

# Structural modelling of protein-DNA interactions

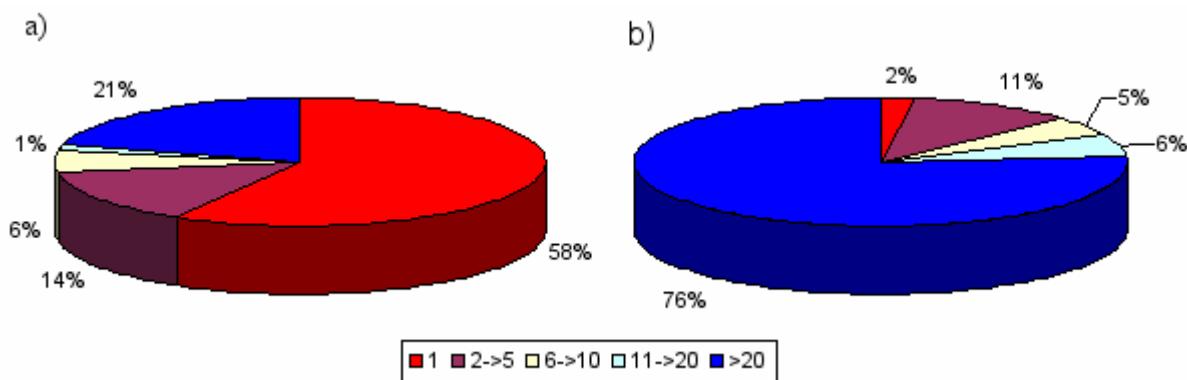
Richard Gamblin and Richard Jackson

A large number of *in silico* models of protein-DNA interactions reflect a simplistic view of DNA recognition, whereby proteins recognise a specific sequence of nucleotides. These models use aligned binding sites to generate a position specific scoring matrix (PSSM), in what is essentially a table giving the prevalence of nucleotides at each position in the section of bound DNA. This can then be used to characterise and identify other such binding sites in DNA sequences.

This representation of recognition is, however, far simpler than present *in vivo*. DNA is a negatively charged bi-polymer projecting into three dimensional space rather than a one dimensional series of letters, and it is this that proteins must negotiate to distinguish specific regions. The difference between the PSSM models of transcription factor binding and the *in vivo* reality provides an explanation as to why PSSM binding site predictions are plagued with erroneous predictions. Recent studies into protein-DNA complex structures have indicated the presence of trends in amino acid-DNA base interactions, and preliminary results from the latest models of protein-DNA recognition, which utilise this type of information, appear promising.

We have developed a novel method with the aim of quantifying structural features that confer specific binding properties not evident from sequence similarity alone. Using a catalogue of 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 was applied to create new PSSM-type models, termed structurally derived matrices (SDMs).

Assessment of the functional differences between the SDM and PSSM models revealed that SDM predictions 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 for a diverse set of experimentally characterised binding sites (see Fig. 1).

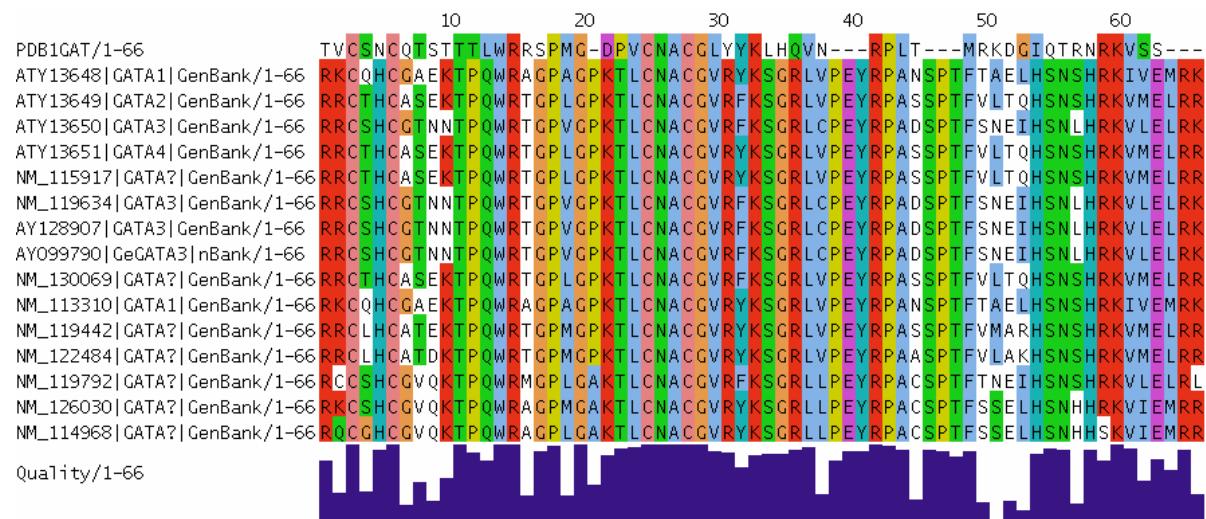


**Fig. 1. Binding site predictions by a) the PSSM model and b) the SDM model.** Correctly predicted binding sites as the top ‘hit’ appear in red, correctly predicted binding sites in the top 2 to 5 ‘hits’ appear in maroon, those in the top 6 to 10 appear in yellow, those in the top 11 to 10 appear in light blue and finally binding sites predicted outside the top 20 ‘hits’ appear in blue.

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. Indeed, although PSSM models have their limitations, they do perform very well on short sections of DNA sequence when representing a well defined binding site.

Recently we shifted our broad spectrum analysis and focussed our attention on the GATA family of transcription factors in *Arabidopsis thaliana*. The members of this class IV zinc finger protein are typified by the GATA motif that they selectively bind. PSSMs representative of this DNA recognition site are available for mammalian systems, however consideration of the length of sequence to be searched in *A. thaliana*, coupled with the abbreviated nature of the motif, mean that binding site predictions made with the PSSM model could never achieve statistical significance.

Consequently an alternative approach was taken, which involved investigation of the DNA binding domain of the *A. thaliana* GATA factors, using multiple sequence alignment and homology modelling techniques. Our findings suggested that these proteins interact with their cognate DNA in fundamentally the same way at the molecular level.



**Fig. 2. Multiple sequence alignment of *Arabidopsis thaliana* sequences aligned against a mammalian template.** The domain selected from the mammalian sequence is the zinc finger DNA binding domain as found in the 1gat PDB structure.

Our most recent efforts have involved developing a molecular mechanics model which will allow us to study the energy terms present in protein-DNA interactions. Using a modified version of the MultiDock interface refinement tool, we are currently investigating the components of these interactions to identify key features that are important to binding and recognition.

## Collaborators

Professor P. M. Gilmartin, Centre for Plant Science, University of Leeds, UK.

## Funding

This work was supported by the BBSRC.