Volume 13 Number 9 1985 Nucleic Acids Research

Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984

Athel Cornish-Bowden*

Department of Biochemistry, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, UK

1. INTRODUCTION

With the introduction of methods of rapid nucleic acid sequence determination, synthesis of mixed oligonucleotide probes and computer-assisted analysis of nucleic acid sequences, the use of a single symbol to designate a variety of possible nucleotides at a single position has become widespread over the last few years. Whereas the use of, for example, the symbols R and Y to designate purine (A or G) and pyrimidine (C or T) ribonucleotides respectively [1] is generally accepted, no agreed symbols exist for the other possible combinations. Indeed, a plethora of diverse systems have proliferated in the last few years [2-11]. It is striking that, in one extreme case, the combination (C or G) has been represented by at least five different symbols [2,3,4,8,11]. A standardized set of symbols is thus required to prevent confusion.

The symbols are intended to be applicable to both deoxyribonucleic and ribonucleic acids. Thus it is important to note from the outset that the recommended symbols will not discriminate between DNA and RNA, and the symbol T will be employed at all positions where U might appear in the RNA. Similarly, no distinction will be made in the symbols between base, nucleoside and nucleotide. Sequences may be assumed to have a deoxyribose backbone (DNA) unless specified otherwise. These changes from earlier recommendations [1] reflect great advances in techniques for sequencing DNA, so that RNA sequences are now commonly deduced from the corresponding DNA sequences. Since the standard representation of a DNA sequence may be converted to the corresponding RNA sequence by the simple expedient of

*As secretary of the Nomenclature Committee of the International Union of Biochemistry, which thanks a panel convened by Dr Richard Lathe for drafting these recommendations. The members of the committee and of the panel are given in the Acknowledgements.

substituting T by U, it is not envisaged that data banks based on computer storage facilities will inevitably contain entries for both DNA and its RNA equivalent. Authors should always, however, make it clear which strand of DNA or RNA a given sequence refers to, and in circumstances where confusion between DNA and RNA is likely the sequence may be prefixed with the lower-case letter d or r, as in the previous recommendations [1].

As the present recommendations present unique alphabetic symbols for each nucleotide combination, the use of upper- and lower-case letters as equivalents does not lead to confusion. However, such use may cause confusion between r (ribo-) and R (purine), and care must be taken in those rare cases where the various symbols are used in combination. In general, it should be emphasised (i) that upper-case symbols are advocated, and (ii) that the present recommendations are not intended to prejudice any possible future use of contrasting upper- and lower-case letters for specific purposes.

It was previously [1] recommended that hyphens should be used to represent 5'-3' phosphodiester linkages in known nucleotide sequences. As there is now little danger of confusion between codon triplets and nucleotide sequences this recommendation is no longer considered necessary. Hyphens may therefore be omitted from sequences, and are omitted from all sequences in this document. In addition it may be assumed that all sequences are presented 5' to 3' unless otherwise specified, although specific mention of this fact is not discouraged.

Although several diverse systems of symbols for incompletely specified bases already exist in the literature, this presentation makes no systematic review. Details of the previous recommendations may be found in ref. 1, and of systems that have been used in the literature in refs. 2-11.

2. APPLICATIONS OF A STANDARD NOMENCLATURE

2.1 Recognition sequences in DNA for restriction enzymes

Most restriction enzymes and their corresponding methylases recognise simple unique nucleotide sequences in DNA. For example, EcoRI and BamHI recognise the sequences 5'-GAATTC-3' and 5'-GGATCC-3' respectively. Nevertheless, a growing class of enzymes includes those that recognise series of derivative sequences, where two or more bases may be present at a particular position in the recognition sequence (for a complete listing see ref. 12). For instance, the enzyme AvaI recognises four different sequences 5'-CCCGGG-3', 5'-CCCGAG-3', 5'-CTCGGG-3' and 5'-CTCGAG-3'. The recognition sequence for AvaI may thus be represented as 5'-CYCGRC-3', where Y represents

a pyrimidine and R represents a purine, as recommended previously [1]. However, several newer enzymes recognise combinations that are not covered by the existing symbols. For instance, AccI recognises the sequence 5'-GT(A or C)(G or T)AC-3' [13]. SduI recognises the sequence 5'-G(A or G or T)GC(A or C or T)C-3' [14]. The present symbols are intended to cover these possibilities.

