Binding site recognition and disorder prediction in protein function annotation

D.T. Jones*, L.J. Ward and J.S. Sodhi

University College London

The completion of the first draft of the human genome in 2001 has been heralded as a major breakthrough that will finally enable researchers throughout the world to answer intriguing and elusive questions relating to the mechanism that govern complex biological processes. The hope is that the information generated from the numerous genome projects worldwide will be harnessed to further our understanding and applied to beneficial and therapeutic use through computer aided biological research.

A crucial area of research which will allow these aims to be reached is the prediction of the biological function of sequenced genes. Relationships between protein structure and function have been well documented over the last 30 years, however the diversity and complexity presented by nature poses several challenging problems. Gene products from different species may exhibit the matching biological functions, but may show little or no sequence similarity, perhaps due to convergent evolution. It may be that although there is little overall structural and sequence similarity key active sites are conserved allowing similar functions to be carried out. The well-studied example of the serine proteinases illustrates this they have been shown to conserve a specific catalytic triad atomic arrangement across a range of fold and sequence families. Further, analyses of structural superfamilies have shown that members of the same family can function quite differently.

Analyses of functional regions within proteins will not only allow the development of more reliable genome annotation tools but also enhance the knowledge base of the biological role of proteins at a cellular level. Such understanding will be a key stepping stone in the development of techniques and pharmaceuticals to target diseased genes and their products as well as proteins from pathological organisms.

Structural classification of proteins has proved to provide extremely valuable information complementing sequence data and providing additional insights into the structure function relationship of proteins. Fold classification methods such as SCOP, CATH and FSSP reveal are examples of strategies where the general goal is to classify proteins according to their global fold. Such classification schemes have highlighted the fact that there are only a finite number of different protein folds. Further protein with similar architectures may have little or no sequence similarity and function quite differently. Thus although the structural scaffold of proteins may reveal a great deal of information, ranging from evolutionary to functional aspects the exact details regarding the specificity of biological function may be lacking. Indeed the precise mechanism of function may only be determined by an in depth analysis of key functional sites. Examples of such regions include enzyme active sites, metal binding sites, ligand binding clefts and indeed interacting regions between two proteins, protein-small molecule or protein DNA, albeit over a much larger area.

In this context, I will describe an extensible method we have developed for applying neural networks to the problem of functional site recognition in protein structures (Sodhi *et al*., 2004). Results of applying this method to metal binding sites and DNA binding proteins will be discussed. The eventual goal of this project will be to build a pipeline for automatically annotating possible functional sites in protein structures (either theoretical models or experimentally

determined structures from structural genomics projects).

The second part of my talk will look at the prediction of protein disorder from amino acid sequence and how this might influence our ability to predict gene function (Jones & Ward, 2003; Ward *et al*., 2004; Lise & Jones, 2005). It has been one of the central tenets of structural biology that the function of a protein is determined by its three-dimensional structure. However, a large proportion of protein sequences now appear not to code for well-defined structures at all, and may adopt a non-globular structure or even be entirely unfolded (disordered) in solution. The abundance of these "disordered" regions suggests that they are, to some extent, evolutionarily conserved and therefore likely to possess biological function. An analysis of the disordered regions predicted in the yeast genome provides an insight into the biological roles of disordered proteins and suggests that disorder may be an important missing ingredient in predicting protein function from sequence.

References

Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L. & Jones, D. T. (2004). Predicting metal-binding site residues in low-resolution structural models. *J. Mol. Biol.*, **342**, 307-320.

Jones, D.T., & Ward, J.J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, **S6**, 573-578.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635-645.

Lise, S., & Jones, D. T. (2005). Sequence patterns associated with disordered regions in proteins. *Proteins*, **58**, 144-150.