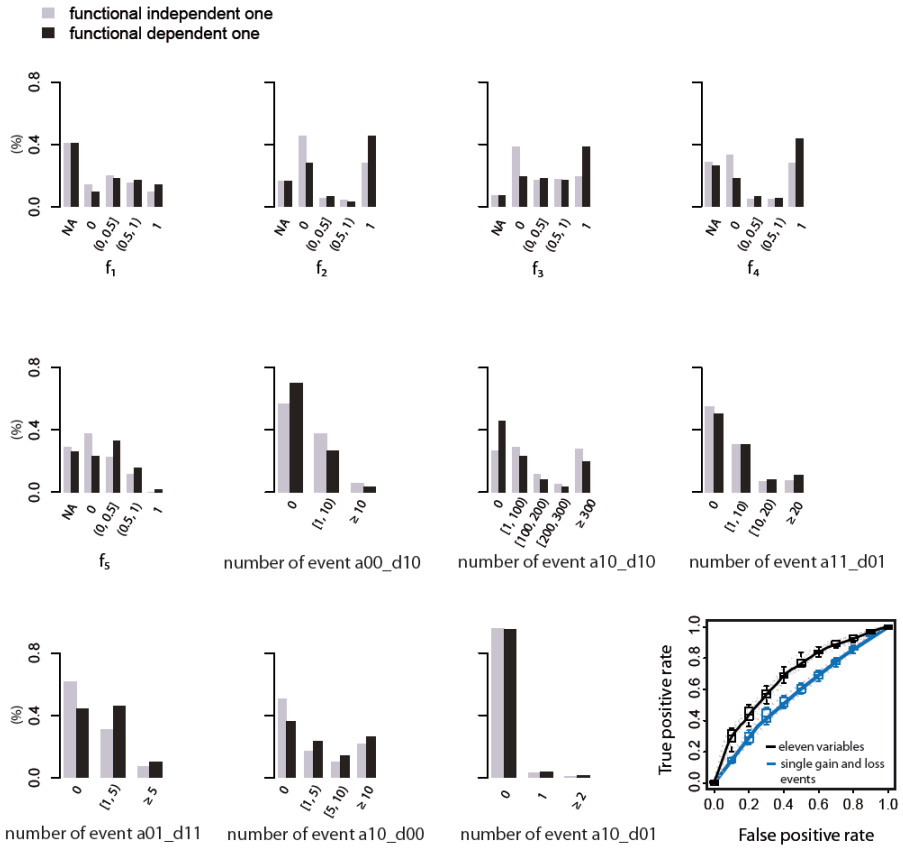
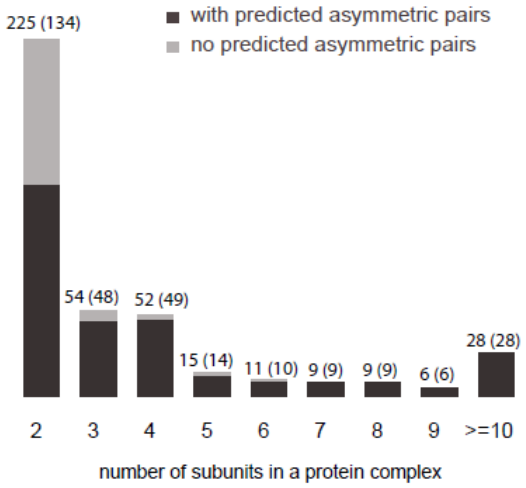
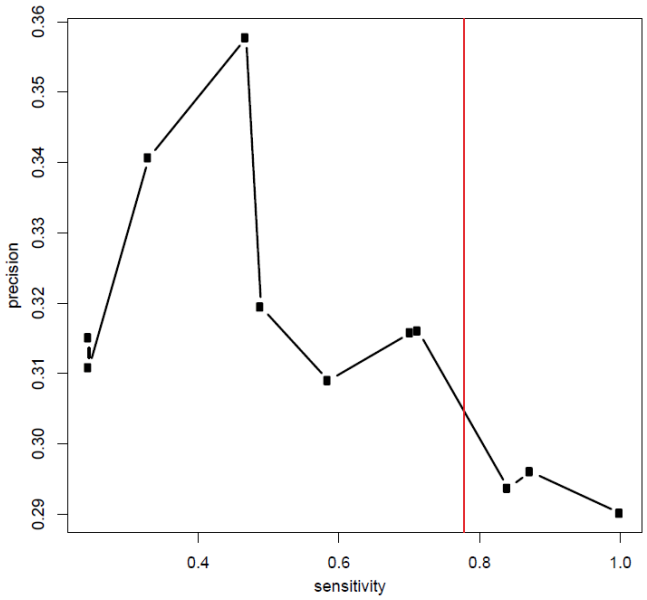
Supplementary Information



