BIGPROBE*: a computer program that predicts the sequence of long oligonucleotide probes with high reliability

Mark Dubnick, L. Kevin Lewis and David W. Mount

Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721, USA

Received August 17, 1987; Revised and Accepted October 23, 1987

ABSTRACT

We have written a computer program, BIGPROBE, which facilitates the design of long nucleic acid probes from the partial or complete amino acid sequence of a protein. BIGPROBE relies upon information on codon usage, intercodon dinucleotide frequency, and potential probe self-complementarity. We have examined the accuracy with which the program predicts coding sequences using sample human and rat genes and probe lengths of 30-60 nucleotides. Rat probe sequences selected by BIGPROBE using either codon usage or dinucleotide frequency data alone averaged 86-92% homology with the known exons of the corresponding gene sequences. Predictive accuracy with rat gene probes could be improved to 89-94%, depending upon probe length, by applying codon usage and dinucleotide frequency data in combination. Similar accuracy was achieved for human genes.

INTRODUCTION

Advances in the design, synthesis, and utilization of oligonucleotides as hybridization probes have led to successful isolation of the genes for many proteins whose partial or complete amino acid sequence was known. The vast majority of reports describing isolation of genes from eucaryotic organisms have involved screening cDNA libraries of relatively low complexity (1), but improved strategies have made oligomer-screening of genomic DNA libraries more practical (e.g. 2-6). Two approaches to probe construction have so far achieved the best results. In the first method, a mixed set of short oligonucleotides (usually 11-20 bases in length) is synthesized which contains every possible sequence that can code for a small region of a protein. Under stringent hybridization and wash conditions only one of the sequences within the mixture will form a stable, perfect duplex with the target gene. This strategy has worked well in experiments with cDNA libraries, but has been much less successful with more complex genomic DNA libraries.

The second approach involves synthesis of a single, long oligonucleotide probe. Selection of bases for those amino acids which have multiple codons is usually determined by most or all of the following: (1) analysis of codon usage in the organism and/or tissue of interest; (2) calculation of intercodon dinucleotide frequencies (dinucleotide between base 3 of a codon

to base 1 of the next codon in the gene), especially CpG suppression in mammalian DNA; (3) promotion of potential G:T base-pairing; (4) requirement that any possible contiguous internal base-pairing within the oligonucleotide be minimized. Mismatches between the probe and target gene are expected; however, the large size of the oligonucleotide and its high percentage of homology with the gene can confer enough specificity when used in conjunction with slightly relaxed hybridization conditions.

This latter approach has been used in screening experiments with both cDNA (7-15) and genomic DNA (2-5) libraries. In most studies, the probes used were between 30 and 60 nucleotides in length. Measurement of the accuracy of probe design reveals that the percentage of homology between successful probes and their target genes ranged from 72.2% to 93.1%, the average being 83%. It is worth noting that the lower end of this range is only slightly better than the 67% accuracy expected if the first two unambiguous bases of most codons were predicted correctly, but this result may be somewhat misleading; previous experiments have indicated that small, uninterrupted regions of about 10 or more base-pairs may be required to achieve stable hybrids (6,7,8,10).

Several computer programs have been described to aid in selection of nucleic acid probes based on amino acid sequence (16-18), but they are intended for use with short probe lengths. They make use of knowledge of the redundancy in the genetic code to select stretches of coding sequence with relatively low ambiguity. One of them (18) uses codon usage data to calculate the probability that particular members within a probe mixture are correct.

BIGPROBE is the first program designed for use with the unique, long probe approach to the problem. It is novel in its ability to make use of intercodon dinucleotide data, and to inspect and attempt to correct potential regions of probe self-complementarity. Variation of the relative contributions of codon usage and dinucleotide frequency information to probe selection is possible. In addition, the program will examine the best candidate probe sequences for self-complementarity, make allowable substitutions to alter the regions involved, and re-evaluate the products' scores. This function is extremely important since the local concentration of complementary sequences in the same molecule can be very high, and a probe which has base-paired extensively with itself may never hybridize to its target in the library. Finally, we have shown that the program can, on average, predict rat and human gene probe sequences (30-60 bases long) with greater than 88% accuracy when the probes are designed from complete protein sequences.