2.2 Recognition sequences in DNA for other enzymes

Restriction enzymes are highly specific for particular nucleic acid sequences. For many other enzymes the specificity is rather more lax, however, and the symbols are intended to meet in part the need for presenting a schematic summary of the sequence features. For instance, sequences recognised by the RNA polymerase of Escherichia coli may be presented as the juxtaposition of two sequences 5'-AA(A or T)NTNNN(C or G)TTGACA-3' and 5'-(T or G)NNTATAAT-3' separated by 13 to 16 nucleotides (adapted from refs. 15, 16), where N represents any nucleotide. A similar treatment may be applied to the recognition sequences for other DNA binding proteins such as repressor molecules.

2.3 Recognition sequences in RNA for enzymes involved in translation

RNA sequences are, as mentioned above, most conveniently represented as their DNA counterparts. Thus the basic elements of a translation initiation site in Escherichia coli may be represented by 5'-(G or A)(G or A)GGGNNNNAN(C or T)ATGNN(A or T)NNNNN(C, T or G) (adapted from ref. 17). Similarly, translation initiation sites in eukaryotic mRNAs tend to conform to the sequence 5'-ANNATG(G or A)-3' [18].

2.4 Codon degeneracy

Although there are 64 possible triplet codons, there are only 20 different amino acids coded by them. Thus most amino acids are inserted into a growing polypeptide chain in response to 2 or more different triplets in the mRNA (ref. 19 for a general review). For example, proline is coded by 5'-CCN-3' and alanine by 5'-GCN-3'. In other cases the pattern may be more complex, such as for isoleucine, which is coded by 5'-AT(T, C or A)-3'. Note that certain amino acids (e.g. serine) may be coded by two distinct groups of triplets [here 5'-TCN-3' and 5'-AG(T or C)-3'], which cannot be adequately represented as 5'-(T or A)(C or G)N-3' (see Table 4). It is to be noted that synthetic oligonucleotide probes for detecting protein-coding sequences often involve the preparation of 'mixed probes'. Here a mixture of two (or more) nucleotides is incorporated at a single position in the oligonucleotide to take account of the redundancy of the genetic code (for instance ref. 20). It

is anticipated that a single-letter code might be used to designate such mixtures.

2.5 Construction of ancestral sequences by parsimony procedures [21]

Where two descendants differ in nucleic acid sequence at a particular position (for instance A in one and G in the other), the putative ancestral sequence can be represented [10] using a single-letter code, in this case R.

2.6 Other uses

The symbols are intended to be useful for all purposes in which the exact identity of a nucleotide may vary. Thus uncertainties encountered with primary nucleic acid sequence data may, in some cases, be represented using standard symbols.

3. ALLOCATION OF SYMBOLS

In the choice of symbols the following considerations have been taken into account: (i) conformity to previous IUPAC-IUB nomenclature [1]; (ii) logical derivation; (iii) ease of memorisation; (iv) availability of symbols on a standard typewriter keyboard; (v) historical precedence.

3.1 Guanine, adenine, thymine, cytosine: G, A, T, C

These one-letter symbols have previously been established [1] and are generally used. There is, however, a problem of discriminating between the upper case letters G and C on poorly copied sequences. Nevertheless, the use of alternative symbols for G (such as a barred-G, ⌀) is not recommended. Discrimination between the lower-case letters is much clearer. Note that T and U may, in general, be considered as being synonyms, though care should be taken to avoid ambiguity in circumstances where it is likely, e.g. in discussing artificial hybrids of DNA and RNA and in cases where specific distinction between T and U is advisable.

