A knowledge based model for protein-DNA interactions: a structural approach

Richard Gamblin* and Richard M. Jackson

University of Leeds

Recent studies into protein-DNA complex structures have indicated the presence of trends in amino acid-DNA base interactions (Luscombe *et al.*, 2001). Indeed, using a statistical analysis of pair-wise contacts, it has been shown that empirical interaction potentials can be derived and used, in conjunction with a threading procedure, to identify DNA binding sites (Kono and Sarai, 1999). These preliminary results suggest great promise for DNA recognition from structural data without the need for extensive experimental characterisation, for example by by gel retardation assay.

We have developed a method with the aim of quantifying structural features that confer specific binding properties not evident from sequence similarity alone. Using hydrogen bonding and non-bonded contact patterns from a non-redundant set of protein and DNA complex structures, an overall statistical knowledge based model was developed to represent specific amino acid-DNA base/ backbone interactions. This overall model was then applied with interaction data from specific protein-DNA complexes to create a diverse set of new models, termed structurally derived matrices (SDMs).

Experimental models of protein-DNA interactions based on binding assay data generally reflect a simplistic view of DNA recognition whereby proteins identify a specific sequence of nucleotides. These sequence based models use an alignment of DNA binding sites to generate position specific scoring matrices (PSSMs), a numerical representation of the nucleotide frequency at each position in the binding site. This model can then be used to characterise and identify putative binding sites in DNA sequences (Stormo, 1988).

The functional capacity of each of our SDMs was assessed by searching it against a set of nucleotide sequences in which the recognition sites had been pre-determined experimentally. As a control, an equivalent PSSM, representing the same binding site was also used to search each sequence. This assessment revealed that the binding site predictions made by the SDMs were significantly poorer than the equivalent PSSM predictions. The SDMs correctly predicted binding sites as the top 'hit' in only 2% of cases, compared with 58% of cases by the equivalent PSSM.

Our findings suggest that, while there is clearly some information to be obtained from analysis of these intermolecular interactions, application at the amino acid-DNA base level to a matrix-type model is much worse than PSSM models currently available.

References

Kono, H. and Sarai, A. (1999). *PROTEINS: Structure, Function and Genetics*. **35**, 114-131.

Luscombe, N. M., Laskowski, R. A., and Thornton, J. M. (2001). *Nucleic Acids Research*. **29**, 2860-2874.

Stormo, G. D. (1988). *Annual Review of Biophysics and Biophysical Chemistry*. **17**, 241-63.