Understanding RNA

Background
Ribonucleic acid (RNA) is a nucleic acid similar to DNA and essential for life. Some scientists believe that the earliest forms of life may have relied on RNA rather than DNA to store genetic information. We have a good understanding of DNA and its relation to biology, and considerable knowledge about the mechanisms governing protein functions. We can use the same tools to discover the base sequence of an RNA molecule and its 3D structure. In other words, we can get detailed information about the composition, geometry and topology of RNA, but we do not understand the function of most of the RNA molecules we find in the real world. Solving this problem could lead to a massive step forward in understanding biological systems.

Project
The problem of categorising RNA has a number of elements. Firstly, we must be able to measure the similarity between two molecules as this is at the heart of all pattern recognition and machine learning techniques. The initial goal of the PhD is to develop a new method of measuring this similarity using information from the sequence, geometry and topology. We can then discover which of these factors is important in determining the type and function of the RNA.

The second goal is to apply machine learning to the problem of classifying or categorising a particular RNA molecule. This work is in collaboration with the University of the Balearics, who have compiled a hierarchical classification of 419 known RNA structures. We will use this data to learn about the categories of the thousands of unclassified RNA molecules. The final task will be to develop an interactive classification tool for biologists which can suggest likely classifications of an unknown RNA as well as the most similar analogues from a large database of RNA molecules.

In summary, the aims of the PhD are
- Devise a new method of measuring the similarity of RNA molecules using information gathered from their sequence, geometry and topology.
- Use this measure to determine the relative importance of the three sources of information.
- Exploit machine learning and a database of known classifications to categorise unknown RNA molecules.
- Develop an interactive tool to support classification and exploration by biologists.