Ansgar Schuffenhauer,* Nathan Brown, Paul Selzer, Peter Ertl, and Edgar Jacoby
Novartis Institutes for BioMedical Research, CH-4002 Basel, Switzerland
J. Chem. Inf. Model., 2006, 46 (2), pp 525–535
DOI: 10.1021/ci0503558
Publication Date (Web): December 14, 2005

Abstract

Following the theoretical model by Hann et al. moderately complex structures are preferable lead compounds since they lead to specific binding events involving the complete ligand molecule. To make this concept usable in practice for library design, we studied several complexity measures on the biological activity of ligand molecules. We applied the historical IC50/EC50 summary data of 160 assays run at Novartis covering a diverse range of targets, among them kinases, proteases, GPCRs, and protein–protein interactions, and compared this to the background of “inactive” compounds which have been screened for 2 years but have never shown any activity in any primary screen. As complexity measures we used the number of structural features present in various molecular fingerprints and descriptors. We found generally that with increasing activity of the ligands, their average complexity also increased, and we could therefore establish a minimum number of structural features in each descriptor needed for biological activity. Especially well suited in this context were the Similog keys and circular substructure fingerprints. These are those descriptors, which also perform especially well in the identification of bioactive compounds by similarity search, suggesting that structural features encoded in these descriptors have a high relevance for bioactivity. Since the number of features correlates with the number of atoms present in the molecule, also the number of atoms serves as a reasonable complexity measure and larger molecules have, in general, higher activities. Due to the relationship between feature counts and densities on one hand and biological activity on the other, the size bias present in almost all similarity coefficients becomes especially important. Diversity selections using these coefficients can influence the overall complexity of the resulting set of molecules, which has an impact on the biological activity that they exhibit. Using sphere-exclusion based diversity selection methods, such as OptiSim together with the Tanimoto dissimilarity, the average feature count distribution of the resulting selections is shifted toward lower complexity than that of the original set, particularly when applying tight diversity constraints. This size bias reduces the fraction of molecules in the subsets having the complexity required for a high, submicromolar activity. None of the diversity selection methods studied, namely OptiSim, divisive K-means clustering, and self-organizing maps, yielded subsets covering the activity space of the IC50 summary data set better than subsets selected randomly.