Automatic Spatial Pattern Extraction with Application to Mining Protein Structures and Neuron Morphologies

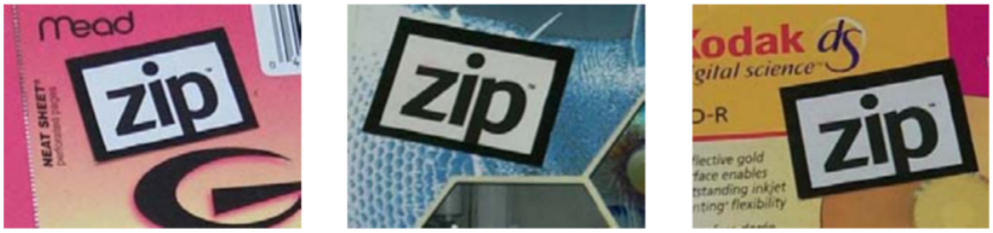
Wei Qian

One of the most amazing capabilities of human beings is to extract common spatial patterns from observations and use these patterns to make inference. For example, even if you do not know Woody and I tell you that below are two pictures of him, you will most likely be able to recognize him since he is the only person (i.e., spatial pattern in our research context) showing up in both pictures:

source: http://bit.ly/1Qu0RUV & http://bit.ly/1ppcfqX

My research journey started with asking: can we develop an intelligent computer system that is capable of doing so? More specifically, can we empower a computer to automatically learn spatial patterns (e.g. Woody in the above example) from a set of samples (e.g., the two pictures above)?

Automatic discovery and detection of spatial patterns is one of the most challenging topics in the modern Artificial Intelligence research. It has broad applications to a variety of fields, such as biology, economics, finance, humanity, psychology, and so on. A year ago, I started pursuing this research topic under the supervision of Professor Pengyu Hong of Computer Science. In 2004, Professor Hong published a paper where he described a mathematical model for representing spatial patterns and a machine learning technique for learning such a spatial pattern model from a set of examples. The learned mathematical model can be applied to detect if a new sample contains the same pattern. For example, his approach was capable of capture the "ZIP" logo which undergoes various transformations and is embedded in diverse backgrounds as shown in the following images:



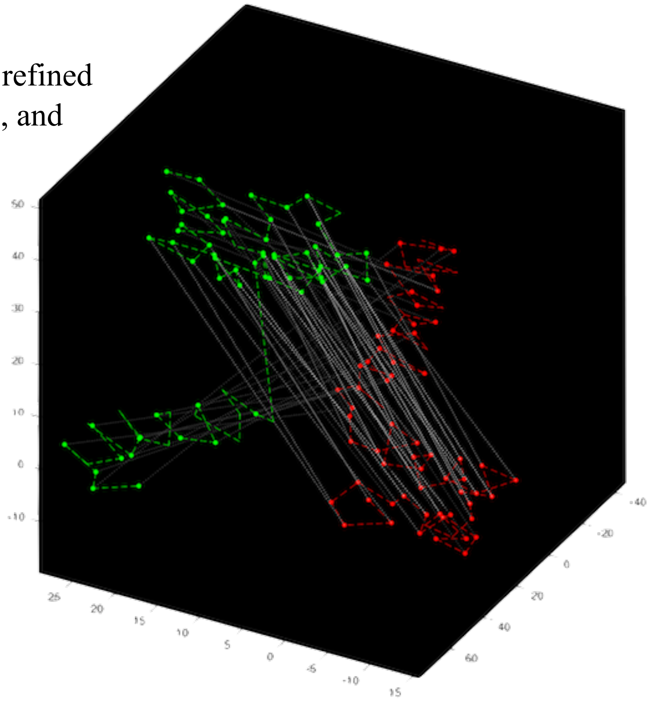
This work provides a solid foundation for me to ask: can I extend this approach to other application domains besides image processing? As a computer science and neuroscience student, I am always fascinated by interdisciplinary research where I can apply my computer science knowledge and skills to tackle challenges in biology. Therefore, I decided to investigate the spatial patterns in proteins and neurons.

