where

$$\mu = \frac{N_1(N_1 + N_2 + 1)}{2}$$

$$\sigma = \sqrt{\frac{N_2\mu}{6}}$$

and T is the smaller sum of ranks in an ordered list of the merged set of activities for the two half-decades. For these census data, $T = 1468$, $N_1 = 32$, $N_2 = 67$, and Z was therefore 0.98. By reference to a normal distribution table, this corresponds to a 68% confidence that the average affinities are different. Note that there were three virtual screens for which the reported high-potency hits were not sufficiently well characterized (vide infra); if these three values are removed, it is 84% likely that the average affinities for 2005–2008 are less potent than for 2000–2004.

Recommendations from census results

In the previous section, we saw that distributions of hit activities differed between the periods 2000–2004 and 2005–2008 and that there was a 68% probability that the average potency decreased for the latter half-decade. One hypothesis that might explain these statistically significant differences is that docking algorithms have gotten worse over the decade; however, the number of hits <10 μM has remained relatively constant across the years, an observation that is more consistent with the hypothesis that docking algorithms as a class have performed at a consistent level across the decade. Although there have almost certainly been modifications and improvements to specific docking algorithms, the census data for prospective virtual screens suggest that those improvements have been incremental at best. A second hypothesis to explain increased average hit IC_{50}s is that docking has been applied to harder targets during the second half-decade. Although I have not painstakingly catalogued the target classes in all ninety-nine virtual screens, a quick scan of the specific targets in these screens suggests that differences in target class do not explain differences in average activity. Enzymes are the most represented class of targets, with kinases, proteases, transferases, phosphatases, and so forth having been screened regularly during the years in this survey. If anything, targets were more difficult earlier in the decade, with a few brave (foolhardy?) researchers applying docking to lead identification for ion channels, protein/protein interactions, and even G-protein-coupled receptors (GPCRs). It is also plausible to hypothesize that the pattern of activities, especially in 2007 and 2008, is due to a greater willingness to publish computational studies with less positive results. If so, that would be a valuable practice for the field as it would allow a more accurate assessment of how docking algorithms perform in the real world of prospective screens rather than in retrospective tests. And one final hypothesis must be that docking has become a tool for the unwary; docking programs have gotten easier to use, performance improvements have been made to individual docking programs, more structural data and larger collections of purchasable compounds have become available, and all of these factors have led to more opportunities for less experienced users to give it a try. No matter what the underlying explanation for differing distributions, a closer examination of the eighteen most potent hits in the census, listed in Table 7.3, suggests strategies that might improve the chances for identifying submicromolar hits from docking-based virtual screens.

To identify docking strategies that might have been conducive to the identification of submicromolar hits, all ninety-eight references were read but the eighteen references in Table 7.3 were examined much more carefully. Of the eighteen virtual screens represented in Table 7.3, the vast majority sought enzyme inhibitors; of the three nonenzyme protein targets, two virtual screens sought competitive binders to the estrogen receptor and the third sought ATP-competitive antagonists of the chaperone Hsp90; none of the eighteen virtual screens targeted the much more challenging ion-channel, protein/protein interaction, or GPCR target classes. Of the enzyme targets in these eighteen virtual screens, the most highly populated classes were kinases (4) or oxidoreductases (3). One might therefore hypothesize that the secret to success would be to carry out a docking-based virtual screen against a protein kinase. Assessing that hypothesis more closely, 22% of the targets in the <1 μM activity bin are kinases; in contrast, 20% of the targets in the combined 1–10 μM and 10–100 μM bins are kinases, while there are no kinases among the nine virtual screens with hits >100 μM. The more likely hypothesis, then, is that a virtual screen of a protein kinase is likely to produce hits with measurable experimental activity, but those hits are as likely to be 10 μM as 10 nM. Instead, exact methodological details of how a virtual screen was run seems a more important factor in success rates for identifying the more potent hits:

- Of the virtual screens in Table 7.3, only four used any variant of the NCI database as a source for searchable compounds, while a higher proportion of screens with less potent hits searched databases from that source. Instead, the virtual screens with submicromolar hits were more likely to search in-house or corporate collections for which care had been taken in choosing the compounds that populate the search database. Given the 2005 publication of the ZINC virtual screening database,[141] there is no longer any reason for even those researchers without access to large corporate collections to not search a carefully chosen and curated database.
- Once a database for searching had been selected, the virtual screeners of Table 7.3 generally prefiltered that database to remove unappealing molecules. In some cases, this filtering was as simple as the removal of reactive or non-drug-like molecules. In other cases, search databases were filtered to remove compounds incompatible with chemical features of the protein binding site – for example, removing anionic compounds before