

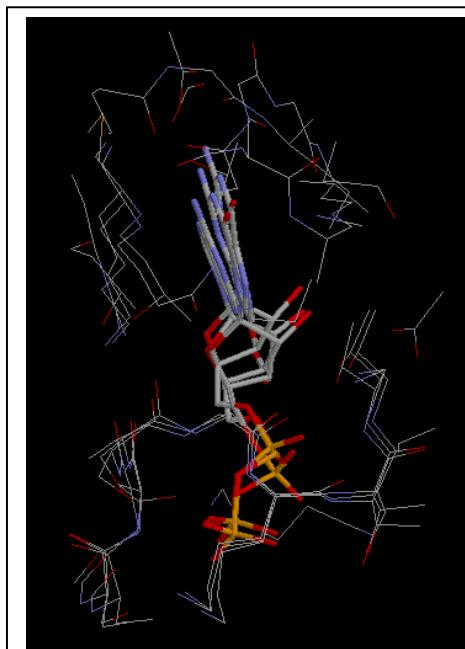
## Searchable database containing comparisons of ligand binding sites at the molecular level for the discovery of similarities in protein function

Nicola Gold and Richard Jackson

Structural genomics projects produce large amounts of data of which some are solved structures of hypothetical proteins of unknown function. The aim of this project is to aid the characterisation of these proteins by structure based prediction of protein function based on common modes of molecular recognition.

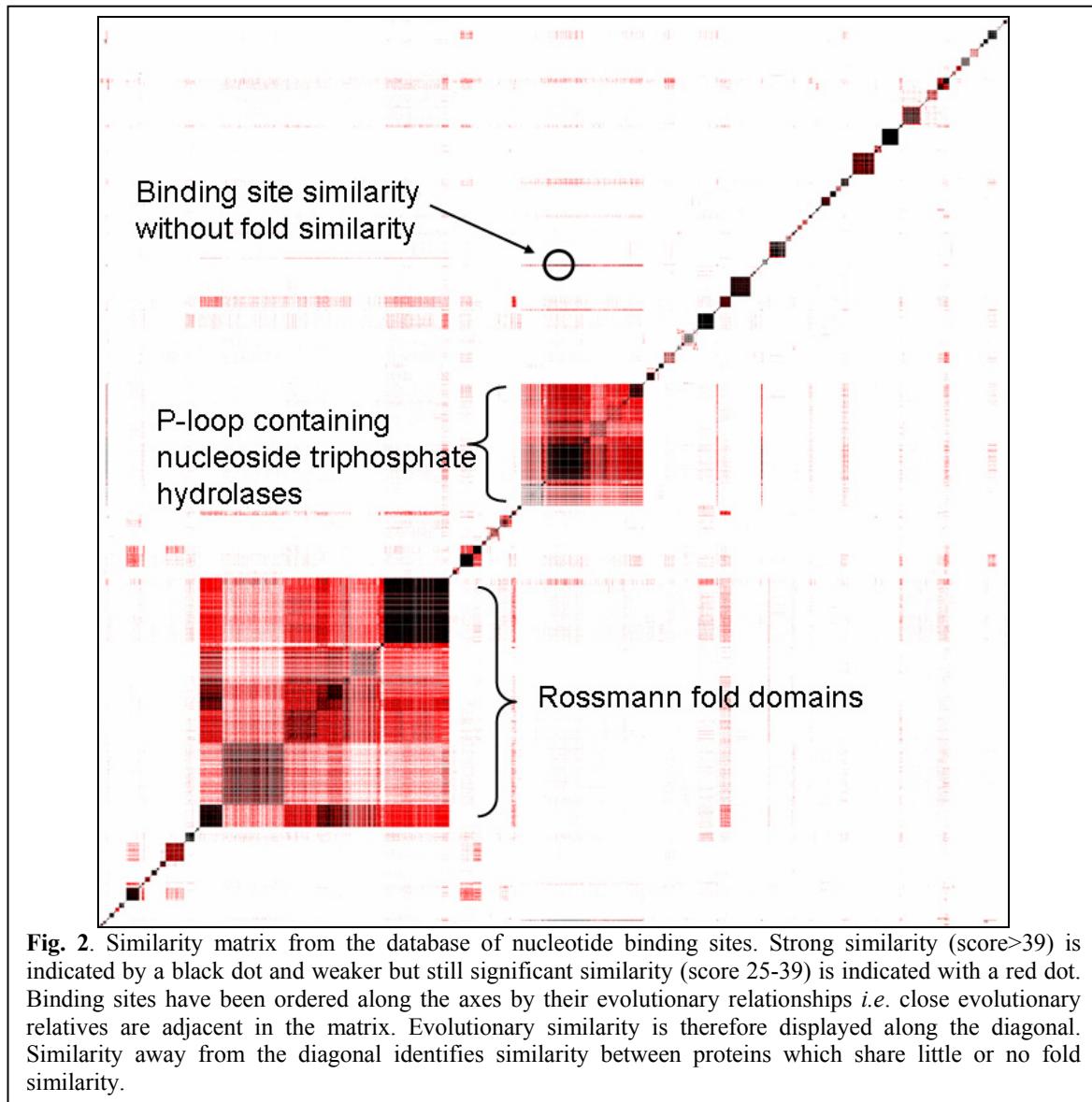
The current project extends previous work which demonstrated a method based on geometric hashing to compare the structures and properties of ligand binding sites and assess the extent of their similarity. In particular the binding site of the phosphate moiety of the large class of nucleotide ligands (ATP/ADP, GTP/GDP, FAD, NAD) was studied and is now extended to include the entire ligand binding sites of these proteins. The current project uses geometric hashing (described previously) to give a similarity score, a superposition, RMSD and equivalenced atoms for each pair of compared binding sites. These data are stored in a World Wide Web accessible database which is searchable with a PDB code and ligand information (such as ligand name, number and chain). Submission of these data rapidly returns a ranked list of similar ligand binding sites with the most similar at the top. Each hit is coloured according to its similarity to the query's overall fold and SCOP family. Interesting hits can then be selected and superimposed on the query allowing further examination and visualisation with molecular graphics packages (Fig. 1). A multiple alignment of structurally equivalenced atoms is also provided.

A frequency distribution of scores in the database gives two score thresholds which are used in a similarity matrix (Fig. 2). Here, strong similarity (score>39) is indicated by a black dot and weaker, but still significant similarity (score 25-39), is indicated with a red dot. Binding sites have been ordered along the axes by their evolutionary relationships *i.e.* close evolutionary relatives are adjacent; therefore similarity between close family members is displayed along the diagonal. This representation allows us to detect binding site similarity in the absence of sequence or fold similarity (off the diagonal). Similarity can, for example, be found between the binding sites of Elongation factor Tu and phosphoenolpyruvate carboxykinase (Fig. 1) which have different overall folds but share a common structural P-loop.



**Fig. 1.** Superposition of the GDP binding site of Elongation factor Tu protein (1dg1), the GTP binding site of signal sequence recognition protein Ffh (1ng1) and the ADP binding site of phosphoenolpyruvate carboxykinase (1k3c). 1dg1 and 1ng1 are from different SCOP families sharing the same superfamily and fold, whereas 1k3c is classified in a different superfamily and fold group.

Future work will see this method and database extended to include ligand binding sites from other proteins. This database and comparison method can then be used to discover new similarities indicating potential functional relationships between proteins and may uncover binding site similarities in proteins previously thought to be unrelated.



### Publication

Brakoulias, A. and Jackson, R.M. (2004) Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins; Structure function and bioinformatics*, **56**, 250-260

### Funding

We wish to acknowledge the support of the BBSRC.