Molecular Informatics and Image Analysis

Rahul Singh
Department of Computer Science and Center for Computing in Life Sciences
San Francisco State University, San Francisco, CA 94132
rsingh@cs.sfsu.edu

Abstract

Recent advances in structure elucidation techniques have led to a significant upsurge in the availability of three dimensional molecular structures. This includes structural information on both large molecules such as proteins and enzymes as well as small molecules used in drug discovery. At the state of the art, molecular information analysis involves a lot of 3D image like data which can be both volumetric and surface-based. Analysis, management, and assimilation of such data lead to algorithmic problems where techniques motivated by research in image analysis, and computer vision can be applied. Promulgating this perspective, we introduce the area of molecular informatics and consider its fundamental problems where image and pattern analysis techniques can play a critical role.

1. Introduction and Molecular Informatics

The role of imaging and image-analysis (including computer vision), is rapidly gaining prominence in investigations related to life sciences. This is not only true for established areas such as proteomics (gel-analysis, localization and immunohistochemistry), cell-based assays, and biomedical imaging, but also for problems such as biological phenotype analysis [3]. However, an area, where the power of interfacing image analysis techniques with life sciences has yet to be fully leveraged, is that of *molecular informatics*.

Advances in techniques such as high-resolution nuclear magnetic resonance (NMR), synchrotron radiation sources, and crystallography have led to a significant upsurge in the availability of three dimensional molecular structures. The abundance of structural information in repositories such as the PDB compliments the even greater amount of structural information available for small-molecules typical to drug discovery. For example, state-of-the-art drug discovery pipelines commonly investigate millions of molecules against multiple targets during initial assays [4]. The availability of such large volumes of structural information is introducing fundamental re-thinking on how biological and chemical investigations can be designed. On the biological side for instance, it can be conjectured that technological advances leading to easier acquisition of structural information, are possible harbingers of novel paradigms where structure-based reasoning about the biochemical characteristics of a protein occurs in parallel with experimental studies in determining its function at cellular and biochemical levels [2]. Similarly, on the drug discovery side, availability of hitherto unavailable quantities of structural information, is spurring the growth of mechanistic and rational drug discovery.

Molecular informatics is an area that deals with the challenges of algorithmic analysis, and management of structural information and possibly non-structural data that may be directly associated with it. It thus spans biological and bio-chemical sciences along with the corresponding disciplines of Bioinformatics and Chemoinformatics. Examples of such investigations include:

- Bio-chemical and computational exploration of the molecular structural space consisting of known (synthesized or non-synthesized) structures.
- Development of computational structure-property models that relate variations in molecular structure to variations in molecular activity or properties
- Querying of molecular structural databases and management of associated information.

It should be noted in this context, that the classical approach to understanding molecules is based on modeling them using the Schrodinger wave equation:

$$E\psi = H\psi, \psi = \psi(t) \quad (1)$$

In this equation, E denotes the energy of the system, H is a self-adjoint operator in Hilbert space, called a Hamiltonian and Ψ encodes the probability of the outcome of all possible measurements made on the system. The main practical challenge in applying the Schrodinger equation to real-world molecules lies in its computational complexity. Researchers have therefore often resorted to simplifications by treating molecules as objects and representing them through characteristics such as shape, volume, connectivity, pose, and a variety of other attributes that are typically represented using vector space models. It is precisely the use of these alternate ways of modeling molecules, which opens up possibilities of applying techniques from pattern/image analysis and information retrieval to this area.

2. Key Problems

Problems where image analysis and information retrieval techniques can play a significant role include:

1. *Molecular representation*: Common ways of molecular representations include graph-based representations

(both as 2D and 3D graphs) and more complex 3D surface-based representations. These choices require tradeoffs between complexity and reasoning power and introduce specific technical challenges. For instance, for surface-based representations, the difficulty lies in defining an intrinsic coordinate system over the curved molecular surface that maps a point on the curved surface to a point on a standard coordinate system. This issue is closely related to the problem of modeling arbitrary curved objects, encountered in computer vision.