3.2 Purine (adenine or guanine): R

R is the symbol previously recommended [1].

3.3 Pyrimidine (thymine or cytosine): Y

Y is the symbol previously recommended [1].

3.4 Adenine or thymine: W

Although several diverse symbols have been used for this pair, (and for the reciprocal pair G+C), only two symbols have a rational basis, L and W: L derives from DNA density (light; G+C — heavy — would thus be H); W derives from the strength of the hydrogen bonding interaction between the base pairs (weak for A+T: G+C — strong — would thus be S). However, the system recommended for the three-base series (not-A = B, etc., see below, section 3.8) rules out H as this would be not-G. W is thus recommended.

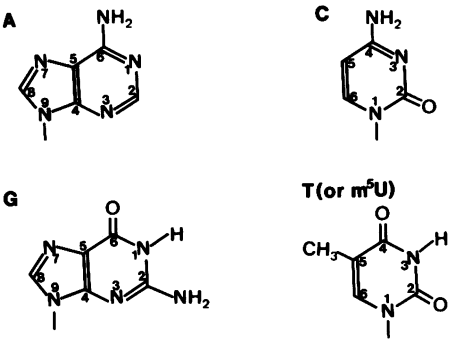

Fig. 1. Origin of the symbols M and K. The four bases are drawn so as to show the relationship between adenine and cytosine on the one hand, which both have aMino groups at the ring position most distant from the point of attachment to the sugar, and between guanine and thymine on the other, which both have Keto groups at the corresponding position. The ring atoms are numbered as recommended [24-26], although for the present purpose this has the disadvantage of giving discordant numbers to the corresponding positions.

3.5 Guanine or cytosine: S

The choice of this symbol is discussed above in section 3.4.

3.6 Adenine or cytosine: M

There are few common features between A and C. The presence of an NH_2 group in similar positions on both bases (Fig. 1) makes possible a logically derived symbol. A and N being ruled out, M (from aMino) is recommended.

3.7 Guanine or thymine: K

By analogy with A and C (section 3.6), both G and T have Keto groups in similar positions (Fig. 1).

3.8 Adenine or thymine or cytosine: H

Not-G is the most simple means of memorising this combination and symbols logically related to G were examined. F and H would both be suitable, as the letters before and after G in the alphabet, but A would have no equivalent to F. The use of H has historical precedence [2].

3.9 Guanine or cytosine or thymine: B

Not-A as above (section 3.8).

3.10 Guanine or adenine or cytosine: V

Not-T by analogy with not-G (section 3.8) would be U but this is ruled out to eliminate confusion with uracil. V is the next logical choice. Note that T and U may in some cases be considered to be synonyms.

3.11 Guanine or adenine or thymine: D

Not-C as above (section 3.8).

3.12 Guanine or adenine or thymine or cytosine: N

This symbol is suggested by the sound of the word "aNy". The use of X to represent an unknown base is acknowledged, but this is not recommended as the symbol refers to xanthine [1]. Occasionally it may be desirable to distinguish between unspecified (N) and unknown (X), but if X is used for this purpose it should be explicitly defined.

4. OTHER ACCESSORY SYMBOLS

There are a number of instances in which additional symbols may be required for routine work. Although this section provides a number of suggestions, these do not form part of the present recommendations.

First, we consider the uncertainty as to whether a base exists at a certain position or not. A symbol denoting "G or A or T or C or no nucleotide", for example ? or +, might be used to define regions of uncertainty of limited variable size in a recognition sequence (see for instance ref. 22). Alternatively, one of these symbols might be used as a modifier to denote uncertainty: ?A might, for instance, denote "A or no nucleotide at this position". Second, the unambiguous absence of a nucleotide introduced into a sequence for alignment or comparison purposes alone could be represented by :, though a simple space has much to recommend itself. Third, a specified number of unknown nucleotides might be represented by a symbol such as = in conjunction with numerals, so that, for example, "=300=" might denote the presence of 300 unknown nucleotides. Fourth, the symbol N (unknown or unspecified) may be replaced by the hyphen - in circumstances where rapid visual discrimination between "known" (essential) and "unknown" (non-essential) sequences is desirable. The value of this may be judged by comparing "NNNNNCNNGNTNN" with "-----C--G-T--", for example. Note that the use of the lower-case letter n may avoid the necessity for an additional symbol, as in "nnnnnCnnGnTnn".

