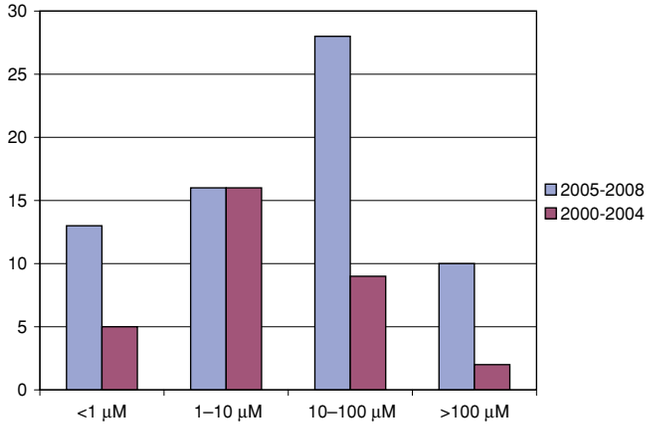

Figure 7.2. Docking-based virtual screens for which experimental results were reported, January 2000–July 2008; distribution of potencies by half-decade.

of compounds with activities less than 10 μM was roughly constant across the census period; the largest increase in number of reports was for compounds with IC_{50} greater than 10 μM. Given the increase in number of hits with poorer activity, it at first blush appears that docking did not improve but instead stagnated or even deteriorated during the census period. Because there were as many as ninety-nine individual reports in the census, a sufficient amount of data was available to more carefully assess whether the apparent trend toward less potent compounds was statistically significant.

The census data were further aggregated by half-decade to smooth out the annual fluctuations in number of reported screens and in distribution of hit activities, particularly for the dips in number of reports in 2002, 2004, and 2006. The binned data for the two half-decades, 2000–2004 and 2005–2008, are shown in Figure 7.2. The distribution for 2000–2004, magenta bars in Figure 7.2, is roughly symmetric with a peak at the 1–10 μM bin, while the 2005–2008 distribution, periwinkle bars, peaks at 10–100 μM but has a substantial tail on the more potent, lower IC_{50} side of the graph. Both the shapes of the distributions and the average activities appear to differ between the first and second half of the decade.

Two statistical tests were applied to assess the significance of the apparent differences in distributions and averages between the two half-decades. Differences in distribution were assessed using the χ^2 test for consistency in $2 \times K$ table.[138] The null hypothesis for this test asserts that the histograms in Figure 7.2 reflect samples drawn from two underlying distributions that are identical; a χ^2 value substantially greater than zero would support the visual impression that the two half-decade distributions differ. Expected frequency distributions for the two half-decades were computed according to Equation (7.11):

$$e_{ij} = \frac{N_i n_{ij}}{N_1 + N_2}, \tag{7.11}$$

Table 7.2. Docking-based virtual screens for which experimental results were reported; average IC_{50} values for each half-decade

Time period	Median activity	95% Confidence interval
2000–2004	5 μM	2–20 μM
2005–2008	13 μM	5–20 μM

where i takes on the values 1 and 2 representing each of the half-decades, j ranges from 1 to 4 representing each activity bin in the histogram, n_{ij} is the number of samples for a specific activity bin in a specific half-decade, N_1 is the total number of samples for 2000–2004, and N_2 is the total number of samples for 2005–2008, resulting in e_{ij}, two new distributions in which the observed frequencies in Figure 7.2 have been normalized by the fraction of all reported virtual screens that occurred during each half-decade. The χ^2_3 statistic for three degrees of freedom was computed from these normalized expected frequencies according to Equation (7.12):

$$\chi^2_3 = \sum_{j=1}^{4} \frac{(n_{1j} - e_{1j})^2}{e_{1j}} + \sum_{j=1}^{4} \frac{(n_{2j} - e_{2j})^2}{e_{2j}}. \tag{7.12}$$

The computed χ^2_3 statistic for the distributions in Figure 7.2 is 7.17, which corresponds to a cumulative probability of 0.93.[139] We can therefore reject the null hypothesis at the $p < 0.07$ significance level and conclude that the underlying distributions likely differ between the periods 2000–2004 and 2005–2008.

For the 2005–2008 period, an increase was seen for the number of hits in both the high-potency "<1 μM" and the poorer potency "10–100 μM" categories. Therefore, although the shapes of the distributions differ, the null hypothesis that average activities for the half-decades are identical remains plausible. Medians were used to examine average activities because medians are robust to experimental errors (a likely issue with data for many different proteins measured in many different labs), robust to outliers such as the three no-hit examples to which a measured activity value could not be assigned, and robust to nonnormal distributions – we have no reason to expect a Gaussian distribution for these data. Median activity values were computed along with the 95% confidence intervals for those averages (Table 7.2). The median activity for the second half-decade was less potent than that for the first half-decade, but there was significant overlap of the range in which the true average probably lies. The Wilcoxon-Mann–Whitney nonparametric rank-order test was therefore used to assess the statistical significance of the differences in average affinity[140]:

$$Z = \frac{|\mu - T| - 0.5}{\sigma}, \tag{7.13}$$