

A new probabilistic model for binding site similarity analysis: applications in understanding ligand cross-reactivity and the functional classification of the protein kinase family

Sarah Kinnings¹, John R. Davies^{2*}, Kanti V. Mardia²
Charles C. Taylor² & Richard M. Jackson¹

¹Institute of Molecular and Cellular Biology, University of Leeds

²Department of Statistics, University of Leeds

The large-scale comparison of protein-ligand binding sites is problematic, in that measures of structural similarity are difficult to quantify and are not easily understood in terms of statistical similarity that can ultimately be related to structure and function. We present a binding site matching score the Poisson Index (PI) based upon a well-defined statistical model (Davies *et al.*, 2007). PI requires only the number of matching atoms between two sites and the size of the two sites – the same information used by the Tanimoto Index (TI), a comparable and widely used measure for molecular similarity. We apply PI and TI to a previously automatically extracted set of binding sites to determine the robustness and usefulness of both scores. We found that PI outperforms TI; moreover, site similarity is poorly defined for TI at values around the 99.5% confidence level for which PI is well defined. A difference map at this confidence level shows that PI gives much more meaningful information than TI. We show individual examples where TI fails to distinguish either a false or a true site pairing in contrast to PI, which performs much better. TI cannot handle large or small sites very well, or the comparison of large and small sites, in contrast to PI that is shown to be much more robust. Despite the difficulty of determining a biological 'ground truth' for binding site similarity we conclude that PI is a suitable measure of binding site similarity and could form the basis for a binding site classification scheme comparable to existing protein domain classification schema. Methods for analysing complete gene families in the drug discovery process are becoming of increasing importance, because similarities and differences within a family are often the key to understanding functional differences that can be exploited in drug design. Constituting around 1.7% of the human genome, the protein kinase family is one of the largest enzyme families and plays key roles in almost all signalling pathways. Since the deregulation of these signalling pathways often leads to disease, the control of protein kinase activity is a principle focus of pharmaceutical research. The vast majority of kinase inhibitors target the ATP-binding site. However, the high degree of sequence and structural conservation amongst the protein kinases means that the design of potent, selective kinase inhibitors is a significant challenge. We have developed a large online database for the retrieval of ligand binding site similarities (Gold and Jackson, 2006). These are extracted automatically from the Macromolecular Structure Database using a geometric hashing algorithm (Gold *et al.*, 2007). We have undertaken a large-scale comparison of protein kinase ATP-binding sites. This has allowed us to discover binding site similarity in different sub-families of protein kinase that are not evident from sequence similarity alone. It has also enabled us to quantify the effect of how different drug molecules bind to the same binding site and influence the local binding site conformation. We propose a relevant classification of the protein kinase family based on the similarity of their binding sites. Not only does this classification highlight features that are important for the potency and selectivity of kinase inhibitors, but it also allows us to predict possible cross-reactivity among the protein kinases.

References

- Brakoulias A., Jackson RM (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
- Davies J.R, Jackson R.M, Mardia K.V, and Taylor C.C. (2007). The Poisson Index: a new probabilistic model for protein ligand binding site similarity. *Bioinformatics*, **23**: 3001-3008
- Gold, N.D., Jackson, R.M. (2006). SitesBase: A database for structure-based protein-ligand binding site comparisons, *Nucleic Acids Res.*, **34**: D231-234
- Gold N.D., Deville K., Jackson, R.M. (2007). New opportunities for protease ligand-binding site comparisons using SitesBase. *Biochem Soc. Trans.*, **35**: 561-565