Computer selection of oligonucleotide probes from amino acid sequences for use in gene library screening

Junhui Yang¹, Jianhong Ye² and Douglas C. Wallace^{2,3*}

¹School of Engineering Science and Mechanics, Institute of Technology of Georgia, Atlanta, GA 30332, ²Department of Biochemistry and ³Department of Pediatrics, Division of Medical Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA

Received 23 August 1983

ABSTRACT

We present a computer program, FINPROBE, which utilizes known amino acid sequence data to deduce minimum redundancy oligonucleotide probes for use in screening cDNA or genomic libraries or in primer extension. The user enters the amino acid sequence of interest, the desired probe length, the number of probes sought, and the constraints on oligonucleotide synthesis. The computer generates a table of possible probes listed in increasing order of redundancy and provides the location of each probe in the protein and mRNA coding sequence. Activation of a next function provides the amino acid and mRNA sequences of each probe of interest as well as the complementary sequence and the minimum dissociation temperature of the probe. A final routine prints out the amino acid sequence of the protein in parallel with the mRNA sequence listing all possible codons for each amino acid.

INTRODUCTION

Since R. Wu first proposed that oligonucleotide probes could be deduced from known amino acid sequences (1) and such a probe was used to identify the yeast cytochrome c gene within a gene library (2), the use of chemically synthesized oligonucleotide probes in cloned gene identification has experienced rapidly increasing popularity. Interest in this strategy has been greatly increased by the rapid refinement in methods for oligonucleotide synthesis (3). Oligonucleotide probes are now used not only to screen genomic libraries, but also to screen complementary DNA (cDNA) libraries prepared from mRNA (4,5) and as primers which when annealed to mRNA mixtures permit the selective extension of the oligonucleotide using the mRNA as a template (6,7).

The success of all of these strategies depends on the specificity of the oligonucleotide probe used, which is in turn a

function of the redundancy and length of the probe and its affinity for the desired gene, cDNA or mRNA sequence. At present optimal probe sequences are sought by hand by "reverse translating" the amino acid sequence into the mRNA sequence employing all possible codons and then looking for regions of minimal redundancy. This procedure is both prone to error and time consuming. To increase the reliability of this essential step, we have devised a computer program, FINPROBE, which rapidly and automatically identifies optimal oligonucleotide probes complementary to the nucleic acid sequences which could code for the protein. In this paper we describe the principles on which this program is based and provide as an example an analysis of the carboxy-terminal 94 amino acids of the beef heart mitochondrial ADP-ATP translocator (8).

MATERIALS AND METHODS

This program was developed on an IBM Personal Computer having 320 KB diskette drives and a NEC Spinwriter 3010 printer as peripherals. The program was written in IBM Personal Computer Advanced Basic (Version A 1.10, Copyright IBM Corp., 1981, 1982) and requires a diskette drive and at least 11 kilobytes RAM (10 kilobytes for the program and 0.6 kilobytes for handling each 100 amino acids on file) to run.

RESULTS AND DISCUSSION

Parameters of the Program

This program scans a given amino acid sequence for those regions which could be coded by the least number of possible mRNA sequences. Generally, this favors regions rich in amino acids having few alternative codons (e.g., methionine, tryptophan, etc.). The program behaves as if it "reverse translates" the amino acid sequence into all possible nucleotide sequence combinations and then selects regions of a user prescribed length (e.g., 14 nucleotides) with the least number of combinations.

Once the program is started, the user is confronted with a FUNCTION SELECTION MENU. This menu has four functions. Function 1 is "Handle Amino Acid Sequence File". Function 2 is "Choose and List Probes". Function 3 is "Print a Table of

AA/mRNA/PROBE". Function 4 is "Print the Whole Sequence of AA/mRNA".

Function 1 (Handle Amino Acid Sequence File) must be selected first and permits the input and editing of the amino acid sequence of interest. Each amino acid file is given by users an IBM DOS file name and is saved on diskette. If the diskette already has a file of that name, it is read into the core memory. Amino acid sequences are entered at 10 amino acids per line in either standard one letter or three letter notation. One letter abbreviations are entered as a continuous series of letters while three letter abbreviations must be separated by a space. The amino acid sequence can be edited by making corrections, insertions, or deletions.

Function 2 (Choose and List Probes) actually identifies the least redundant probes of a specified length which occur throughout the amino acid sequence of interest. The user designates the length of the probe (L), the number of probes to be sought (up to 50), the region of the amino acid sequence file to be searched (specify first and last amino acid of the region of interest separated by a comma) and the mode of calculation of the least redundant probes (M, MD, MDT). The program then calculates the number of nucleotide sequence combinations (SC) for each region of the protein and prints out a table of the least redundant probes.