ALGORITHM

Although BIGPROBE's use will be demonstrated in a variety of circumstances in the following sections, the description of the algorithm will assume that all of the information the method can use is available. When BIGPROBE is actually running and data such as dinucleotide or codon frequencies cannot be supplied, the portion of the equation they affect is not used.

BIGPROBE scans an ambiguous DNA sequence derived by back-translation of a protein sequence. The ambiguous DNA sequence

should begin at a codon boundary, i.e. the first nucleotide of the first codon is the first character of the back-translated sequence. The program scans the DNA sequence and chooses a unique unambiguous sequence whose bases are determined by codon usage and/or intercodon dinucleotide frequencies for the organism, tissue, or type of gene.

In this analysis, the probability that each of the four nucleotides will be found at a particular position in the predicted sequence is determined for every base position. No calculations are necessary at unambiguously assigned base positions in the back-translated DNA. The base at such positions is assigned probability 1, and the other three are assigned probability 0.

The probability for each possible base at an ambiguous position results from a combination of one or two factors. If the base is first or third in the codon, the probability depends upon both the frequency of the codon which would result from the choice of that base and the dinucleotide frequency for the intercodon dinucleotide which would result. If the base is second in a codon, the probability depends only upon the codon frequency. The dinucleotide frequency comes into play only if the ambiguous position is either the first or third in a codon because the dinucleotide frequency data for the internal dinucleotides (first to second base and second to third base) are completely subsumed by the codon usage frequency data.

For each ambiguous position, the probability P of each of the possible bases occurring at that position is calculated as:

$$P = CF + DI \quad \text{-----} \quad (1)$$

In equation (1), C and D represent the user-input ratio of the relative weights of codon use and dinucleotide frequency in the probability calculation. The value of D is input by the user as a number between 0 and 100% which expresses the fraction used to weight the contribution of dinucleotide frequency. The contribution of codon frequency (C) is calculated as 100% minus the value of D. F, the codon factor is simply the codon frequency as read from a usage table.

When more than one codon has a particular base at the first or third positions, which is true for ARG, LEU, and SER codons, the sum of the two codon frequencies is used. I, the dinucleotide factor, is given by:

$$I = W / (W + X + \dots + Z) \quad \text{-----} \quad (2)$$

where W is the frequency for the dinucleotide resulting from the base chosen and W through Z are the frequencies for all of the dinucleotides which are possible, depending on the substitutions allowable for the ambiguous base at this position. These frequencies are read from the intercodon dinucleotide frequency table.

For example, suppose that we are trying to calculate the base probabilities in an asparagine codon. The following codon begins with a 'C'. The ASN codon would be represented as 'AAY' in the back-translated sequence. The 'Y' symbol represents a pyrimidine. Obviously, the probability that the first base in the codon is 'A' is 1, and that it is one of the other bases is

0. The same applies for the second base. The third base might be either 'C' or 'T'. The probability that it is 'C', $P(C)$, is obtained from the probability equation (1) above:

$$P(C) = C_f + D_i \text{ ----- (3)}$$

where f is the codon frequency of 'AAC'. The dinucleotide factor, i , for 'CC' in this case, is given by:

$$i = w / (w + z) \text{ ----- (4)}$$

where w is the frequency of 'CC' and z is the frequency of 'TC' from the dinucleotide frequency table. Similarly, the probability that the third base is 'T', $P(T)$, is:

$$P(T) = C_g + D_j \text{ ----- (5)}$$

where g is the codon frequency of 'AAT' and j , the dinucleotide factor for 'TC', is defined by:

$$j = z / (w + z) \text{ ----- (6)}$$

A procedure of the type described above is performed for each ambiguous base for the entire back-translated sequence. In order to analyze all probes of the desired length, subsequences of that length are scored by summing the probabilities of all the bases in each. The subsequence scores correspond to the number of bases we expect to assign correctly, on average, in that particular probe. They are sorted in descending order, and the desired number of the top-scoring probe sequences are reported to the user.

