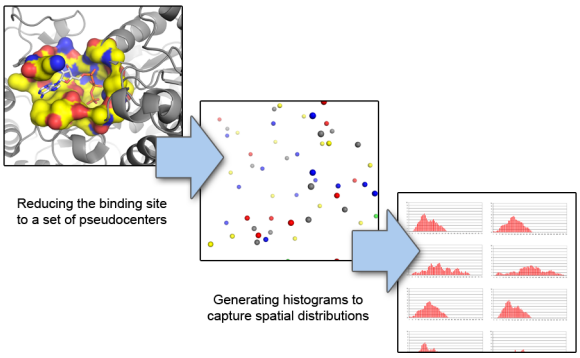
Efficient Comparison of Protein Binding Sites using Distance Histograms

Timo Krotzky, Gerhard Klebe

Institute of Pharmaceutical Chemistry, Philipps-Universität, 35032 Marburg/Lahn, Germany



Efficient determination of structural similarities between protein binding sites is one of the remaining challenges in computational chemistry and drug design as it can help to understand selectivity considerations and to predict unexpected cross-reactivity. The mutual comparison is often approached by using graphs as a way to represent and calculate metrics such as the maximum shared common subgraph to estimate a degree of similarity. Cavbase [1, 2] was developed as a tool for the automatic detection, storage and classification of putative binding sites. Cavbase assigns so-called pseudocenters to the cavity-flanking amino acids, which characterize their physicochemical properties with respect to molecular recognition. Subsequently, the pseudocenters are used as graph nodes to accomplish mutual binding site comparisons. However, the modeling of protein binding sites by means of graphs tends to be computationally very demanding, which often leads to very slow computations of the similarity measures. While this is acceptable when just a couple of structures are compared, it becomes inadequately slow for large data sets or the screening of entire databases.

In this study, we propose a Pocket Comparison using Spatial Distributions (PoCuSD), a new modeling formalism for Cavbase structures which allows for an ultrafast comparison procedure that performs similarity calculations very efficiently. Here, protein binding sites are represented by sets of distance histograms based on specific spatial reference points [3]. They characterize the distribution of pseudocenters within the cavity and can be both generated and compared with linear complexity. Attaining a speed of approximately 20,000 comparisons per second, similarity calculations across large data sets and even screenings of entire databases become easily feasible.

We demonstrate the discriminative power and the very fast runtime of this method by carrying out several classification and retrieval experiments. Among others a well studied data set of protein cavities binding either ATP or NADH is used for a classification experiment, where PoCuSD results in a considerably higher rate of correct classifications than many of the hitherto approaches while requiring only a fraction of their runtime.

[1] S. Schmitt, D. Kuhn, G. Klebe, *J Mol Biol*, **2002**, *323*, 387-406.
[2] M. Hendlich, F. Rippmann, G. Barnickel, *J Mol Graph Model*, **1997**, *15*, 359-363.
[3] P. J. Ballester, W. G. Richards, *Proc. R. Soc. A*, **2007**, *463*, 1307-1321.