Doctoral thesis summary

Maciej Dziubiński

"Development of selected mesoscopic physical models with the aid of machine learning methods and their applications in studies of molecular systems",

in Polish,

"Rozwijanie wybranych mezoskopowych modeli fizycznych z wykorzystaniem metod uczenia maszynowego i ich zastosowania w badaniach układów molekularnych"

Introduction and aims of the thesis

This dissertation is concerned with the development and application of unsupervised machine learning methods in the field of theoretical biophysics and bioinformatics. The machine learning approach offers a powerful framework for extracting and purifying valuable information from large, multi-dimensional sets of data generated in simulations and experiments of biomolecular systems. It is not, however, the case that ready-made machine learning methods offer infallible means of dealing with all sorts of complex, and partially chaotic data encountered in computational biophysics and structural biology. Large portion of this work is devoted to the adaptation of unsupervised machine learning techniques to our particular purposes.

In this dissertation, we employed unsupervised machine learning strategies dealing with two problems arising in theoretical biophysics and bioinformatics. The first problem was the identification of quasi-rigid structural parts in proteins, whereas the second one was devoted to discovery of internal cooperation of molecular subsystems that propels a conformational transition. Both problems involved dynamical properties of molecular systems, and the analyses presented in this dissertation allowed for a simplified description of these phenomena.

Both, simulation and experimentally acquired data become increasingly complex, in particular because:

- the systems under study are progressively larger,
- and their structural transitions comprise seemingly chaotic paths in multidimensional configurational spaces.

For both these reasons, we need to facilitate methods of extracting the underlying structure of the data, to gain perspective, and to augment our 3-dimensional perception in order to pose new hypothyses about structural and functional properties of the biomolecular world. The overall aim of this dissertation was the simplification of complex, dynamical properties of molecular systems. We applied unsupervised machine learning techniques to two such problems.

One problem arising in computational biophysics is the identification of dynamic domains, where the input consist of configurations produced in the course of an MD simulation, or an NMR experiment. Although highly mobile and flexible, many proteins have structurally static sub-regions. By expressing protein's conformations in terms of quasi-rigid parts that move with respect to one another, but remain internally rigid,

we were able to simplify the description of its dynamic nature. The second problem tackled in this dissertation is the identification of molecular subsystems that play active roles in structural transformations of the whole systems. The focus of this study was more technical, as the molecules analyzed therein were small. However, we were also aiming at discovering conceptual and methodological caveats that might hinder analyses of larger molecules such as proteins.

Dynamic domains

ResiCon was built upon several ideas, some of which were essential for its high-quality results. The first one made use of the concept of a contact between residues (that had been proven to be advantageous in an earlier application [20]) to define the geometrical variability. Thereupon, the second idea was to propose a specific spectral clustering algorithm, whose output were clusters corresponding to dynamic domains of a given protein structure. In its current implementation, ResiCon requires information about all pairs of configurations to calculate the geometrical variability.

The similarity measure derived from geometrical variability was intended to express not only structural shifts, but – more importantly – the strength of physical interactions between pairs of amino-acids. This similarity measure was indeed effective, as it lead to compact, high-quality dynamic domains, but as we discuss in Chapter 2, its interpretability proved to be especially useful. As we investigated peculiar artifacts in the results of the spectral clustering algorithm, we encountered discontinuities in the clusters. Because of the similarity measure's interpretability, these "artifacts" lead to valuable observations about the mechanics of structural transitions of the HIV-1 protease.

The spectral clustering method offered a criterion for selecting an optimal number of clusters. One should note that popular methods – GeoStaS and PiSQRD – were vulnerable to the slightest changes in parameter values, often yielding absurd clusterings, whereas ResiCon was able to discover relevant, rigid domains, or a lack of such. ResiCon is available at *http://dworkowa.imdik.pan.pl/EP/ResiCon*.

Molecular cogs

In Chapter 3 we presented a novel methodology for extracting knowledge on internal mechanics of small molecules undergoing structural transions. This was intended as a "proof of concept", with a stronger focus on potential capabilities of the proposed methodology and indicating technical problems and pitfalls. We have learned that the similarity measure used for identifying molecular cogs leads to reasonable results, and often highlights unexpected properties of the system at hand. But – perhaps more importantly – the similarity measure has a clear interpretation in terms of free energy change, associated with a structural transition.

Thus, we got a meaningful similarity measure, for which we could apply an adequate clustering method. For that purpose, we proposed a custom objective function, for which we found optimal solutions using a genetic clustering algorithm.

Summary

This work presented the difficulties and benefits of adapting unsupervised machine learning to analyzing biophysical data. The examples provided in Chapters 2 and 3 showed solutions to three main problems that are likely to be encountered in such investigations:

1. choice / definition of a similarity measure,
2. right clustering procedure and the optimal number of clusters,
3. verification of results.

Our similarity measures were strongly based on physical properties of the analyzed systems. This was crucial as it brought well defined meaning to the resulting partitionings. Next, we chose the appropriate clustering algorithms that worked well with the proposed similarity measures. However, we were not be able to judge the quality of the clustering (and thus, the clustering procedures) without an external method of verifying the results. Therefore, these three aspects of an analysis involving unsupervised machine learning are interrelated and cannot be considered separately.

Maciej Dziubiński