Characterization of the distribution of potential short restriction fragments in nucleic acid sequence databases

Implications for an alternative to chemical synthesis of oligonucleotides

Chang-Shung Tung*+ and Christian Burks*°

*Theoretical Biology and Biophysics Group and +Experimental Pathology Group, MS K710, Los Alamos National Laboratory, University of California, Los Alamos, NM 87545, USA

Received 23 June 1986

We have searched the GenBank nucleic acid sequence database for potential short restriction fragments. All possible oligonucleotides up to length five are found at least once flanked by known restriction recognition patterns. Thus, searches in the database for a specific sequence corresponding to a desired oligonucleotide would often point to one or more sources of short, retrievable fragments containing that sequence. These results underscore the potential of nucleic acid sequence databases in planning experiments.

Nucleic acid sequence database *Oligonucleotide synthesis* *Pattern recognition* *Restriction endonuclease*

1. INTRODUCTION

As available nucleic acid sequence databases (such as those of EMBL [1] and GenBank [2]) have grown, so has their usefulness to molecular biologists. Currently, the most frequent use of the databases involves querying the established data with a new sequence to determine whether or not identical or similar sequences are already known [3,4]; a related use involves extracting a small number of known and presumably similar sequences and aligning them. Both of these activities are primarily post-experimental, representing a final analysis of data.

We wish to draw attention to the potential use of nucleic acid databases as pre-experimental tools. In particular, we propose that a researcher desiring one or more short restriction fragments could examine the database (or a list compiled from it) for evidence of such a fragment and then isolate the fragment from the clone that was originally used for sequencing. Such an approach is supported by our finding that the known nucleic acid sequence data are rich in potential short restriction fragments.

2. METHODS

We searched the November 1985 release of the GenBank database [2] for all oligonucleotides up to length 20 that were flanked by representatives of a set of known restriction endonuclease recognition patterns [5,6]. Mutually complementary oligonucleotides were catalogued as a single oligonucleotide species. All oligonucleotides were catalogued independently of the restriction recognition patterns they were flanked by. For example, consider the two sequences, 'CCGG/CGC/GGATCC' and 'GCGCGC/GCG/TGATCA', each consisting of an internal oligonucleotide flanked on both sides by restriction recognition patterns. These would have been catalogued as a single species, 'CGC'. We used a reduced set of 68 recognition patterns, listed in table 1, consisting of

° To whom correspondence should be addressed

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/86/$3.50 © 1986 Federation of European Biochemical Societies

unique patterns that were associated with an endonuclease site coincident with the span of the pattern. Phrased in terms of pattern recognition, we searched for patterns defined not by alphabetic similarity but by the known ability of the corresponding nucleotide sequence to specify interaction with a set of functionally related enzymes; when any two of these patterns were separated by a string of less than 21 characters, we alphabetically catalogued the separating string.

