**On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites**

Murad Nayal, Barry Honig, Proteins: Structure, Function, and Bioinformatics, Published online 13 Feb 2006

**Summary by: Ankur Dhanik**

Many strategies used in the drug discovery process, such as virtual screening or de novo design, call for the identification of the drug-binding sites based on structure. One possible application of particular interest involves the targeting of a drug to a protein-protein interface. Identifying potentially 'druggable' sites in such regions would be of great interest. Another application that is likely to be of increasing importance is the identification of the best target in a particular pathway that one wishes to perturb. This paper introduces a new method for the identification and the accurate characterization of protein surface cavities. The method is encoded in the program SCREEN (Surface Cavity Recognition and Evaluation) . The cavity detection methods proposed so far has inherent limits. The results obtained using grid-based cavity finding methods are sensitive to grid spacing, as well as position and orientation of the protein in the grid. Cavity detection methods that depend on fitting spheres in the cavity space are not well suited for the detection of wide cavities. A large cavity requires large spheres to span it, which are then likely to spill over ridges and merge with spheres belonging to other cavities. Methods that use a simplified protein envelope, such as an ellipsoid or convex hull, are susceptible to the presence of protein extensions or arms, which will stretch the protein away from the molecular surface. SCREEN defines surface cavities geometrically in terms of the empty space between the protein's molecular surface and an envelope surface constructed by rolling an intermediate size spherical probe (of dimensions akin to that of a a typical small ligand). Properties of interaction sites are utilized as predictors of drug binding sites. Previous studies fall short of extensive examination. Only a few properties are examined in any one study and this makes it difficult to establish the role that individual properties play in binding. In this study, 408 attributes were computed for each cavity. The properties include various measures of size(6), electrostatics (94), hydrogen bonding (34), hydrophobicity and polarity (42), amino acid composition (21), rigidity (26), secondary structure (5), and cavity shape (180). A machine learning technique, Random Forests, is used to train a classifier to distinguish drug-binding from non-drug-binding cavities using the computed cavity property profile. Random Forests technique is robust to the presence of a large number of irrelevant variables, it does not require prior scaling of input variables, and it can detect and take advantage of high-order interactions. In this study SCREEN analyzed 99 non redundant, comprehensive protein-ligand complexes from the PDB. The drug-binding cavities were predicted with a balanced error rate of 7.2% and coverage of 88.9%. A technique using only cavity size for prediction gave a balance error rate of 15.7% and coverage of 15.7%. The paper provides examples of instances where the Random Forests based classifier was able to predict drug-binding cavities that would have been missed using simple 'largest surface cavity' criterion. Only 18 of the 408 cavity attributes had a statistically significant role in the prediction. Of these 18 important attributes, almost all involved size and shape rather than physiochemical properties of the surface cavity, A SCREEN web server is available at http://interface.bioc.columbia.edu/screen.