In addition, the use of the oblique or slash / may present advantages in the definition of the precise cleavage sites of restriction endonucleases. For instance, the cleavage specificity of the common enzyme EcoRI might be represented by G/AATTC, where cleavage occurs in both strands of the self-symmetrical sequence between the G and A residues.

It is emphasised that the symbols appearing in this section do not form an integral part of the recommendations and must therefore be defined explicitly in the context in which they are used.

Table 1. Summary of single-letter code recommendations

Symbol	Meaning	Origin of designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Ketone
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

5. DISCUSSION

The present nomenclature, summarised in Table 1, has been formulated to deal with incomplete specification of bases in nucleic acid sequences. In cases where two or more bases are permitted at a particular position the nomenclature permits the allocation of a single-letter symbol. The nomenclature may also be applied where uncertainty exists as to extent and/or identity. For double-stranded nucleic acids Table 2 permits the allocation of symbols to the complementary strand. Examples are given whereby the nomenclature is applied to sequences recognised by certain type II restriction endonucleases (Table 3) and to uncertainties in deriving a nucleic acid sequence from the corresponding amino acid sequence (Table 4).

Two applications fall outside the scope of the nomenclature and these are considered separately below.

5.1 Interpretation of primary sequencing data

In certain cases the nomenclature permits uncertainty encountered in nucleic acid sequencing to be represented using a single letter code. At least two applications, namely "either X or XY" and "probably X", are not adequately handled. Specialised nomenclatures have been developed specifically for such purposes (for instance ref. 3).

Table 2. Definition of Complementary Symbols

Symbol	A	B	C	D	G	H	K	M	S	T	V	W	N
Complement	T	V	G	H	C	D	M	K	S*	A	B	W*	N*

*In certain cases the symbol and its complement are identical.

Table 3. Single-letter code recognition sequences for several type II restriction endonucleases recognising multiple sequences*

Enzyme	Recognition sequence
AccI	G T/M K A C
AcyI (HgiDI)	G R/C G Y C
AflIII	A/C R Y G T
AvaI	C/Y C G R G
AvaII	G/G W C C
BanI (HgiCI)	G/G Y R C C
BanII (HgiJII)	G R G C Y/C
CfrI (GdiII)	Y/G G C C R
EcoRII	/C C W G G
HaeI	W G G/C C W
HaeII	R G C G C/Y
HgiAI	G W G C W/C
HindII (HincII)	G T Y/R A C
NciI (CauII)	C C/S G G
NspI (Nsp-7524-I)	R C A T G/Y
NspBII	C M G/C K G
SduI (Nsp-7524-II)	G D G C H/C
XhoII	R/G A T C Y

*Recognition sequences are presented 5'-3': the exact position of cleavage is indicated here by /. Data from ref. 12.

5.2 Modified nucleotides

In a number of organisms DNA and RNA are modified at certain positions. For instance, the DNA of Escherichia coli is usually methylated at the 5-position of the adenine residue in the sequence 5'-GATC-3' [23]. The present nomenclature does not allocate any specific symbol to these modified nucleotides for the following reasons. 1. The presence or absence of a given modification depends upon the location of the DNA. Sequences modified in one organism may not be modified in another. 2. Modification is usually statistical, in that only a proportion of possible sites for modification may actually be utilised in vivo. Modification of a nucleotide or base in a given polynucleotide is not a function of the sequence per se. Although it is recognised that stable RNA species (tRNA and rRNA) often carry a constant pattern of post-translational modification, the present nomenclature is not intended to overlap with or supplant existing systems. It would probably be impossible to devise a simple and logical system that avoided all conflict with previous usage. One should therefore recognise that such conflict is possible and take steps to prevent it from generating confusion, for example in relation to the symbols B, D and S, which have been recommended for 5-bromouridine, 5,6-dihydrouridine and thiouridine respectively [1], or W, which is sometimes used for wyosine [27].