Table 1

Reduced set of restriction recognition patterns

Pattern[a]	Enzyme[b]	Pattern[a]	Enzyme[b]
GACGTC	*Aat*II	CTGCTG	*Eco*P15
GTMKAC	*Acc*I	AGACC	*Eco*PI
CGCG	*Acc*II	GGTCT	*Eco*PI
GRCGYC	*Acy*I	RRATYY	*Eco*RI′
CTTAAG	*Afl*II	GCNGC	*Fnu*4HI
ACRYGT	*Afl*III	GANTC	*Fnu*AI
TTTAAA	*Aha*III	CTCCAG	*Gsb*I
AGCT	*Alu*I	CTGGAG	*Gsb*I
TCGCGA	*Ama*I	WGGCCW	*Hae*I
TGCGCA	*Aos*I	RGCGCY	*Hae*II
CYCGRG	*Aqu*I	GACGC	*Hga*I
GGTNACC	*Asp*AI	GCGTC	*Hga*I
GGNCC	*Asu*I	GTCGAC	*Hgi*CIII
TTCGAA	*Asu*II	CTAG	*Mae*I
TGATCA	*Atu*CI	ACGT	*Mae*II
ATGCAT	*Ava*III	GTNAC	*Mae*III
CCTAGG	*Avr*II	CGATCG	*Nbl*I
ATCGAT	*Ban*III	CCATGG	*Nco*I
CTGCAG	*Bce*170	CATATG	*Nde*I
AAGCTT	*Bbr*I	CATG	*Nla*III
TGGCCA	*Bal*I	GGNNCC	*Nla*IV
GGCC	*Blu*II	GCGGCCGC	*Not*I
GATC	*Bsa*PI	RCATGY	*Nsp*(7524)I
GCGCGC	*Bse*PI	CMGCKG	*Nsp*BII
GDGCHC	*Bsp*1286	GTAC	*Rsa*I
CCGG	*Bsu*1192I	AGTACT	*Sca*I
GCGC	*Cfo*I	CCNGG	*Scr*FI
RCCGGY	*Cfr*10I	GCATC	*Sfa*NI
YGGCCR	*Cfr*14I	GATGC	*Sfa*NI
GTYRAC	*Chu*II	TACGTA	*Sna*BI
CCTNAGG	*Cvn*I	TCGA	*Taq*I
CTNAG	*Dde*I	CAARCA	*Tth*111II
GATATC	*Eco*32I	TGYTTG	*Tth*111II
CAGCAG	*Eco*P15	TCTAGA	*Xba*I

[a] The single-letter code for designating alternatives at a single position is that recommended by IUPAC [7]; in those cases where inverting and complementing a given pattern did not result in the identical pattern, the inverted complement was treated as a separate pattern

[b] This reduced list was compiled from Roberts [8]; the original list was reduced by excluding (i) patterns identical to or subsumed by other patterns and (ii) patterns not coincident with the associated cleavage site

The search and sorting routines (EXTCT, DIVD, COMB) were written in C and run on a VAX 11/780 under the UNIX operating system. The 12 divisions of the database were searched separately to keep the sizes of the intermediate output files manageable. Finding and cataloguing the oligonucleotides in the database, which contained over 5.5 × 10^6 nucleotides, took roughly 100 h of VAX time.

3. RESULTS AND DISCUSSION

The total number of found distinct oligonucleotide species (as defined above) is presented as a function of oligonucleotide species length in fig.1. These data are compared in fig.1 with the total possible number of distinct oligonucleotide species, $T(l)$, as a function of oligonucleotide species length, l,

$$T(l) = 2^{(l-1)} \times (2^l + \delta) \qquad (1)$$

where $\delta = 0$ for odd l and $\delta = 1$ for even l. Up to length 5, all possible distinct oligonucleotide species were found; for lengths 6 and 7, greater than 85% of all possible species were found. For longer species, although the total number found remains relatively constant (rising from $10^{4.17}$ to $10^{4.29}$ and then dropping to $10^{4.14}$), the percentage of possible species found rapidly diminishes, dropping below 1% for lengths greater than 10. However, as the sequence data and quantity of known distinct restriction patterns continue to grow, so will the percentage values for the longer oligonucleotide species. The sequence data are currently growing at close to two million nucleotides per year [2], and known restriction recognition specificities at about ten per year [5,8].

We also preserved a list of the oligonucleotides with their restriction pattern flanks and an indication of the GenBank entry in which they were found (these data are available on request). This list contains, as an example, the sequence 'GGAGTTCGAGACCAG' nested in a potential

