

MotifMiner: A General Framework for Efficiently Identifying Common Substructures in Molecules *

Matt Coatney and Srinivasan Parthasarathy
Computer and Information Science, Ohio State University
{coatney,srini}@cis.ohio-state.edu

October 21, 2002

Abstract

1 Abstract

Much of biochemical research involves using structural similarities and differences between molecules to classify or predict function. Traditionally, researchers have identified these patterns, or motifs, manually using chemical expertise. However, with the massive influx of new chemical data and the ability to gather data for very large molecules, there is great need for techniques that automatically and efficiently identify commonly occurring structural patterns in molecules. Previous automated substructure discovery techniques have suffered from several limitations. They incorporate domain-specific chemical knowledge, preventing them from being applicable for arbitrary datasets. Also, they do not address scalability or noise, which is critical for macromolecules. In this paper, we present MotifMiner, a general framework for automatically identifying common motifs in most any chemical dataset. We describe both our application framework and algorithm for incrementally identifying common motifs,

as well as demonstrate the flexibility of our system by analyzing several disparate domains, including protein, transfer-RNA, drug, and crystal lattice datasets.

2 Introduction

Current biochemical research is driven by the structure-activity relationship (SAR), which dictates that the three-dimensional structure of a molecule exactly determines its function. The mechanisms of drug activity, toxicity, and disease can typically be described in terms of structure-level interactions between portions of molecules. The ability to characterize important, frequently occurring structural patterns (also known as structural classes or motifs) is therefore a critical first step in determining these mechanisms, classifying molecules based on function, and predicting the function of new molecules.

In the past, researchers have identified these structural motifs manually, sifting through hundreds of compounds with similar behavior visually looking for patterns. However, recent technological advances have led to an explosion of structural data that has far surpassed the feasibility of manual analysis. For example, pharmaceutical companies have corporate

*This work was partially supported by an Ameritech Faculty Fellowship.

databases with millions of compounds, many of which have never been thoroughly analyzed. Macromolecule databases such as the Protein Data Bank (PDB, <http://www.rcsb.org/pdb>) also have increasing numbers of large, complex, noisy structural coordinate data that are of interest to genetics and proteomics researchers. In addition, molecular simulations of chemical phenomena have produced massive amounts of structural data. An example of this is the simulation of crystal lattice defects, useful especially in predicting and minimizing defects in silicon wafers used in computers.

What is needed is a mechanism for automatically and efficiently detecting motifs in chemical data. Initial attempts have focused primarily on the pharmaceutical domain and have not addressed flexibility, scalability, and noise handling required for considering alternative chemical domains. In this paper we introduce MotifMiner, a general framework that is capable of efficiently identifying frequent substructures in a variety of structure datasets. Our approach is even capable of finding motifs in macromolecules, which have an incredibly large search space. For instance, proteins can have thousands of atoms and billions of possible multi-atom substructures.

The framework transparently supports many different chemical domains by providing a series of parsing utilities, and new formats can be supported simply by adding a new parser. These parsers convert the disparate structural data into a common internal representation suitable for efficient mining. This representation only considers atoms and their three-dimensional locations; physical bonds are discarded in favor of virtual "mining bonds", which describe spatial relationships between pairs of atoms. This has two main advantages; it allows those datasets without bonds (i.e. from X-ray crystallography) to be analyzed, and more importantly it allows detection of *non-covalent* interactions. Previous techniques have typically only allowed physically con-

nected substructures, which do not capture non-covalent spatial interactions.

The algorithm uses an incremental frequent pattern mining algorithm to drastically reduce the search space and provide for scalable analysis of large molecules. Substructure matching can be relaxed to allow for approximate motif definition, thus handling noisy coordinate data like that obtained by X-ray crystallography. The algorithm also has several computational optimizations that assist in improving performance and scalability. The original algorithm identified frequent intra-molecular substructures; it has since been expanded to support analysis of frequent inter-molecular substructures as well.

To demonstrate the general utility of MotifMiner, we examine several fundamentally different datasets. We identify frequent intra-molecule motifs in several different proteins and a transfer-RNA. We then identify frequent inter-molecule motifs in three classes of drugs. Last, we identify unique defects in silicon crystal lattices by identification and removal of the highly uniform lattice bulk.

