Empirical methods in protein modelling

Michael J E Sternberg*[1,2], David Balding[3], Phillip Carter[1], Keiran Fleming [1], Herny Gabb [2], Suhail Islam [1,2], Richard Jackson[2], Lawrence Kelley [1,2], Lianne Mayor[3] Robert MacCallum [2], Arne Muller [1,2], Florencio Pazos [1], Graham Smith [1,2].

1 Centre for Bioinformatics, Imperial College London
2 Biomolecular Modelling Laboratory, Cancer Research UK
3 Department of Epidemiology and Public Health, Imperial College London

Today the genome sequencing projects have derived the sequences for more than 1 million proteins. In addition, experimental methods have revealed atomic coordinates for more than 20,000 proteins Understanding and exploiting these data is now central to progress in biology and this requirement has stimulated the development and expansion of bioinformatics. Of particular importance is that bioinformatics is essential to formulate *in silico* hypotheses for experimental study. A powerful and widely-used strategy to interpret the data is the use of empirical methods derived from careful analysis and treatment of the rich but highly complex data sources. This talk will describe several protein modelling studies that to a varying degree employ empirical methods.

A central concept in bioinformatics is that there are families of proteins that have evolved by divergence from a common ancestor (known as homologues). Homologous proteins generally share a similar three-dimensional structure. In addition, homologues, particularly if they are closely related ($> 30\%$ identity of the individual residues), have similar function. But distant homologues can have quite different functions.

The first topic to be described is the development of an automated pipeline to annotate protein sequences in the genomes (Muller *et al.*, 2002). Methods include the recognition of homology via statistical profiles (implemented in the algorithm PSIBLAST) and scanning against hidden Markov models (PFAM). Analysis of the occurrences of protein families in the different genomes reveals insight into the evolution and specialisation of different species. The talk will describe a recent statistical analysis to search for clusters of homologous proteins close within the human genome (Mayor *at al.* 2004). It was well know for a a few protein families that different homologues cluster within the genome. Our analysis shows that there is a wide variation in the extent of clustering amongst the most common families. This may have resulted from different evolutionary pressures on different functional classes of molecules.

Sequence-based methods are unable to detect all homologues and there is considerable interest in developing methods that exploit structural information to detect remote homologues. The approach is called fold recognition or threading. We will describe an approach (3D-PSSM) we have developed using profiles to detect remote homologues which is available via a web server (Kelly *et al.*, 2000; Bates *et al.*, 2001). Fold recognition algorithms are often evaluated via international blind trials (CASP) and the developments of these trials will be reported. Of note that in the last trial (2002) meta-servers that pooled results from individual algorithms were remarkably successful (Kinch *et al.*, 2003).

Next the talk will outline the current status of the package 3D-DOCK that aims to predict the 3D structure of a protein-protein complex starting from the coordinates of the unbound components. The first step (FTDOCK) is a global search for complementary shape with favourable electrostatic interactions using the Fourier correlation approach (Gabb *et al.*, 1997). This yields

a list of possible complexes and generally within this list there will be at least one complex that is close to the true structure. The subsequent steps aim to reduce the number of complexes in this list that one would need to consider to have a good chance of including g a good prediction. First, empirical residue-residue pair potentials are used to reduce the list (RPSCORE) (Moont *et al.*, 1999). Then a rigid-body refinement coupled with optimisation of side-chain / side-chain packing is performed (MULTIDOCK) (Jackson *et al.*, 1998). The talk will describe the status of the above approach on a test data set. Finally the results of the recent blind trial of protein-protein docking (CAPRI, www.capri.ebi.ac.uk) will be reported (Smith & Sternberg, 2003); Mendez *et al.*, 2003).

References

Muller, A., MacCallum, R. M. & Sternberg, M. J. E. (2002). Structural characterization of the human proteome. *Genome research*, **12**, 1624-1641.

Mayor, L. R., Fleming, K. P., Muller, A., Balding, D. J. & Sternberg, M. J. E. (2004). Clustering of protein domains in the human genome. *J.Mol. Biol.*, in the press.

Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 501-522.

Bates, P. A., Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. (2001). Enhancement of protein modelling by human intervention in applying the automatic prgrams 3D-JIGSAW and 3D-PSSM. *Proteins*, **45**, Supplement 5, 39-46.

Kinch, L. N., Wrabl, J. O., Krishna, S. S., Majumdar, I., Sadreyev, R. I., Qi, Y., Pei, J., Cheng, H. & Grishin, N. V. (2003). CASP5 assessment of fold recognition target predictions. *Proteins* **53**, Supplement 6, 395-409.

Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, **272**, 106-120.

Moont, G., Gabb, H. A. & Sternberg, M. J. E. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**, 364-373.

Jackson, R. M., Gabb, H. A. & Sternberg, M. J. E. (1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.*, **276**, 265-285.

Smith, G. R. & Sternberg, M. J. (2003). Evaluation of the 3D-Dock protein docking suite in rounds 1 and 2 of the CAPRI blind trial. *Proteins* **52**, 75-9.

Mendez, R., Leplae, R., De Maria, L. & Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*, **52**, 51-67.