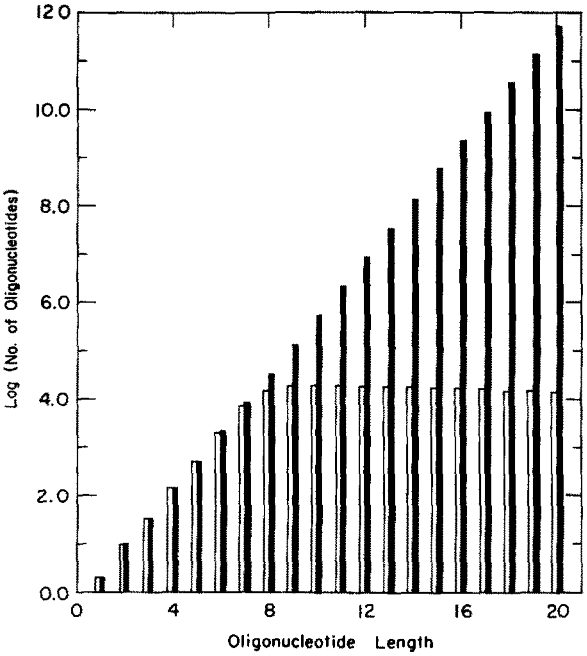

Fig.1. Number of oligonucleotide species as a function of oligonucleotide species length expressed as number of nucleotides. Empty bar, number of distinct oligonucleotide species (as defined in the text) found in GenBank; shaded bar, number of possible distinct oligonucleotide species (see eqn 1).

short restriction fragment (flanked by *Dde*I and *Scr*FI patterns) within the human gastrin gene sequenced by Ito et al. [9] and presented in GenBank entry HUMGAST2 (EMBL/GenBank accession no. K01254). The duplex oligonucleotide corresponding to this sequence was synthesized by Murphy and Baralle [10] for use in a study of box B variants.

Short oligonucleotides are being used increasingly [11,12] in the isolation, sequence determination, and characterization of DNA. This includes their use as primers for sequencing of DNA and RNA, as hybridization probes, as building blocks in the construction of synthetic genes, as elements for directed mutational analysis, and as the objects of structural studies in crystals and in solution. The ability to synthesize chemically these oligonucleotides [11] has kept apace of the demand for them and has culminated in so-called 'gene machines' that carry out completely automated syntheses [12].

One can envision circumstances – such as particular requirements for quantity and purity, or lack of access to a gene machine – where acquiring, isolating, and amplifying a clone indicated by a database search would provide a reasonable alternative to chemical synthesis. Furthermore, we anticipate both that the preservation of clones will continue to become more standard and centralized, and that the technology of amplification and isolation of DNA fragments will move further towards automation: developments such as these would also support the alternative suggested here. Finally, we note that although the initial search for all potential short restriction fragments in a large database is computationally intensive (as described above), a subsequent search for an individual sequence in the output list is relatively quick. With a minimal allocation of resources, a potentially valuable alternative source of the desired oligonucleotide might be identified.

ACKNOWLEDGEMENTS

We are grateful for enlightening conversations with G. Dean, J. Fickett, and R. Moyzis; for suggestions about the manuscript from J. Fickett and R. Roberts; and for the graphical expertise of B. Atencio. This work was done under the auspices of the US Department of Energy. C.-S.T. was also supported by a grant (GM26857) from the National Institutes of Health to G.C. Salzman.

REFERENCES

[1] Hamm, G.H. and Cameron, G.N. (1986) Nucleic Acids Res. 14, 5–9.
[2] Burks, C., Fickett, J.W., Goad, W.B., Kanehisa, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C.-S. and Bilofsky, H.S. (1985) Comp. Appl. Biosci. 1, 225–233.
[3] Smith, T.F., Waterman, M.S. and Burks, C. (1985) Nucleic Acids Res. 13, 645–656.
[4] Goad, W.B. (1986) Annu. Rev. Biophys. Chem. 15, 79–95.
[5] Roberts, R.J. (1985) Nucleic Acids Res. 13, r165–r200.
[6] Kessler, C., Neumaier, P.S. and Wolf, W. (1985) Gene 33, 1–102.
[7] Cornish-Bowden, A. (1985) Nucleic Acids Res. 13, 3021–3030.

[8] Roberts, R.J. (1984) Nucleic Acids Res. 12, r167–r204.

[9] Ito, R., Sato, K., Helmer, T., Jay, G. and Agarwal, K. (1984) Proc. Natl. Acad. Sci. USA 81, 4662–4666.

[10] Murphy, M.H. and Baralle, F.E. (1984) J. Biol. Chem. 259, 10208–10211.

[11] Itakura, K., Rossi, R.B. and Wallace, R.B. (1984) Annu. Rev. Biochem. 53, 323–356.

[12] Caruthers, M.H. (1985) Science 230, 281–285.