We here report a statistical fact in support of the basic assumption underlying our model. The matching condition we employ dictates a certain correlation between the sets of regulated genes by each TF: if the binding sequence of a TF (A) is embedded in that of a TF (B), then the set of genes \( \{G_i\}_B \) regulated by TF\(_B\) in our model is a subset of \( \{G_i\}_A \). A similar investigation of the yeast databases listed below reveals that the top 50% of the TF pairs related by the sequence inclusion relation above, rank in the top 3% when all the TF pairs are listed according to the overlap of their \( \{G_i\} \) sets. The actual ranking of the TF pairs obtained among all possible pairs of 102 TFs with known binding sequences is shown in Fig. 1.

On the other hand, the more straightforward expectation that TFs with short binding sequences should regulate more genes is not verified by the same data. This curious fact probably points to certain sequence correlations arising from the duplication and divergence processes \(^\dagger\) that distort the occurrence statistics of the binding sequences in PRs. Note that the result in Fig. 1 is robust to such deviations from the unbiased probabilities for the occurrence of different strings.

Figure 1: Correlation between the sets of proteins regulated by the TFs with similar binding sequences. The vertical axis is the percentage overlap of the two sets of genes regulated by an arbitrary pair of TFs, which are ranked on the horizontal axis according to their overlap. The red vertical lines mark those pairs of TFs that are also related by binding sequence inclusion. The accumulation of the red lines to the left of the graph is indicative of the correlation described in the text.