ones. At present, the size of individual elements in chips is in the sub-micron scale. Further shrinking these elements brings them dangerously close to the scale of extended defects, which could render the chips useless. These extended defects are created by boron implantation and thermal process in

chip production. The mechanism of their formation from smaller point defects is still unknown. We seek to understand the evolution, classification and propagation of defects that affect the physical and chemical properties of Silicon.

We study the dynamics of point defects in silicon lattice using Molecular Dynamics (MD) simulations[?], which have thus far generated tens of millions of time frames detailing important phenomena such as defect formation, interaction and diffusion. However extracting such interesting events from simulation results is extremely difficult given the size of the data sets, the complexity of defect structures, as well as the fact that these events occur in multiple time scales.

Si has four valence electrons. It shares these electrons with four neighboring silicon atoms resulting in a crystalline diamond like structure. Defects exist in regular lattice which force it to deviate from its regular structure. Currently our focus is on detecting interstitial and vacancy defects. An Interstitial defect [?, ?] is caused by one or more atom within lattice being located at some non-crystallographic position whereas valency defects are due to absence of one or more silicon atoms from lattice. This causes structure to change local bonding configuration, energy and potential.

Our goal is to develop techniques to detect, isolate, identify, and track defects in datasets generated by MD simulations[?]. Our focus is currently on the Si lattice however our algorithm is general enough to handle other semiconductors and metal alloys.

In section 2 we describe work related in Data Mining literature. Section 3 and 4 describe our methods and some of the results.

2 Background

The field of knowledge discovery and data mining, spurred by advances in data collection technology, is concerned with the process of deriving interesting and useful patterns from large datasets. Most work in this field has focused on analyzing financial and business datasets. More recently, there has been wide spread interest toward developing techniques that can analyze large scientific and biomedical datasets[?]. Here, the objective is to mine information that can fundamentally advance the science through mining techniques.

Frequent pattern mining is an important class of data mining techniques[?]. In several domains, a frequently occurring pattern is often a pattern that is of paramount interest. Examples included frequent customer behavioral patterns, recurring protein structure or drug structure patterns etc. There have been a variety of algorithms that model frequent associations in transactional data, frequent sequences in scientific and transactional data, and frequent structures in biochemical and physical data. The seminal work by Agrawal and his colleagues at IBM largely targeted business datasets[?]. More recently work by Dehaspe *et al.*[3], Djoko *et al.*[7] and Wang *et al.*[9], explored frequent substructures in small chemical compounds. However, recently macromolecules [2, 5, ?, 8, 10], like proteins and nucleic acid have also received increased attention.

In this work we leverage the ideas presented by Li, Parthasarathy and Coatney [2, ?] for mining molecular dynamics simulation data. One twist is that in that work the authors are actually seeking for frequent patterns that relate to the protein function. Here, we are interested in defects that do not occur frequently. Instead, we use the above frequent pattern approach to *prune* the atoms that are not part of the defect, enabling us to localize the defects. We exploit the noise handling capability of algorithm for finding defect in MD simulation. One of the noteworthy points of algorithm is its relative independence to underlying semiconductor domain which al-

lows it to easily adapt to alloys. We discuss this algorithm¹ briefly next.

3 Proposed solution

The Silicon(Si) lattice is represented as a 3-d graphs where nodes are atoms and edges represent the bonds between atoms. The underlying idea for defect detection is based on fact that much of the Si lattice is composed of bulk atoms where bulk atoms are defined as atoms which conform to perfect silicon lattice structure. In Figure 1 atoms marked as white are bulk and black colored atoms are defect. After pruning the bulk we will be left with defect.

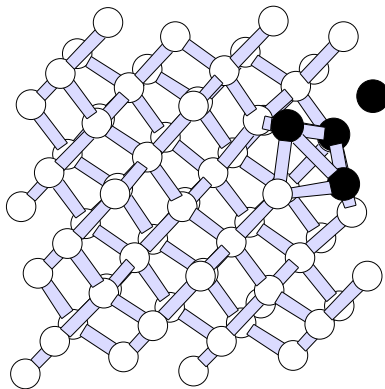


Figure 1: Defect in Silicon

As mentioned earlier, to prune the bulk we use a novel variant of frequent atomsets mining. Our approach relies on using frequent atomsets to prune the bulk atoms. Ideally, what should be left over are the defect atoms.