Each of the top-scoring probe candidates is inspected for internal self-complementarity. G-T base pairing may be considered at the user's option. When a region of the sequence is found to be complementary to another region, the number of bases with perfect match are counted. If this number exceeds the maximum to be tolerated, then a recursive procedure, called Resolve, attempts to make allowable substitutions at the sites of ambiguous bases to create mismatches in the stem which the probe molecule might form. Resolve computes the most probable non self-complementary probe sequence. If self-complementarity cannot be avoided, Resolve computes instead the sequence with the shortest possible self-complementary regions. Resolve stops when either the most probable non-self-complementary sequence is found, or when the most probable configuration with the shortest potential stem is identified. When one of these states is achieved, the candidate probe sequence's score is adjusted to reflect the change in probability that the substitutions have effected. If the probe sequence still retains significant self-complementarity, the user will be notified of that fact in the final output. Since the original score was optimal, such changes always decrease the score of a candidate probe. The changes are stored for each probe candidate so that the global predicted sequence remains optimal.

After all of the probe candidates are checked for self-complementarity and adjusted as necessary, the scores of all of the probe subsequences are re-sorted. The changes required by

considering self-complementarity have usually decreased the scores of some candidates to the extent that some new probe candidates replace them in the top-scoring group. The checking, adjusting, and sorting are repeated until a group of probe candidates which have the top probability scores, and either have no self-complementarity, or have had such a condition reduced as far as possible, is ready. Then, the candidates are reported to the user. Finally, the entire original and optimal predicted sequences are listed.

PROGRAM DESCRIPTION

The BIGPROBE program runs under MS DOS on an IBM PC or compatible computer. The program is written in the C programming language, and should be easily ported to any computer system with a C compiler. On the PC, BIGPROBE was compiled with the Lattice C compiler.

When running BIGPROBE, the user is first prompted for the input sequence file name. The sequence can be in the form of a single-letter amino acid sequence, or a back-translated ambiguous DNA sequence. The ambiguous base symbols recognized are those recommended by the Nomenclature Committee of the International Union of Biochemistry, as extended for the DNA and Protein Sequence Analysis Programs (DM programs) available from this laboratory (19). File format for the sequence input files is the same as that used by the DM programs as well as the Pearson/Lipman programs and others. In brief, any line beginning with ';' or '>' is interpreted as a comment and ignored. Sequence lines can be of any length, and can contain spaces, tabs, and other white space characters. If the sequence is that of a nucleic acid, the character 'l' should follow the sequence, indicating a linear nucleic acid. If the character '*' follows the sequence, it is interpreted as an amino acid sequence. If no signal character follows the sequence, the user is asked to clarify the situation. If the sequence is of amino acids, the program back-translates this sequence.

The user is then prompted for the name of a codon usage frequency table and, optionally, a dinucleotide frequency table. The formats of these tables are similar to that of the sequence files. A ';' or '>' in the first column of a line indicates a comment. The codons or dinucleotides are placed one per line, in any order. The letters of the dinucleotide or codon are situated in the first two or three columns of the line, as appropriate, in upper or lower case. The frequency is expressed as a decimal number on the same line, and is separated from the codon or dinucleotide characters by one or more spaces. Codon frequencies are represented as numbers between 0 and 1 and express the frequency of using that particular codon to code for its amino acid. Dinucleotide frequencies are represented as decimal numbers between 0 and 1. They express the overall frequency at which the dinucleotide occurs between base three of each codon and base one of the next codon.

After the files described above are successfully read, the program asks the user a few questions about parameters for this run. These are: the weight to give the dinucleotide factor (only asked if using an intercodon dinucleotide frequency table), the length of desired probes, whether to check probes for self-

complementarity, and the number of top-scoring probes to report. If the self-complementarity option is in use, the program queries for the maximum run of base pairing to be tolerated, and whether G-T base pairing is allowed. Finally, BIGPROBE asks for an output file name.

Data output goes to the terminal screen and a user-specified disk file. The output consists of a ranked list of probe sequences, their location in the entire sequence, and score. This information is first related for those probes predicted before consideration of self-complementarity, and then for the probes after such consideration. If any of the probe sequences in the second listing still retain significant regions of self-complementarity, an asterisk appears in the first column of the line. Furthermore, the bases so affected have '+' printed below them in the sequence listing. The final part of the output is a listing of the entire back-translated sequence and the entire predicted sequence.

