Evaluation of the thesis “Predicting function of genes and proteins from sequence, structure and expression data” by Torgeir R. Hvidsten

Prof. Anna Tramontano
University of Rome “La Sapienza”

The thesis by Torgeir R. Hvidsten describes several computational biology and bioinformatics methods aimed at understanding the sequence – structure – function relationship in biological molecules.

Functional assignment is a central problem in the life sciences and this thesis is a good example of how computational approaches can lead to a better understanding of biological problems as well as to useful practical results.

Function can be defined at several levels, for example by describing the biological process to which a given biological entity takes part or by describing the specific molecular activity of the molecule. This thesis explores both these aspects, by first describing novel methods for analyzing the temporal expression patterns of genes and then by devising a method to assign a structural class to the encoded proteins.

I found the thesis very well balanced and extremely clearly written. The Introduction to molecular biology is simply excellent. It is apparent that the candidate made an enormous effort to simplify very intricate biological concepts as to make them understandable to computer scientists and he was able to do so without loosing in accuracy. At the same time he described the algorithms and statistical approaches in a language accessible to biological scientists, again not at the expenses of accuracy. This part of the thesis really impressed me and I believe it demonstrates that the new generation of scientists are achieving one of the most important goals of the field: the ability to effectively establish a communication link between the computational and biological world which is indispensable if we want to exploit the power of computation in making sense of the complexity of biology.

The methods described in the thesis are all at the forefront of computational biology and of bioinformatics.
The first two papers address the issue of how to relate the temporal expression pattern of genes with their biological function and it does so by using a rule learning approach based on the rough set theory, an approach pioneering by the laboratory where the thesis work was performed. The innovation consists in using rule induction obtained by analyzing the measured biological effects in terms of combinations of causes. The results are excellent and competently evaluated. A further advantage of the selected approach is that the generated rules can be used to interpret the observed effect in biological terms.

The similar temporal expression pattern of different genes is the result of common control mechanisms that have exquisite specificity and great robustness. The candidate analyzed this aspect of the problem in paper III, where he set to find signatures for these common regulatory mechanisms, taking into account the fact that gene expression regulation must be combinatorial in nature to achieve specificity and robustness.

In the next two papers, the candidate looked at the relationship between the protein sequences and their structures, another central problem in computational biology. Should we be able to correctly infer the three-dimensional structure of all the biological entities, we could complete our understanding of the genomic data, correlate each of the molecules with a specific molecular function, understand how they interact with each other, etc.

The problem is very complex due to the physics of the process by which a protein sequence achieves its three-dimensional structure. One relatively recent observation is that the menu of possible three-dimensional arrangements of proteins seems to be limited and therefore it might be possible to assign a given protein to classes describing its architecture. Once again this is a problem of extreme interest and it is attracting a concentration of efforts from a large community of scientists. The papers described here are based on a novel, and very interesting, idea of combining sequence and structure conservation to derive signatures for each of the known protein architectures. The results are extremely interesting and open the road to new possibilities in the field.

All the efforts in protein structure prediction are based on the assumption that the knowledge of the structure of a protein will help in elucidating its function. The generality of this assumption is still to be proven, and is raising enormous interest also because of the establishment of community wide efforts to determine the structures of as many proteins as possible (structural genomics). The last paper approaches this problem in a very clever way, by relating the presence of structural motifs, identified with the approaches described in papers IV and V, to molecular function.

Altogether, this thesis addressed several related problems of extreme importance for biology using very interesting computational approaches.

Two final remarks need to be made. The first is that the thesis is very well written and logically organized, notwithstanding the inherent difficulty of relating different approaches and different experimental data. The second, even more important, is the fact that all the results were accurately evaluated and tested, and in genomic research, where different large data set coming from different experimental laboratories have to be integrated this is of paramount importance.

Altogether the thesis properly describes excellent, innovative and creative research and I have no doubt that the candidate should be invited to present his dissertation.

Anna Tramontano