The algorithm starts with generating 2-atom sets from the lattice. Theoretically, there can be $\binom{n}{2}$ 2-atom sets in an n-atom lattice. However the interaction between two atoms decreases as the distance increases. Therefore we introduce a user specified range criteria which specifies the maximum allowable Euclidean distance between the two atoms for them to be considered as a atomset. For Si lattice this range is estimated

¹For further details on the algorithm the reader is referred to Parthasarathy and Coatney[?].

to be 3 times the bond length. This condition is strictly enforced while generating 2 -atomsets but relaxed while building larger atomsets. Thus two atoms may be included in a atomset even if they do not satisfy the range criteria as long as they are connected by atoms satisfying range.

The pruning of infrequent k -atomsets is conducted by counting the number of each k -atomsets and retaining only those whose count is above a user specified minimum support. Then, we prune the frequent atomsets from the original lattice.

Our algorithm is robust to noise inherently present in datasets obtained from simulation. We discretize the Euclidean distance between two atoms into equiwidth bins and the number of bins measure the resolution. The binning or quantization not only enhances performance but also allows minor deviations in distances due to lower resolution.

Due to presence of noise, exact matching of atomsets for pruning is prohibitive. Hence an approximate match is sought. We use Recursive Fuzzy Hashing (RFH) for this purpose. RFH is a technique in which atomsets are hashed to its exact and neighboring bins. The number of neighboring bins is an user-specified parameter which specifies the number of neighboring bins in which hashing is done.

4 Results

We now show the effectiveness of the RFH method. All the experiments are performed on files containing metastable defect structure. These metastable defect structure are obtained using wavelet analysis, which outputs the time-averaged positions of atoms in the time frames during which the defect is stable.

Figure 2a shows a Si lattice with a single I3-defect. This type of defect is induced by 3 Silicon atoms present in non-crystallographic locations. Since we simulate the defects at about 1000K, there is still considerable thermal fluctuation in the time-averaged position of atoms. At present, each structure found in the simulation is first brought down to zero temperature. How-

ever, this requires significant increase in computational time. Our approach works well even in presence of such thermal noise. Figure 2b shows the detected defect in a noisy Si lattice. All detected defect atoms are marked black.

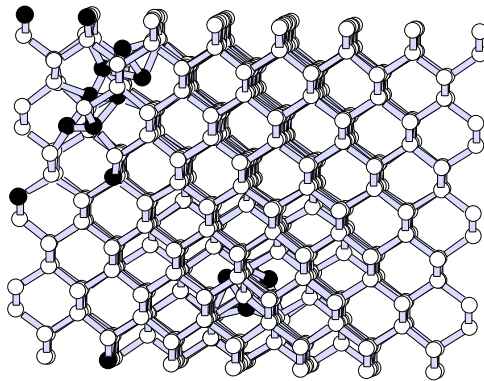


Figure 3: 512-atom Si lattice with I3 and I1 defect

Two defects in a single lattice are also detected correctly Figure 2c shows a 66-atom lattice with one I3 and one I1 defect. The approach works well with large lattice also. Figure 3 shows a 512-atom lattice with two defects. Since the algorithm is not using domain specific knowledge it adapts well for alloys also. Figure 4 shows Ni_3Al with one interstitial defect which is correctly identified and marked.