Finding Areas of Minimum Ambiguity

In some instances, it is desirable to synthesize all possible combinations of coding sequences as a mixed population of probe molecules. This procedure is usually done for short to medium probe lengths, so that the relative amount of the species which matches the target sequence exactly is still a reasonable fraction of the population. In order to use this approach most effectively, BIGPROBE can be used to scan the ambiguous sequence back-translated from the amino acid sequence for regions of minimum ambiguity. After starting BIGPROBE, the default choice is made at the 'Codon frequency file name' prompt. The default selects a file named default.cdn, which contains a codon usage table that gives equal weight to every codon coding for a given amino acid. At this point, a special prompt will ask, "Show probe sequences in ambiguous code?" If 'Y' is answered, the output will consist of a list of probe sequences of the desired length, with their locations in the entire back-translated gene sequence. These probes will be ranked by their degree of ambiguity, based only on the redundancy inherent in the genetic code. The specific sequence given for each probe will be in the ambiguous base code. This sequence can then be used as the input for a DNA synthesizer, setting the equimolar mixes of bases in the places of their corresponding ambiguous base characters.

Designing probes based on codon usage

When designing long nucleic acid probes, the chances of selecting a sequence which is sufficiently similar to the target to allow detection can be increased by using a table of codon usage data. This data tends to be genome specific (20), and is readily available in the literature (21). In order to be used by BIGPROBE, the codon usage data must be entered in a text file format, as specified above.

To use BIGPROBE in this way, 'BP' is typed to invoke the program. Prompts for input sequence file name and name of the codon usage table file name are then requested. A 'n' response is given to the dinucleotide table question and the probe length is entered. The next prompt asks for the probe sequences to be checked for self-complementarity. If 'Y' is answered, two questions are asked. First, what is the longest stretch of perfect self-complementarity that can be tolerated under

hybridization conditions being used? If the probe folds back on itself and forms a stable stem loop structure, its bases will not be available to anneal to its target molecule in the library. The effect is exacerbated by the fact that the local concentration of complementary sequences within the same molecule is far greater than that of complementary target sequences present at low concentration in a genomic or cDNA library. Of course, short regions of self-complementarity, consisting of just a few bases, will continually form, disassociate, and reform in the hybridization mixture. Such transient stem structures should not compromise the effectiveness of the probe. The default choice for the limit of allowable stem length is 5. The second question concerns the allowance of G-T base-pairs under the hybridization conditions being used. The default is not to allow such base pairing. The next question sets the number of probe candidate sequences to be reported. If several probes are desired from an amino acid sequence, it is a good idea to choose those probes which do not overlap.

BIGPROBE often identifies several probes which overlap each other in a region of low ambiguity, so the choice of number of probes to report should be large enough to assure that probes arise from several regions which do not overlap. Finally, a request is made for an output filename.

Designing probes based on codon usage and intercodon dinucleotide frequency

Several statistical studies have established that dinucleotide frequencies in mammalian coding DNA sequences are not random (20,22). Consideration of these data in conjunction with codon usage data may well result in a more accurate prediction of the sequence coding for a polypeptide (6, and our Results). BIGPROBE permits total control over the method used to combine these two factors. The program is used in the same way as for using codon usage data, including providing the name of the file containing the codon table. If 'Y' is answered to the next question, concerning the use of a dinucleotide frequency table, the user is prompted for the name of the file with the table. The format of such a file is explained above. Then, one is asked for a dinucleotide weight factor which is entered as a number between 0 and 100 percent. If 0 is entered, the analysis gives no weight to the dinucleotide factor, just as if no dinucleotide table is used. If 50% is entered, equal weight will be given to codon frequency and dinucleotide frequency. Finally, if 100% is entered, only the dinucleotide factor will be used, and the codon usage table will not be used at all. The combining of codon usage and dinucleotide frequency is only used on ambiguous bases which occur at positions 1 or 3 in a codon. Of course, bases which are not ambiguous are not affected by either. The rest of the prompts for using BIGPROBE in this mode are identical to those used with a codon usage table, as detailed above.

Method of Testing

We tested the performance of the empirical methods described above with a variety of parameters using complete DNA coding sequences extracted from GenBank, release 44.0. We examined sequences from ten rat and ten human proteins. Human genes used were HUMALBG, HUMMHDR5B, HUMESTR, HUMP53T, HUMGFIB, HUMP971, HUMKEREP, HUMERPA, HUMC3, and HUMAPOE4. Rat genes were RATMLC2,

RATPHH, RATHDP, RATPKL, RATCATL, RATMT12C, RATGLTPG, RATUCP, and RATCYPOXM. A table of mammalian intercodon dinucleotide frequencies (22) was used for both human and rat sequences. A human codon usage table was used (23), and a rat table was calculated from published data (21).