The modes of calculation of oligonucleotide redundancy (SC) reflect current constraints on oligonucleotide synthesis techniques. The first mode (M) calculates the number of sequence combinations which would be generated if a probe for the region was synthesized by sequentially adding one base (monomer) at a time or with mixtures of bases being added at points where an ambiguity was encountered. The second mode (MD) determines the number of combinations expected if a probe was synthesized using either sequential addition of single bases or addition of mixtures of presynthesized dinucleotides. Inclusion of dinucleotides effects the degeneracy of probes which encompass serine residues. Serine has six codons varying in all three codon positions. Inclusion of serine in the M mode results in 16 base combinations (2x2x4) while in the MD mode the first two

bases can be added as a pair of dimers (UC and AG) resulting in 8 base combinations. The third mode (MDT) calculates the number of combinations if the probe could be synthesized by sequential addition of mononucleotides or presynthesized dimers or trimers. This assumption has the greatest effect on sequences which include leucine, arginine or serine, where mixtures of trimers including all six codons would be used. Inclusion of presynthesized trimers has the effect that the MDT mode calculates the actual number of mRNA sequences which could code for the amino acid sequence of the probe. Consequently, this calculation mode will be of greatest value to users who wish to synthesize each of the possible mRNA sequences separately for independent use (9).

In addition to these modes, the program will also permit the user to specify the use of G-T pairing (1,11) in M, MD and MDT calculations. In this GT subroutine, a G is inserted in the probe when a U/C ambiguity is encountered in the mRNA and a T when a G/A ambiguity is encountered. This modification greatly reduces probe redundancy. However, G-(T/U) pairs also reduce the stability of RNA-DNA and DNA-DNA hybrids (10,11,12,13,14). The output lists those probes with the fewest G-T pairs first.

Once the amino acid sequence has been scanned in a particular mode, a table is printed listing the requested number of probes in increasing order of redundancy and listing the probe locations within the amino acid and nucleotide sequences. Such a table is printed in Fig.1.

Function 3 (Print a Table of AA/mRNA/PROBE) provides detailed information on the probes of interest. The user designates the number of the probe and the program prints the amino acid sequence of interest, all possible nucleotide sequences for each amino acid (mRNA), all possible complementary nucleotide sequences (probe), and the minimum dissociation temperature (T_d) of the probe. T_d is valuable in determining hybridization conditions and is calculated by the empirical formula $T_d = 2^\circ\text{C}$ times the number of AT base pairs + 4°C times the number of GC base pairs (10). A value of 2°C is given to positions where either an AT or GC base pair may be located. T_d is not calculated when G-T pairing is permitted. (Fig.2).

```

      FILENAME:  ADP-ATP.TR

      SCANNING AA SEQUENCE FROM      : 1
      LENGTH OF PROBES                : 14
      OLIGONUCLEOTIDE SYNTHESIS METHOD : MD

      # OF          COMBINATION OF          PROBE STARTING POSITION
      PROBES        OLIGONUCLEOTIDE MIXTURES  mRNA(FROM 5'END) AA(FROM N-END)

      #1             4                    100 - 113      34(1) - 38(2)
      #2             8                    97 - 110       33(1) - 37(2)
      #3             8                    98 - 111       33(2) - 36(3)
      #4            12                    28 - 41        10(1) - 14(2)
      #5            12                   130 - 143       44(1) - 48(2)

```

Fig.1. Output of Function 1. A list of the five least redundant 14 base oligonucleotide sequences found by the MD calculation mode within the carboxy-terminal 94 amino acids of the ADP-ATP translocator. Numbers in parentheses under AA(FROM N-END) give the number of bases included in the first and last codon.

The final function 4 (Print the Whole Sequence of AA/mRNA) prints out the entire amino acid sequence of the protein and prints in parallel all possible codons for each amino acid. This information is helpful in localizing the position of probes of interest (Fig.3).

Structure of the Algorithm for Calculating the Degeneracy of Oligonucleotide Probes

The program starts at the designated amino acid address (N) and calculates the corresponding nucleotide sequence address of

```

      PROBE #1          FILE NAME:  ADP-ATP.TR

      AMINO ACID        34
      (N-TERMINAL)      M      M      M      Q      S

      mRNA 5'          100
                        AUG      AUG      AUG      CAA      UC
                                      G

                                      AG

      PROBE 3'          TAC      TAC      TAC      GTT      AG
                                      C

                                      TC

      Td= 38 ~C          LENGTH:  14
      COMBINATION:      4

```

Fig.2: Output of Function 3. Detailed information on Probe 1 of Figure 2 showing the amino acid, mRNA and probe sequences and the Td value.

