Chapter 44
Kernel Generative Topographic Mapping of Protein Sequences

Martha-Ivón Cárdenas
Universitat Politècnica de Catalunya, Spain

Iván Olier
The University of Manchester, UK

Alfredo Vellido
Universitat Politècnica de Catalunya, Spain

Xavier Rovira
Universitat Autònoma de Barcelona, Spain

Jesús Giraldo
Universitat Autònoma de Barcelona, Spain

ABSTRACT

The world of pharmacology is becoming increasingly dependent on the advances in the fields of genomics and proteomics. The –omics sciences bring about the challenge of how to deal with the large amounts of complex data they generate from an intelligent data analysis perspective. In this chapter, the authors focus on the analysis of a specific type of proteins, the G protein-coupled receptors, which are the target for over 15% of current drugs. They describe a kernel method of the manifold learning family for the analysis of protein amino acid symbolic sequences. This method sheds light on the structure of protein subfamilies, while providing an intuitive visualization of such structure.

INTRODUCTION

It has been just over 10 years since the publication of the first draft of the human genome decoding. The detailed description of the human genome is a milestone for science in general and for medicine in particular. It has opened the doors to new approaches to the investigation of pathologies that hold the promise of the advent of truly personalized medicine. Through these doors, though, a new challenge for intelligent data analysis has also entered.

Over the last decade, medicine has become a data-intensive area of research. One in which new data-acquisition technologies and a wider variety of investigative goals coalesce to make it one of the most important challenges for intelligent data analysis (Lisboa *et al*., 2004). The *-omic's* sciences have contributed the most to this data deluge, stemming from microarrays in genomics, protein chips and tissue arrays in proteomics, etc. As very explicitly reported in (Kahn, 2011): *[...]*

DOI: 10.4018/978-1-4666-3604-0.ch044

the need to process terabytes of information has become the rigueur for many labs engaged in genomic research.

Arguably, drug research has contributed more to the progress of medicine during the past century than any other scientific factor (Drews, 2000). One of the main areas of drug research is related to the analysis of proteins. The function of the proteins depends directly on their 3D structure, which is embodied in their amino acid sequence. Such 3D structure is difficult to unravel, though. Alternatively, protein sequences can be the direct object of our analysis, and they are easy to acquire. The analysis of the gene-family distribution of targets by drug substance reveals that more than 50% of drugs target only four key gene families, from which almost the 30% correspond to the G protein-coupled receptors (GPCRs) family. This family regulates the function of most cells in living organisms and is the focus of the work reported in this chapter. The grouping of GPCRs into types and subtypes based on sequence analysis may significantly contribute to helping drug design and to a better understanding of the molecular processes involved in receptor signaling both in normal and pathological conditions.

The challenge of managing the complexity of these types of data invites us to go one step further than traditional statistics and resort to intelligent pattern recognition approaches. In particular, statistical pattern recognition and machine learning methods bear the potential to both scale well to large databases and to deal with non-trivial types of data. Sound statistical principles are essential to trust the evidence base built with any computational analysis of medical data (Lisboa, 2002). Statistical machine learning methods are already establishing themselves in the more general field of bioinformatics (Baldi, 2001).

This work is specifically motivated by the need of defining a robust probabilistic method for grouping and visualizing symbolic protein sequences. As mentioned in (Schölkopf, Tsuda & Vert, 2004), there is no biologically-relevant manner of representing the symbolic sequences describing proteins using real-valued vectors. This does not preclude the possibility of assessing the similarity between such sequences. Kernel methods can be used to this purpose if understood as similarity measures.

In the following sections, we report our work on grouping and visualization of GPCR protein sequences using a kernel variant of a nonlinear model of the manifold learning family. A suitable kernel for this type of data is described. The visualization of the sequence data and the grouping results can be a useful tool in the quest for interpretability. The reported results reinforce the veracity of this statement.

FROM PROTEINS TO DRUGS

Introduction

As stated in (Overington, Al-Lazikani, & Hopkins, 2006), there is a paradox in the fact that an industry such as pharma that spends yearly more than US $50 billion on R+D, has not been able to generate enough knowledge about the set of molecular targets that are the object of its products. That is why drug target discovery has of late received much attention in different areas of biochemistry-related drug research.

Lately, drug target discovery has received much attention from different areas of biochemistry-related drug research contributing more to the progress of medicine than any other factor. This is the result of advances in chemistry, pharmacology, and the clinical sciences. Molecular biology and genomics are now at the forefront on drug research. This has been exponentially amplified by developments in information, communication, and computation technologies. Genomics, proteomics, and the bioinformatic tools that support them, can provide us with knowledge of suitable targets for medicines yet to be designed and, therefore, with a more proactive leverage on the process of drug design.