Coding sequences were translated into amino acid sequences using the DM5 Sequence Analysis Programs from this laboratory, version 5.06. These protein sequences were used as input to BIGPROBE and the output oligonucleotides were compared to their counterparts in the genes. Data were collected for the probe sequence which BIGPROBE scored as most probable for each gene and for the average of the top ten probes selected by BIGPROBE. Tests were performed using oligonucleotide lengths of 30, 40, 50, and 60 bases to predict rat gene sequences at dinucleotide/codon weights of 0%/100%, 20%/80%, 40%/60%, 50%/50%, 60%/40%, 80%/20%, and 100%/0%. Results for the ten rat genes were averaged, as were results for the ten human sequences.

RESULTS

The ability of BIGPROBE to select optimum probe sequences was tested using 10 human and 10 rat genes obtained from Genbank. Oligonucleotide lengths of 30, 40, 50, and 60 bases were evaluated. In all experiments, potential self-complementarity was limited to stems of five base-pairs and G:T base-pairing was not considered. Probes were characterized by calculating the percentage of bases which matched those bases in the actual gene sequence. Those probes receiving the best scores by the program and the ten highest scoring probes for each gene were analyzed for accuracy and results were averaged.

Results obtained using either codon usage alone (Table 1) or intercodon mammalian dinucleotide frequencies alone (Table 2) indicate that these parameters are sufficient to predict rat and

RAT				
Probe size:	30	40	50	60
Mean top score(n=10):	89.3	89.3	89.0	85.7
Mean top 10 score(n=100):	89.7	88.1	88.6	87.1
HUMAN				
Mean top score(n=10):	90.0	90.8	90.8	89.2
Mean top 10 score(n=100):	90.6	90.1	89.1	88.3

Table 1. Evaluation of probe sequences selected by BIGPROBE from amino acid sequences. Bases at ambiguous positions were assigned from tables of codon usage. The database consists of human and rat coding sequences extracted from Genbank. n is the number of probes generated. The values shown reflect the percentage of bases within each probe which are identical to those bases in the exons of the sequenced gene.

RAT				
Probe size:	30	40	50	60
Mean top score(n=10):	92.3	89.3	89.0	87.2
Mean top 10 score(n=100):	91.1	89.1	89.2	86.4
HUMAN				
Mean top score(n=10):	89.3	86.0	85.8	85.8
Mean top 10 score(n=100):	89.2	85.9	85.4	84.8

Table 2. Evaluation of probe sequences selected by BIGPROBE using intercodon dinucleotide frequency data to assign ambiguous bases. Data is scored as in Table 1.

human coding sequences with high accuracy. Homologies of 85.7-92.3% were achieved with top probes selected by BIGPROBE and mean top 10 scores ranged from 84.8-91.1%. Shorter probes generally yielded best results, regardless of the selection criteria employed (compare especially 30 vs. 60 bases). In addition, the ten best probes selected by BIGPROBE for each rat or human gene usually contained as many correct bases as the top probe selected for each gene. Furthermore, although rat probe accuracy did not differ greatly under the two conditions used, human probe selection was clearly better when derived solely from

Codon Usage (%):	100	80	60	50	40	20	0
Dinuc. Freq.(%):	0	20	40	50	60	80	100
Probe size=30							
Mean top score:	89.3	89.7	92.0	93.7	93.7	92.3	92.3
Mean top 10 score:	89.7	89.8	91.0	90.4	91.4	90.5	91.1
Probe size=40							
Mean top score:	89.3	90.0	91.0	91.3	92.0	90.1	89.3
Mean top 10 score:	88.1	89.1	90.1	91.0	90.7	89.8	89.1
Probe size=50							
Mean top score:	89.0	90.0	91.2	91.8	91.8	90.6	89.0
Mean top 10 score:	88.6	88.7	89.7	90.1	90.2	89.7	89.2
Probe size=60							
Mean top score:	85.7	88.0	88.7	89.0	88.5	87.2	87.2
Mean top 10 score:	87.1	87.2	88.4	88.6	88.2	87.0	86.4

Table 3. Effect of varying weighted scores assigned to codon usage data and intercodon dinucleotide frequencies on probe accuracy. The sample set consists of the 10 rat genes used for compilation of data in Tables 1 and 2. Data is scored as in Table 1.

codon usage data. This is most evident with 40, 50, and 60mers.