31-40	R	R	R	M	M	M	Q	S	G	R
	CGU	CGU	CGU	AUG	AUG	AUG	CAA	UCU	GGU	CGU
	C	C	C				G	C	C	C
	A	A	A					A	A	A
	G	G	G					G	G	G
	AGA	AGA	AGA					AGU		AGA
	G	G	G					C		G
41-50	K	G	A	D	I	M	Y	T	G	T
	AAA	GGU	GCU	CAU	AUU	AUG	UAU	ACU	GGU	ACU
	G	C	C	C	C		C	C	C	C
		A	A		A			A	A	A
		G	G					G	G	G

Fig.3. Part of the output of Function 4. A list of the amino acid sequences and all mRNA codons of the ADP-ATP translocator in the region of probe 1 (Fig.1) including amino acids 31 to 50 in the file.

the beginning base (B) = $N \times 3 - 2$ and the ending base = $B + L - 1$ where L is the designated length of the probe. In relation to the amino acid sequence this probe nucleotide sequence is composed of 3 parts (HEAD, MID, and TAIL). The MID region is that portion of the sequence which includes complete amino acid codons. The HEAD and TAIL are those portions at the beginning and end of L which include parts of codons. The number of bases in these regions are calculated by $HEAD = 3 - [(B-1) \text{ MOD } 3]$ and $TAIL = (B+L-1) \text{ MOD } 3$ where a MOD b calculates the remainder of a/b. In cases where HEAD includes a complete codon, it is included in the MID region. Once the amino acids of the HEAD, MID and TAIL have been identified, tables are consulted for the appropriate multiplication factors (MF) for the specified calculation mode (M, MD, MDT) and for each amino acid or portion thereof. The MFs included within the probe are then multiplied to yield the SC expected within that sequence. This SC value is then compared with that of the worst "candidate" already included in the accumulated table of "candidates" for the least redundant probes. (See Fig.1). This algorithm is both flexible and rapid, requiring approximately one minute of computing time per 100 amino acids of sequence.

ACKNOWLEDGMENTS

This work was supported by NIH grant GM33022 and NSF grant

PCM-8340190 awarded to D.C.W. Inquiries about the program should be addressed to D.C.W.

*To whom correspondence should be addressed

REFERENCES

1. Wu, R. (1972) *Nat. New Biol.* 236, 198-200.
2. Montgomery, D.L., Hall, B.D., Gillam, S. and Smith, M, (1978) *Cell* 14, 673-680.
3. Atkinson, T.C. (1983) *BioTechniques* March/April 1983, 6-10.
4. Williams, J.G. (1981) in *Genetic Engineering*, Williamson, R. Ed., Vol. I, pp. 1-59, Academic Press, New York.
5. Noda, M., Takahashi, H., Tanabe, T., Toyosato, M., Furutani, Y., Hirose, T., Asai, M., Inayama, S., Miyata, T., and Numa, S. (1982) *Nature* 299, 793-797.
6. Houghton, M., Eaton, M.A.W., Stewart, A.G., Smith, J.C., Doel, S.M., Catlin, G.H., Lewis, H.M., Patel, T.P., Emtage, J.S., Carey, N.H., and Porter, A.G. (1980) *Nucleic Acids Res.* 8, 2885-2893.
7. Gray, A., Dull, T.J., and Ullrich, A. (1983) *Nature* 303, 722-725.
8. Babel, W., Wachter, E., Aquila, H., and Klingenberg, M. (1981) *Biochem. Biophys. Acta* 670, 176-180.
9. Ohkubo, H., Kageyama, R., Mjihara, M., Hirose, T., Inayama, S., and Nakanishi, S. (1983) *Proc. Natl. Acad. Sci. (U.S.A.)* 80, 2196-2200.
10. Suggs, S.V., Hirose, T., Miyaka, T., Kawashima, E.H., Johnson, M.J., Itakura, K., Wallace, R.B. (1981) in *Developmental Biology Using Purified Genes*, Brown, D.D. and Fox, C.F., Eds., ICN-UCLA Symp. *Mol. Cell Biol.* 23, 683-693.
11. Patel, D.J., Kozlowski, S.A., Marky, L.A., Rice, J.A., Broka, C., Dallas, J., Itakura, K., Breslauer, K.J. (1982) *Biochem.* 21, 437-444.
12. Gillam, S., Waterman, K., Smith, M. (1975) *Nucleic Acids Res.* 2, 625-634.
13. Agarwal, K.L., Brunstedt, J., Noyes, B.E. (1981) *J. Biol. Chem.* 256, 1023-1028.
14. Uhlenbeck, O.C., Martin, F.H., Doty, P. (1971) *J. Mol. Biol.* 57, 217-229.