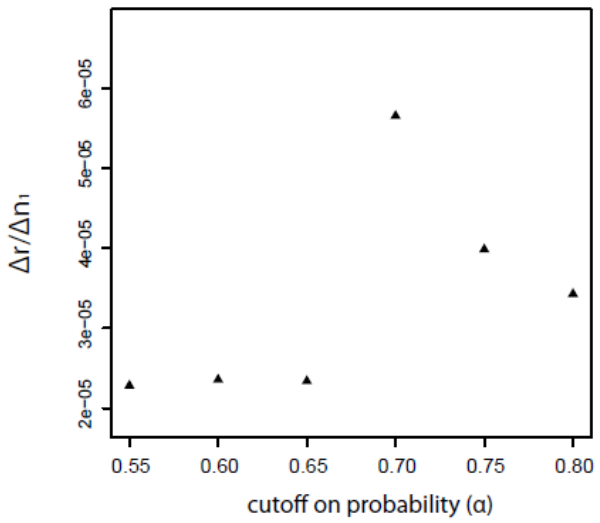
Supplementary Figure S1 Distribution of dependent and independent genes for 11 evolutionary variables. For each evolutionary variable, the distribution of functionally independent genes (grey bars) and functionally dependent genes (black bars) is plotted. For all 5 fractions, the functional dependent one is enriched over the independent one for the scores 1, while the independent one is enriched for score 0. For 2 evolutionary events a00_d10 and a10_d10, it is the functional independent one that occurs more often than the dependent one across number of events (except 0). For 4 evolutionary events a11_d01, a01_d11, a10_d00 and a10_d01, it is the functional dependent one that occurs more often than the independent one across number of events (except 0). The black receiver operator characteristic (ROC) curve shows the performance of the model integrating 11 evolutionary variables. The blue ROC curve shows the performance of the model integrating only single gain and loss events. The model with 11 evolutionary variables integrated has an area under curve (AUC) of 0.7, while the model with only single gain and loss events only has an AUC of 0.568. The error bars in the ROC curve are the standard deviations of cross-validations (n=10) .



Supplementary Figure S2 Distribution of functional asymmetry in protein complexes Each bar shows how many of the complexes contain at least one predicted functionally asymmetric pair, and the number of complexes with asymmetric pairs is in parentheses



Supplementary Figure S3 Prediction precision and sensitivity The black dotted line is the plotting of sensitivity against precision of our prediction model. Each dot represents the precision and sensitivity of our model at a cutoff value of asymmetry (α, see methods). The red line is the sensitivity of the Pandey's model (0.78).



Supplementary Figure S4 Cost of increased correct prediction rate on classified samples For each cutoff on probability of a gene being independent or dependent (α), the number of classified samples, independent and dependent (n_1), and the correctly classified samples (n_2) are counted. The correct classification rate (r) is calculated as n_2/n_1. With the cutoff increasing from 0.5 to 0.8, the correct classification rate increases from 0.644 to 0.72 while the number of classifiable genes decreased from 4800 to 2467. The criteria for cutoff on probability are that i) most samples can be classified by the model and ii) a high correct classification rate. Combining these two criteria, we determined the cutoff for significant classification at α=0.7 with a largest $\Delta r/\Delta n_1$, where the correct prediction rate increases at the least cost of coverage.