An important function of BIGPROBE is its ability to combine different sources of information used to predict coding sequences. This capability is exemplified by the experiments depicted in Table 3. The ten rat genes were re-examined under conditions in which the contributions of codon usage and dinucleotide frequency were varied over a wide range. The ratio of relative weights given to these parameters are indicated as percentages, e.g. codon usage/dinucleotide frequency ratios of 100%/0%, 80%/20%, etc. The values within the table indicate that application of approximately equal weights to the two parameters produces better results than using either one alone (compare 100/0, 50/50, and 0/100). This result was obtained regardless of the size of the oligomer selected.

At the preset stem limit of 5 base-pairs used in these experiments, adjustments due to self-complementarity of the initial top 10 probes selected for each gene became common only with lengths of 50 and 60 bases. In all cases, the accuracy of prediction of initial and adjusted top 10 scores differed by less than 1% (data not shown).

DISCUSSION

We have developed a computerized method for predicting a coding DNA sequence from the amino acid sequence of a protein. We have implemented the method in a computer program, BIGPROBE, which runs on the IBM PC family of computers and is widely transportable. BIGPROBE can make use of a number of types of information to optimize sequence prediction for long oligonucleotide probes.

BIGPROBE gives an investigator great flexibility in using different types of information to design nucleic acid probes for screening cloned gene libraries. The program can make use of codon usage data to maximize the likelihood of predicting the sequence of a target gene correctly. Codon usage data for a number of species is easily available in the literature (21,23), and is generally genome specific (20). After predicting the entire coding sequence of a protein, the program produces a list of candidate probe sequences which are ranked in order of their probability of matching the target sequence.

BIGPROBE can also make use of intercodon dinucleotide frequency information. Such data is available in the literature, and has been shown to differ significantly from random in coding DNA (20,22,24). Moreover, the program allows the user to specify the relative weights used to make predictions based on codon usage and dinucleotide frequency data.

At the user's option, BIGPROBE will check the predicted candidate probe sequences for significant self-complementarity. If such self-complementarity is found, the algorithm attempts to resolve this by substituting allowable replacements for the ambiguous bases in the regions involved. Probe self-complementarity must be considered in order to achieve optimal hybridization because if a probe molecule forms a stable stem loop configuration with itself, the bases involved will not be available to anneal with its target in the cloned library. BIGPROBE is unique in its ability to consider and adjust probe

sequences based on codon usage, dinucleotide frequency, and potential self-complementarity.

The ability of BIGPROBE to predict correctly rat and human coding sequences using either codon usage, dinucleotide frequency, or a combination of these (Tables 1-3) has been examined for probes containing 30-60 bases. This sequence range was chosen based on probe lengths used in past experiments (2-15). Human probe sequences selected as best for each gene by the program were 89.2-90.8% correct when human codon usage was used alone and this range could not be improved by using either dinucleotide frequencies alone or by using combinations of these parameters. Rat probe sequences were predicted most accurately when approximately equal weights were given to contributions of codon usage and dinucleotide frequency (89.0%-93.7%). It is worth noting that, for each rat or human gene, the accuracy achieved among the 10 best probes was very similar.

The values obtained here compare quite well with those found to be experimentally acceptable in past hybridization studies. Interestingly, in subsequent analyses of the ability of BIGPROBE to predict the sequences of entire genes, 80% identity or better was usually obtained. Improvements in predictive capability will require refinements in our knowledge of codon usage and other sequence, structural, and/or functional constraints. We are currently assessing the use of modified codon usage tables which reflect the influence of additional factors (e.g. adjacent codons; see (6)).

ACKNOWLEDGEMENTS

We thank Margaret M. Cavenagh for data compilation and bibliographic assistance. M. D. was supported by postdoctoral fellowships from the American Cancer Society and the Leukemia Society of America during the period this work was done. K. L. was supported by a grant from the American Diabetes Association. D. M. is supported by grant no. GM24496 from the National Institutes of Health.

*Program availability: BIGPROBE is available for a nominal charge as an executable program on an IBM compatible diskette from the Genetics Software Center, Biosciences West 308, Dept. of Mol. and Cell. Biol., Univ. of Arizona, Tucson, Az., 85721.