3 Application Framework

3.1 Structure and Motif Representation

Molecules are represented in a variety of ways, depending on the chemical domain's needs and current technology. One representation is a three-dimensional coordinate graph, where atoms are nodes and bonds are edges. This is applicable primarily to small organic compounds such as drugs, where the complete chemical structure is known. An example is the commonly used MDL MOL format. This type of representation is insufficient for molecules which only have their atom coordinates defined. This may be due to limitations in empirical techniques such as X-ray crystallography or simplifications made in molecular simulation software. Ex-

amples of these include the PDB format for macromolecules and the XYZ format common in MD simulations.

In order to provide a common framework for efficiently mining structural data regardless of source, we convert these and other representations through a series of input parsers into an internal structural representation. This representation stores only the atoms and their three-dimensional Euclidean coordinates. Actual bonds are ignored; instead, during our algorithm we generate virtual bonds known as “mining bonds”. These bonds describe the atom types they connect and the distance between them; thus, actual bonds are implicitly handled through these distance relationships, and even normalized aromatic bonds are handled without specialized chemical domain knowledge. In addition, non-connected spatial interactions are described in this manner.

In our framework, we define several concepts related to substructures. A *motif* is a commonly occurring structural pattern, described by a set of atoms. The motif captures all information of a three-dimensional graph in a form that facilitates quick comparison without the need for coordinate translation. We store three-dimensional information between a pair of atoms, A_i and A_j , in a *mining bond*. The mining bond $M(A_i A_j)$ is a 3-tuple of the form

$$M(A_i A_j) = \{A_i \text{type}, A_j \text{type}, \text{distance}(A_i A_j)\}$$

A k -motif X , which is a substructure containing k atoms, is then defined as a tuple of the form

$$X = \{\mathbf{S}_X, A_1, A_2, \dots, A_k\},$$

where A_i is the i^{th} atom and \mathbf{S}_X is the set of mining bonds describing the motif.

An *atomset* is a concrete instance of a motif for a particular structure. There may be multiple atomsets representing the same motif in a structure, if the substructure occurs more than once. Two atomsets X

and Y are considered to belong to the same motif if $\mathbf{S}_X = \mathbf{S}_Y$. While stereochemistry is not explicitly handled by this representation, it can be implicitly handled by appending a chirality label (e.g. L or R) to the atom type, such that different stereoisomers produce different, non-equivalent atomsets.

In order to keep the search space and atomsets from exploding, we employ a novel yet general chemical domain technique known as range pruning. Essentially, the user specifies a range, in angstroms, beyond which chemical interactions are considered negligible and direct atom-atom relationships can be ignored. This does not prevent atoms beyond this range from being included in larger substructures, so long as there exist intermediate atoms within range of each other and the extremal atoms that satisfy the constraint. For a more thorough discussion and analysis of range pruning, please see [?].

3.2 Noise Handling

One of the hallmarks of our general framework is its ability to handle both slight noise fluctuations as well as high levels of coordinate noise chronic to macromolecule and crystal lattice datasets. For slight noise, we provide simple discretization of the raw Euclidean distance between two atoms. We do so through equiwidth binning, where each bin is of size *resolution* specified by the user.

For more extreme coordinate noise, we define a concept known as fuzziness. With fuzziness, atomsets are considered to be equal if all of their mining bonds are within a certain number of resolution bins of each other. Fuzziness is a non-negative integer; a value of 0 specifies that no fuzziness should be used and that exact matching is enforced, while a value of $n > 1$ allows matching of mining bonds within $2n * \text{resolution}$ angstroms of each other. For many algorithms, this fuzziness results in more expensive computation but is necessary for handling

noisy datasets.

For most small chemical datasets, fuzziness is not necessary. For macromolecules and relaxed crystal lattices, fuzziness of 1 is sufficient to identify even large substructures. Fuzziness values greater than 1 may be useful for coordinate sets with extreme levels of noise, such as unrelaxed crystal lattices. This is currently a field of active research.