ABSTRACT

We describe a program to assist in designing oligonucleotide probes on the basis of protein alignments and the codon usage of the target organism. If necessary, the input sequences can be weighted to neutralise the effect of closely similar sequences or to bias the output in favour of a particular taxon.

Numerous computer programs exist to select the optimal oligonucleotide from a single nucleotide sequence for use as PCR primers or hybridisation probes (e.g. 1–3). However, to isolate genes for new members of a protein family, it is frequently necessary to design oligonucleotides on the basis of protein alignments. Such oligonucleotides will usually correspond to the most highly conserved region of the alignment and should reflect the codon usage of the target organism. A widely used procedure is to simply choose the most frequently used codon for the commonest amino acid at a given position, but this ignores the contribution made by other amino acids. Furthermore, it does not allow individual bases within the codon to vary independently, an unnecessary constraint which leads to further loss of information.

We describe here a program, DNAPROBE, to assist in the design of oligonucleotides in such cases. This uses MSF protein alignments generated by the program PILEUP and codon frequency tables from the Wisconsin (GCG) sequence analysis software package. The basic algorithm calculates the most probable nucleotide sequence corresponding to the alignment by summing the products of the frequency with which a particular amino acid occurs at any one position and the codon usage frequency for that amino acid. Where an alignment contains a cluster of closely related sequences, the user may choose to implement a weighting option that reduces their individual contributions to the output in order that they do not overwhelm the contributions of more distantly related sequences which are under-represented. This may be done in two ways: (i) using the algorithm of Gerstein et al. (4) and the information contained in the dendrogram file written by PILEUP, or (ii) by specifying weights in a separate file which is then read by the program, an option which enables one to give increased weight to sequences known to be more closely related to the target organism.

The user specifies the length of the oligonucleotide required, and the program then scans the most probable nucleotide sequence to find segments of this length having the greatest overall probability of matching the sequence of an unknown gene of the desired class. Users will generally wish to exclude probes derived from parts of the alignment that contain gaps, and the program contains an option to do so. A chosen number of the best oligonucleotide segments are listed, together with their matching probability, their melting temperatures according to the algorithm of Freier et al. (5) and their positions in the sequence. The probability of occurrence of the preferred base at each position in the sequence can be displayed graphically, as well as the cumulative probability for consecutive oligonucleotides over the entire sequence. The program displays the most probable nucleotide sequence together with its translation, which is also aligned to the input protein sequences. This hypothetical translation product is useful for matching oligonucleotides to the input protein sequences, although occasionally it contains amino acids not present in the original alignment as a result of different forces predominating at different positions within a codon; this is not a cause for concern (Fig. 1).

The efficiency of both PCR and filter hybridisation can be enhanced by using inosine (6) or mixtures of bases at positions for which no strong prediction can be made, and to facilitate this kind of choice, DNAPROBE contains an option to display the probability of occurrence for each base at every position of a chosen oligonucleotide.

The program is written in C++, runs under Windows, and may be downloaded from WWW.HENSA.AC.UK

REFERENCES