REFERENCES

1. Itakura, K., Rossi, J.J. and Wallace, R.B. (1984) *Ann. Rev. Biochem.* **53**, 323-356.
2. Wood, W.I., Capon, D.J., Simonsen, C.C., Eaton, D.L., Gitschier, J., Keyt, B., Seeburg, P.H., Smith, D.H., Hollingshead, P., Wion, K.L., Delwart, E., Tuddenham, E.G., Vehar, G.A. and Lawn, R.M. (1984) *Nature* **312** 330-337.
3. Anderson, S. and Kingston, I.B. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 6838-6842.
4. Ullrich, A., Berman, C.H., Dull, T.J., Gray, A. and Lee, J.M. (1984) *EMBO Journal* **3**(2), 361-364.

5. Linnenbach, A.J., Speicher, D.W., Marchesi, V.T. and Forget, B.G. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2397-2401.
6. Lathe, R. (1985) *J. Mol. Biol.* **183**, 1-12.
7. Parker, P.J., Coussens, L., Totty, N., Rhee, L., Young, S., Chen, E., Stabel, S., Waterfield, M.D. and Ullrich, A. (1986) *Science* **233**, 853-859.
8. Coussens, L., Parker, P.J., Rhee, L., Yang-Feng, T.L., Chen, E., Waterfield, M.D., Francke, U. and Ullrich, A. (1986) *Science* **233**, 859-866.
9. Derynck, R., Jarrett, J.A., Chen, E.Y., Eaton, D.H., Bell, J.R., Assoian, R.K., Roberts, A.B., Sporn, M.B., and Goeddel, D.V. (1985) *Nature* **316**, 701-705.
10. Derynck, R., Roberts, A.B., Winkler, M.E., Chen, E.Y. and Goeddel, D.V. (1984) *Cell* **38**, 287-297.
11. Ullrich, A., Bell, J.R., Chen, E.Y., Herrera, R., Petruzzelli, L.M., Dull, T.J., Gray, A., Coussens, L., Liao, Y.-C., Tsubokawa, M., Mason, A., Seeburg, P.H., Grunfeld, C., Rosen, O.M. and Ramachandran, J. (1985) *Nature* **313**, 756-761.
12. McLean, J., Fielding, C., Drayna, D., Dieplinger, H., Baer, B., Kohr, W., Henzel, W., and Lawn, R. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2335-2339.
13. Kubo, T., Fukuda, K., Mikami, A., Maeda, A., Takahashi, H., Mishina, M., Haga, T., Haga, K., Ichiyama, A., Kangawa, K., Kojima, M., Matsuo, H., Hirose, T. and Numa, S. (1986) *Nature* **323** 411-416.
14. Toole, J.J., Knopf, J.L., Wozney, J.M., Sultzman, L.A., Buecker, J.L., Pittman, D.D., Kaufman, R.J., Brown, E., Shoemaker, C., Orr, E.C., Amphlett, G.W., Foster, W.B., Coe, M.L., Knutson, G.J., Fass, D.N. and Hewick, R.M. (1984) *Nature* **312**, 342-347.
15. Jaye, M., de la Salle, H., Schamber, F., Balland, A., Kohli, V., Findeli, A., Tolstoshev, P. and Lecocq, J.-P. (1983) *Nucl. Acids Res.* **11**(8), 2325-2335.
16. Yang, J., Ye, J. and Wallace, D. C. (1984) *Nucl. Acids Res.* **12**(1), 837-843.
17. Lewis, R. M., (1986) *Nucl. Acids Res.* **14**(1), 567-570.
18. Raupach, R. E., (1984) *Nucl. Acids Res.* **12**(1), 833-836.
19. Mount, D.W. and Conrad, B. (1986) *Nucl. Acids Res.* **14**(1), 443-454.
20. Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) *Nucl. Acids Res.* **11**(7), 2205-2219.
21. Maruyama, T., Gojobori, T., Aota, S. and Ikemura, T. (1986) *Nucl. Acids Res.* **14**(supplement) r151-r197.
22. Santibanez-Koref, M. and Reich, J. G. (1986) *Biomed. Biochim. Acta* **45** (6), 737-748.
23. Chen, H.R. and Barker, W.C. (1985) *Trends in Genetics* 221-223
24. Santibanez-Koref, M. and Reich, J. G. (1986) *Biomed. Biochim. Acta* **45**(9), 1105-1109.