Table 4. Triplet correspondence for amino acids (standard genetic code)

Amino acid	Single-letter code	Triplet (5'-3')
Glycine	G	GGN
Alanine	A	GCN
Valine	V	GTN
Leucine	L	YTN (CTN and TTR)*
Isoleucine	I	ATH
Proline	P	CCN
Phenylalanine	F	TTY
Tyrosine	Y	TAY
Cysteine	C	TGY
Methionine	M	ATG
Histidine	H	CAY
Lysine	K	AAR
Arginine	R	MGN (CGN and AGR)*
Tryptophan	W	TGG
Serine	S	WSN (TCN and AGY)*
Threonine	T	ACN
Aspartic acid	D	GAY
Glutamic acid	E	GAR
Asparagine	N	AAY
Glutamine	Q	CAR
Aspartic acid or asparagine	B	RAY
Glutamic acid or glutamine	Z	SAR
Terminator	.	TRR (TAR and TGA)*
Unknown	X	NNN

*The sequence of amino acids is uniquely specified by the nucleotide sequence. Similarly, it is possible to convert an amino acid sequence to a linear order of base uncertainties, but this raises problems with the codons for leucine, arginine, serine and termination. With leucine, for example, the coding triplets are precisely specified by CTN and TTR, but combining these gives YTN, which also includes two phenylalanine codons, TTT and TTC. Thus information may be lost when a amino acid sequence is converted into a single sequence of base-uncertainty symbols. To avoid ambiguity, therefore, it is important to make it clear whenever the triplet YTN, for example, occurs in a sequence deduced from the occurrence of a leucine residue in the corresponding amino acid sequence that it does not include TTT or TTC as possibilities, etc. To emphasise this, it may be helpful to print such triplets in italics.

ACKNOWLEDGEMENTS

This is a document of the Nomenclature Committee of the International Union of Biochemistry (NC-IUB) whose members are H. B. F. Dixon (Chairman), H. Bielka, C. R. Cantor, C. Liébecq (representing the Committee of Editors of Biochemical Journals), N. Sharon, S. F. Velick and J. F. G. Vliegenthart. NC-IUB thanks the panel, whose members were F. Blattner (U.S.A.), N. L. Brown (U.K.), D. L. Brutlag (U.S.A.), W. M. Fitch (U.S.A.), W. Goad (U.S.A.), R. Grantham (France), G. Hamm (Federal Republic of Germany), L. H. Kedes (U.S.A.), R. Lathe (France, convener), D. W. Mount (U.S.A.), J. Schroeder (U.S.A.), R. Staden (U.K.), P. A. Stockwell (New Zealand), for drafting these recommendations. Comments may be sent to any member of the NC-IUB, or to its secretary, A. Cornish-Bowden, Department of Biochemistry, University of Birmingham, P. O. Box 363, Birmingham B15 2TT, England, or to the convenor of the panel, R. Lathe, whose present address is: A.F.R.C. Animal Breeding Research Organisation, West Mains Road, Edinburgh EH9 3JQ, Scotland.

