Thought we have a DP algorithm to compute the hit probability of a seed, it is still useful to have some sense of the “dependency” between the 1’s and its relation with the general hit probability. In the following problem, consider two seeds:

Seed_X = 1101011

Seed_Y = 11111

where

1 for match and 0 for don’t care in the seeds

L = the length of the query and subject sequence

p = similarity of the two sequence = Prob(the two sequence has same symbol)

M = the length of the seed

W = the weight of the seed

assuming that there are only match/mismatch in any pair of symbols and overlapped hits counts as different hits

1. In the two seeds, which seed is likely to be used in PatternHunter?

A: Seed_X since PatternHunter uses spaed seeds

2. If we already have a hit in a pair of sequence, what’s the expected total number of hits we can find in this pair?

A: If we have already a match in the positions of 1’s of a hit, we need less new matches than its actual weight in the region overlapping the first hit. We can regard Seed_Y as “1111100” which has the same window length as Seed_X does, and in the overlapped regions the probability to find a second hit is

For Seed_X: $(p^3 + p^3 + p^3 + p^4 + p^3 + p^4) * 2$

For Seed_Y: $(p + p^2 + p^3 + p^4 + p^5 + p^5) * 2$

the probability is multiplied by 2 since the window can be shifted backwards and forwards.

Outside the region, the probability to find a second hit is independent of the first hit, hence it is p^5 for all possible positions for both seeds.

3. If we can approximate the probability to hit by the formula

E(#hits) = Pr(seed hits) * E(#hits | seed hits)

what’s the order of the seeds in terms of E(#hits)? in terms of Pr(seed hits)? which criterion is more important, why?

A: E(#hits) are $(L – M + 1)\ p^W$ for both seeds, hence E_X(#hits) = E_Y(#hits)

Pr(seed hits) = E(#hits) / E(#hits | seed hits), hence Pr_X(seed hits) > Pr_Y(seed hits)

Pr(seed hits) is more important, since ideally we need just one hit for further extension in each significant alignment of two homologous sequences; in other words, finding more hits in the same alignment does not contribute to the sensitivity of the alignment algorithm.