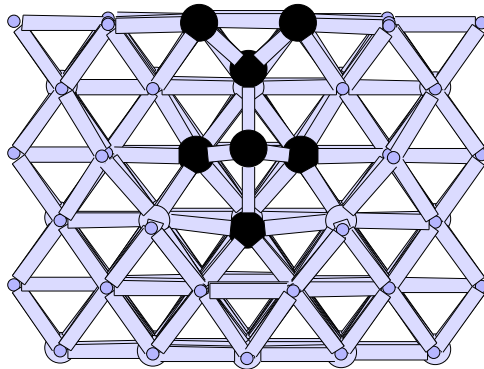


Figure 4: Interstitial Defect in Ni_3Al

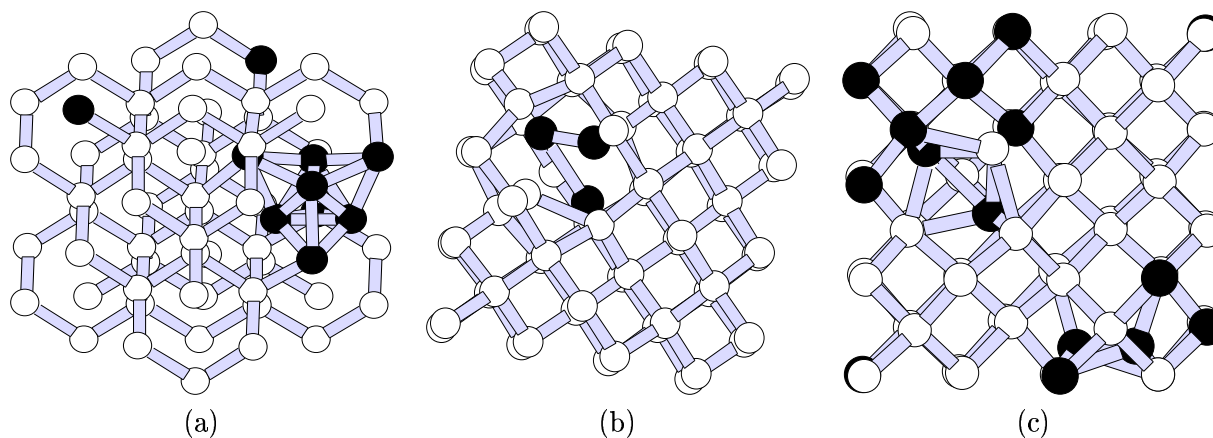


Figure 2: (a) Silicon Lattice with I3 defect (b) Noisy Silicon Lattice with I1-defect (c) Two Defects in 66-atom lattice

5 Summary and Future Work

In this work we describe applications of data mining algorithms to Si and alloy lattices. Our defect algorithm was derived from a frequent atomset mining algorithm. This algorithm provides satisfactory results. However, there is a need for more efficient algorithms which do not discover the bulk before unearthing the defect. Also, in future we will apply our algorithm to even larger lattices having larger number of various defects.

References

- [1] Daniel Fischer et.al. A Geometry-based Suite of Molecular Docking Processes. In *Journal of Molecular Biology*, pages 459–477, 1995.
- [2] H. Li and S. Parthasarathy. Automatically Deriving multi-level protein structures through data mining. In *HiPC Conference Workshop on Bioinformatics and Computational Biology*, 2001.
- [3] L. Dehapse, H. Toivonen and R. King. Finding Frequent substructures in chemical compounds. In *The Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1998.
- [4] R. Agrawal and R. Srikant. Fast Algorithms for mining association rules. In *VLDB*, 1994.
- [5] R. King, A. Karawath, A. Clare and L. Dehapse. Genome scale prediction of protein functional class from sequence using data mining. In *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.
- [6] R. Nussinov and H. Wolfson. Efficient Detection of three dimensional Structural Motifs in Biological Macromolecules by Computer Vision Techniques. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 88, Dec 1, 1991.
- [7] S. Djoko, D. Cook and L. Holder. Analyzing the benefits of domain knowledge in substructure discovery. In *The First ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1995.
- [8] W. Pan, J. Lin and C. Le. Model-based cluster analysis of microarray gene-expression data. In *Genome Biology*, 2002.
- [9] X. Wang et.al. Automated discovery of active motifs in three dimensional molecules. In *The Third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1997.
- [10] X. Zheng and T. Chan. Chemical Genomics: A systematic approach in biological research and drug discovery. In *Current Issue in Molecular Biology*, 2002.
- [11] Y. Lamdan and H. Wolfson. Geometric Hashing: a general and efficient model-based recognition scheme. In *Proceedings of the second ICCV*, pages 238–289, 1988.