REFERENCES

1. IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and Symbols for Nucleic Acids, Polynucleotides and their Constituents. Recommendations 1970. Arch. Biochem. Biophys. 145, 425-436 (1971); Biochem. J. 120, 449-454 (1970); Biochemistry 9, 4022-4027 (1970); Biochim. Biophys. Acta, 247, 1-12 (1971); Eur. J. Biochem. 15, 203-208 (1970), corrected 25, 1 (1972); Hoppe-Seyler's Z. Physiol. Chem. (in German) 351, 1055-1063 (1970); J. Biol. Chem. 245, 5171-5176 (1970); Mol. Biol. (in Russian) 6, 167-174 (1972); Pure Appl. Chem. 40, 277-290 (1974); also pp. 116-121 in Biochemical Nomenclature and Related Documents (1978), the Biochemical Society, London.
2. Fitch, W. M. (1973) J. Molec. Evol. 2, 123-136.
3. Staden, R. (1979) Nucl. Acids. Res. 6, 2601-2610.
4. Clayton, J. and Kedes, L. (1982) Nucl. Acids Res. 10, 305-321
5. Orcutt, B. C., George, D. G., Fredrickson, J. A. and Dayhoff, M. O. (1982) Nucl. Acids Res. 10, 157-174.
6. Patarca, R., Dorta, B. and Ramirez, J. L. (1982) Nucl. Acids Res. 10, 175-182.
7. Stockwell, P. A. (1982) Nucl. Acids Res. 10, 115-125.
8. Tolstoshev, C. M. and Blakesley, R. W. (1982) Nucl. Acids Res. 10, 1-17.
9. McClelland, M. (1983) Nucl. Acids Res. 11, R169-173.
10. Fitch, W. M. (1983) Unpublished manuscript.
11. Brown, N. L. (1983) Unpublished manuscript.
12. Roberts, R. J. (1983) Nucl. Acids Res. 11, R135-167.
13. Zabeau, M. and Roberts, R. J. Unpublished observations, cited in ref. 12.
14. Janulaitis, A., Marcinkeviciene, L., Petrusyte, M. and Mironov, A. (1981) FEBS Lett. 134, 172-174.
15. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353.
16. Siebenlist, U., Dimpson, R. B. and Gilbert, W. (1980) Cell 20, 269-281.
17. Stormo, G. D., Schneider, T. D., Gold, L. and Ehrenfeeucht, A. (1982) Nucl. Acids Res. 10, 2997-3011.
18. Kozak, M. (1983) Microbiol. Rev. 47, 1-45.
19. Ycas, M. (1967) The Biological Code, Wiley-Interscience, New York.
20. Goeddel, D. V., Yelverton, E., Ullrich, A., Heynecker, H. L., Miozzari, G., Holmes, W., Seeburg, P. H., Dull, T., May, L., Stebbing, N., Crea, R., Maeda, S., McCandliss, R., Sloma, A., Tabor, J. M., Gross, M., Familletti, P. C. and Pestka, S. (1980) Nature 287, 411-416.
21. Fitch, W. M. (1971) System. Zool. 20, 406-416.
22. Stormo, G. D., Schneider, T. D. and Gold, L. M. (1982) Nucl. Acids Res. 10, 2971-2996.
23. Geier, G. E. and Modrich, P. (1979) J. Biol. Chem. 254, 1408-1413.
24. International Union of Pure and Applied Chemistry (1979) Nomenclature of Organic Chemistry, Sections A, B, C, D, E, F and H, (Rigaudy, J. and Klesney, S. P., eds.) Pergamon Press, Oxford.
25. IUPAC Commission on Nomenclature of Organic Chemistry (CNOC) Definitive rules for nomenclature of organic chemistry, J. Am. Chem. Soc. 82, 5545-5574 (1960).
26. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Abbreviations and Symbols for the Description of Conformations of Polynucleotide Chains. Recommendations 1982. Eur. J. Biochem. 131, 9-15 (1983); Proceedings of the 16th Jerusalem Symposium, Nucleic Acids, the Vectors of Life (Pullman, B. and Jortner, J., eds.), pp. 559-565, Reidel, Dordrecht, 1983; Pure Appl. Chem. 55, 1273-1280 (1983).
27. Singhal, R. P., Roberts, E. F. and Vakharia, V. N. (1983) Prog. Nucl. Acid Res. and Mol. Biol. 28, 211-252.