Supporting Binding-Sites Discovery via Iterative Database Processing

Ran Tel-Nir, Technology and Information Systems Program, Tel Aviv University, Tel Aviv, Israel
Roy Gelbard, Information Systems Program, Bar-Ilan University, Ramat-Gan, Israel
Israel Spiegler, Technology and Information Systems Program, Tel Aviv University, Tel Aviv, Israel

ABSTRACT

Recognition of binding sites common to a set of protein structures is important for applications such as drug design. Common methods of binding-sites are based on heuristic algorithms that use summarized spatial data and superimposition techniques. However, computational operations generally do not store intermediate data for further calculation and information extraction. The current study presents an alternative approach to binding calculation by introducing a binary representation scheme for three dimensional molecule data and a fast iterative algorithm which obviates the need to calculate and resolve spatial transformations in the binding site extraction process. This is achieved by using relational database indexing methods and an efficient iterative model. This general-purpose iterative algorithm was tested for binding small molecules. The results show that the method can be applied efficiently for binding site extraction, and bio-information extraction. This binary representation improves performance by reducing processing time by 31% compared to typical representations.

Keywords: Binary-Positive Representation, Binding-Site Search, Iterative Algorithm, Protein Structures, Structural Bioinformatics

INTRODUCTION

The extraction of a potential binding site is a major step in solving drug design problems. Matching polyhedra that represent the spatial coordinates of the pseudo-center representation of molecules, is an important stage in the binding site search process. Possible matches for high order polyhedra (i.e. matches of six points in space or more) can consume hundreds of gigabytes of memory. The number of comparisons used for the extraction of matched polyhedra exceeds the limits of current computational power (the problem is NP hard) or computer memory, when done in real time. To overcome this problem, heuristic algorithms have been developed to extract matched higher order polyhedrons from matched triangles using complex algorithms based on manipulating the appropriate spatial transformations (Shatsky et al., 2013).
al. 2006). Among these are the geometric hashing method developed by Wolfson (Wolfson & Lamdan 1988; Wolfson 1990).

These algorithms are based on computational operations done directly in the computer memory which generally does not store intermediate results for further calculation or information extraction.

Relational databases, however, may make it possible to store intermediate results of large data calculations for further use, and manipulate complex computations which use data stored in large databases. This paves the way for extracting possible matched polyhedra by the direct creation of high order polyhedra (n points in space for example). This is done by calculating all the permutations of selecting n points out of m points (where m is the size of the pseudo-centered representation of a molecule). When comparing two molecules for binding site extraction in the direct method we compare all the permutations of selecting n points out of m with the combinations of selecting n points out of m’ (where m’ is the size of the pseudo-centered representation of the compared molecule).

However, when calculating all the possible permutations and combinations of possible matched polyhedra of order n equals six, a huge number of more than 100 Gigabytes of data storage is required for the match extraction process and more than $10^{14}$ comparisons are needed. Thus, the extraction of matched higher order polyhedra is unrealistic through direct polyhedron creation methods. Database methods can extract possible matched polyhedra using large data storage, but an efficient algorithm is also needed to reduce the number of calculations and the size of the intermediate stored data.

The goal of this study is two-fold: (1) design and construct an alternative model for extraction of higher order polyhedra that is not memory bound, and (2) construct a new representation model for bio-data that can be manipulated by standard database technology and thus compatible with data mining methods. The model introduced here implements an iterative algorithm that is based on the creation of possible matched higher order polyhedra from the previously extracted matched polyhedra of lower orders. When using this model of emerging solutions the number of possible matched polyhedra is dependent on the number of extracted matched polyhedra of lower order, which is coherent with the physico-chemical properties of the molecules examined. Moreover, this number is independent of the polyhedron order. The findings show that this iterative recognition model along with the use of advanced binary representation of the bio-data is more efficient than the direct recognition model. The iterative recognition model also extracts better matches and higher polyhedron orders than the direct method. The model and system works faster, performs on a par with existing models and can be a valuable tool in structural bioinformatics research.

**Structural Bioinformatics and Binding-Site Search**

The main source of information about 3D structures of macromolecules (including proteins, peptides, viruses, protein/nucleic acid complexes, nucleic acids, and carbohydrates) is the Protein Data Bank (PDB) (Berman et al. 2000), whose format is the de-facto standard for the exchange of structural information. Most of these structures are determined experimentally by means of X-ray diffraction or nuclear magnetic resonance (NMR) studies (Tisdall, 2001). Today the PDB has more than 96,000 (PDB2 2013) structures, and is expanding constantly with international advances in structural genomics.

The PDB files are human-readable ASCII coded flat files. The databank is organized with one structure per file and it is accessible via the Internet. It contains descriptive information about the protein including its structure, amino acid construction, secondary structure and backbone physical structure. The protein structure file format is column oriented. Thus PDB databanks are merely data stores of huge amounts of data in a primitive representation
Evidence-Based Combination of Weighted Classifiers Approach for Epileptic Seizure Detection using EEG Signals
www.igi-global.com/article/evidence-based-combination-weighted-classifiers/77929?camid=4v1a

Problems-Solving Map Extraction with Collective Intelligence Analysis and Language Engineering
www.igi-global.com/chapter/problems-solving-map-extraction-collective/23069?camid=4v1a