2. *Molecular similarity*: Determining similarity between molecules is essential for identifying molecules that have similar biochemical behavior. This requires, defining an appropriate similarity measure and designing algorithm to compute it. For example, given the graph-based representation of two molecules, their similarity can be computes through graph-matching. Alternatively, the occurrence or frequency of specific motifs in molecular graphs can be characterized using a vector-space model and a similarity measure can be computed on them. For surface-based representations, on the other hand, the similarity formulation is more complex. The molecular similarity problem provides an opportunity to apply and extend research from image and information retrieval to molecular informatics.
3. *Multi-modal nature of molecular properties*: Properties like geometry and charge distributions that may be used for describing molecules have different characteristics. For example, while the geometric representation of a molecule is unique, its field-based properties, defined on molecular surfaces are influenced by superposition-effects. Analyzing such properties, for example, determining their distribution on the molecular surfaces can be done through image analysis techniques.
4. *Molecular pose and conformations:* During analysis, the pose of the participating molecules can be arbitrary. Further, each molecule can alternate between one of its many energetically minimal configurations, called conformations, which are characterized by different bond-angles in a molecule. A molecule can thus be thought of as a "deformable object". Algorithms for molecular analysis need to be pose-invariant and capable of modeling conformers since the bio-chemical behavior of a molecule can vary significantly depending on its conformation.
5. *Efficiency*: It is typical to conduct computation on molecular involving large sets ranging from thousands to millions of molecules (as in drug discovery). Thus, it is imperative for algorithms to be highly computationally efficient.

We summarize below our investigations in successfully developing techniques based on concepts from image and pattern analysis/retrieval to the following two problems in molecular informatics:

- Query-retrieval of molecules using surface-based representations.
- Design of efficiently-computable similarity measures for comparing non-linear molecular properties

Towards the first goal, our research [5], extends 3D object representation techniques from computer vision to systematically map, arbitrary (convex or non-convex) molecular surfaces on a standard spherical coordinate system. Following this, a highly efficient retrieval technique based on histogram intersection from image retrieval is employed to directly match geometric and non-geometric surface attributes. The efficiency of the technique arises from the fact that it does not require explicit pose-optimization. For the second goal, our ongoing research focuses on designing Wavelet-Riemannian similarity metrics [7]. Such metrics (a) capture non-linear molecular features better than currently used measures, (b) are efficiently computable due to their wavelet nature, and (c) can be used for efficient query-retrieval due to their metric character. These techniques are brought together as part of FreeFlowDB [6], a publicly available information management system for drug discovery, which is currently being used for anti-malarial therapeutics development at the University of California, San Francisco [1].

3. REFERENCES

[1] P. Malik, T. Chan, J. Vandergriff, J. Weisman, J. DeRisi, and R. Singh, "Information Management and Interaction in High-Throughput Screening for Drug Discovery", *Database Modeling in Biology: Practices and Challenges* Z. Ma, and J. Chen, eds., Springer, 2006

[2] R. Najmanovich, W. Torrance, and J.M. Thornton, "Prediction of Protein function from structure: insights from methods for the detection of local structural similarities", BioTechniques, Vol. 38, No. 6, 2005, pp. 847 – 851

[3] A. Shimode, I. Yoon, M. Fuse, H. C. Beale, and R. Singh, "Automated Behavioral Phenotype Detection and Analysis Using Color-Based Motion Tracking", *Canadian Conference on Computer and Robot Vision,* pp. 370 – 377, 2005

[4] R. Singh "An Overview of Computational Knowledge Discovery and Pattern Analysis Problems in Contemporary Drug Discovery and Design", *DIMACS Summer School Tutorial on New Frontiers in Data Mining*, 2001

[5] R. Singh, "Reasoning about Molecular Similarity and Properties", Proc. IEEE Proc. IEEE Computational Systems Bioinformatics Conference (CSB), pp 266 – 277, 2004.

[6] R. Singh, E. Velasquez, P. Vijayant, and E. Yera, "FreeFlowDB: storage, Querying, and Interacting with Structure-Activity Information in High-Throughput Drug Discovery", IEEE International Conference on Computer-Based Medical Systems, pp. 75 -80, 2006

[7] E. Velasquez, E. Yera, and R. Singh, "Determining Molecular Similarity for Drug Discovery Using the Wavelet-Riemannian Metric", IEEE Symp. on Bioinformatics and Bioengineering, 2006 (To Appear)