Finding regions of exact identity between two DNA sequences

Ana Vukelić, *Faculty of Food Technology and Biotechnology*

Abstract

The recombination model assumes that recombination events take place in regions of exact identity between the two sequences: Effective Pairing Sequences (EPS). It is assumed that the EPS must be longer or equal than the Minimal Effective Pairing Sequence (MEPS), but no sequence specificity is assumed. A program was written to input two DNA sequences and list all maximal identical segment pairs of different lengths. In order to model recombination the MEPS length parameter was defined so that shorter-segment pairs are not considered as EPS. If it is assumed that *Streptomyces* chromosomes are random 10 Mb sequences of 72% G+C-content, a MEPS length of 16 b would ensure that the probability of encountering a second pairing site in the genome would be less than 10%. However, the genuine genome DNA sequences are less random, so a program was written to count the number of identical sub-sequences of different lengths between two DNA sequences. Pairs of 200 kb sequences were extracted from the chromosomes of the three *Streptomyces* species, *S. coelicolor* A3(2), *S. avermitilis* and *S. scabies*, and the number of identical sub-sequences of lengths 11-60 bases are counted. If the number of sequences counted in the three species is not significantly different (statistical test), the sequences could be considered as typical. So, we can calculate the MEPS length which would give a probability of less than 10% of false pairing.