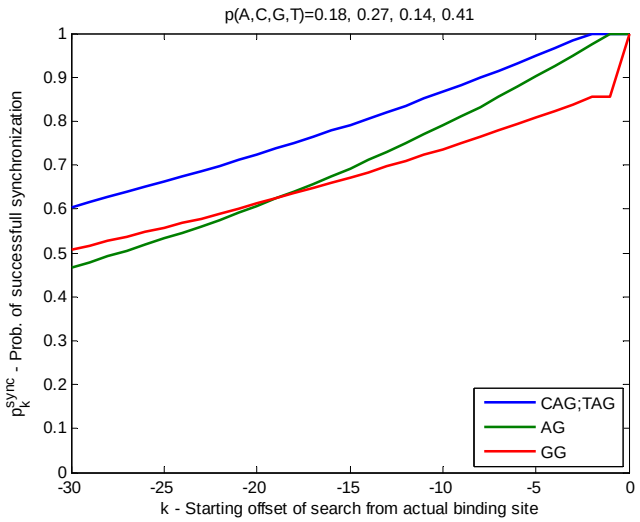
Synchronization Properties of Protein Binding Sites

Pavol Hanus, Inst. for Communications Engineering, Technische Universität München

Protein-DNA and protein-RNA binding sites recognition often takes place in form of sequence specific one-dimensional diffusion along the respective strand. In this sense it strongly resembles a well studied problem of frame synchronization in communications engineering. In this scenario the information frames are embedded into a random data stream and can be recognized by their pre-amble carrying a particular synchronization pattern. The receiver starts listening to the data stream at any point and searches for the next occurrence of the synchronization word in a sliding window process. In fact, the probability of observing a spurious pattern in the random data preceding the actual pre-amble during the sliding process is dependent not just on the pattern length, but also on the pattern itself. The choice of the synchronization word is thus crucial for the reliability of the detection of the information frame. In the course of this work a trellis based algorithm for the evaluation of the synchronization performance of different patterns was used to assess synchronization properties of actual binding sites (e.g. the -35 and -10 promoters in E. Coli recognized by the RNA polymerase or the 3' splice sites detected by the spliceosome). The algorithm was adapted to the possibility of unequal background distribution of nucleotides and multiple valid different binding patterns. It turns out that the patterns employed by nature are actually very good and among those that would be chosen by a communication engineer.

The algorithm operates by calculating the probability of successful synchronization for every offset to the real binding site the search could possibly start from (see Figure). Given the distribution of these offsets it is thus possible to calculate the average synchronization success rate for each pattern of a particular length and to determine the least error prone pattern. The distribution can usually be derived from the dissociation probability of the searching protein from the strand during the one-dimensional diffusion.



The presented Figure depicts the probability plot of the 3' splice site recognition. In [1] it is suggested that the spliceosome performs a directed linear search for the AG splice site starting from the branch point. The splice site is usually situated close by - at most 30 nucleotides downstream. Even though the GG pattern might seem to be the best choice based on the background nucleotide probabilities, since G is the least probable base, for short offsets it is outperformed by AG making it clearly the best choice. The Figure also shows the plot for the typical three bases long 3' splice site consensus YAG.

[1] S. Chen, K. Anderson, M.J. Moore; "*Evidence for a linear search in bimolecular 3' splice site AG selection*," PNAS, Vol. 97, Issue 2, 593-598, January 18, 2000