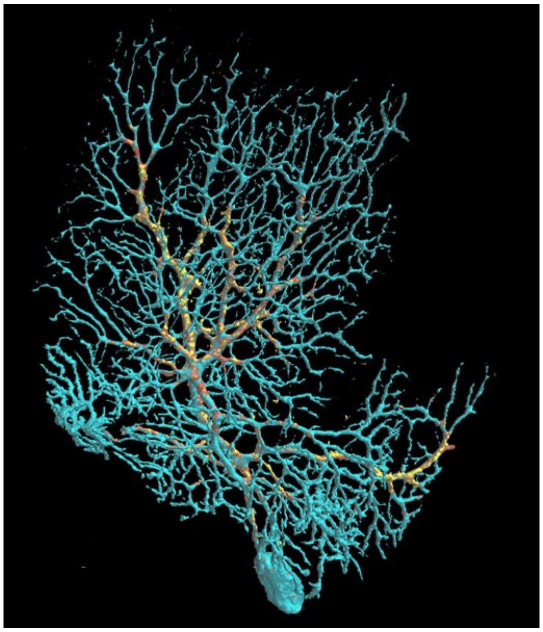
Proteins are macromolecules responsible for nearly every task of cellular life. They are 3D structures consisting of amino acid sequences translated from genes and interact with each other to carry out essential functions, such as catalyzing metabolic reaction, transport molecules, respond to stimuli, and so on. Such interactions often happen at sites (e.g. protein domains) that are conserved between proteins across species. By recombining and rearranging these domains as proteins' basic building block, molecular evolution is able to create proteins with different functions.

Conventional computational methods for studying protein domains between proteins mainly use the sequences and secondary structures of proteins to do pattern matching, which ignore the 3D nature of proteins and cannot take advantage of the available 3D structures of proteins. . To achieve better understandings about protein interactions and their functions, we need to investigate the detailed spatial characteristics of the 3D structures of proteins.

Under this realization, many researchers have started manually matching the 3D structures of proteins. However, this is very time consuming and can be subjective. Therefore, I propose to improve Professor Hong's spatial pattern discovery approach and apply it to investigate the patterns in 3D protein structures. This research can not only lead to better 3D definitions of protein domains, but also may discover novel protein domains missed by conventional computational methods.

In the past year, I have implemented and refined the original model developed by Professor Hong, and recently started applying it to study 3D protein structures. We have obtained quite promising results. An example is shown in the figure to the right, in which two proteins are colored in green (*4Q59 protein*) and red (*4D1E protein*), respectively and each solid dot represents an amino acid. Our method is able to automatically match the domain (*CH1 domain*) shared by them even though their 3D poses are not aligned. The white lines indicate the detailed matches between amino acids in these two proteins. The brighter a line, the stronger a match.





source: http://unc.live/1TbjsHw

Besides proteins, the spatial patterns embedded in 3D neuronal structures are also intriguing. The importance of neuronal morphology in brain function has been recognized for at least a century. For instance, the Purkinje cell structure showing on the left has many branches and such structure is highly related to its function in movement as it collects information from many sense in order to calibrate the movement control. Recent advances in experimental techniques have allowed neuroscience researchers to produce a large number of 3D neuronal structures like this. Our method will be a great tool to analyze those spatial data in depth. We expect our results to shed new lights on the structure-function relation, which is central to many questions in neuroscience.

However, this is a challenge task and there are still several technical obstacles in front of us, this fellowship will allow us to overcome some of them and bring this project to the next level. Since this research requires extensive computation, I plan to use most of the stipend to get a faster computer to run experiment entirely and the rest of the stipend will be used to buy related books, and cover costs to attend seminars held round the Boston area.

In terms of the format, we plan to continue our weekly research meeting during which we will discuss my progress, new discovery, confusion and technical obstacles. In the end of this project, we aim to publish at least one paper out of it.

[redacted]

For more information about me, please visit: *http://www.wesleyq.me/*