Simplicity, function, and legibility in an enhanced ambigraphic nucleic acid notation

David A. Rozak[1] and Anthony J. Rozak[2]

[1]Bacteriology Division, United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, MD and [2]Department of Visual Studies, College of Arts and Sciences, State University of New York at Buffalo, Buffalo, NY, USA

BioTechniques 44:811-813 (May 2008)
doi 10.2144/000112727

We previously showed that an ambigraphic nucleic acid notation, based on symmetrical lowercase Roman characters, permits users to complement DNA by physically rotating the sequence text 180°. This article describes an enhanced ambigraphic notation, which uses concept-related symbol design, rather than the arbitrary set of symbols that constitute the Roman alphabet, to logically encode the four DNA bases and 11 ambiguity characters. As ambigrams, the symbols continue to permit the rapid derivation of complementary sequences and visualization of palindromic DNA. In addition, the new AmbiScript notation uses legibility principles to support the identification of sequence polymorphism and improves writing efficiency by requiring fewer strokes per character than the International Union of Pure and Applied Chemistry (IUPAC) notation.

In 1970 the International Union of Pure and Applied Chemistry (IUPAC) formalized a notation, based on the Roman alphabet, for specifying the arrangement of bases in a genetic sequence (1). The universally adopted system was designed to accommodate the 4 standard DNA bases as well as 11 ambiguity characters, which enable the identification of different bases at a given position (2). We recently showed that symmetries found among lowercase Roman characters could be used to produce an alternative genetic notation with additional analytic advantages (3). Specifically, representing complementary nucleotides with ambigraphic Roman characters such as “b,” “q,” “n,” and “u,” which are configurationally identical when rotated 180°, makes it possible to complement genetic sequences by rotating an entire string of symmetrically encoded nucleotides 180°. The term “ambigram” was first used by Hofstadter to describe sets of symbols that have the same or different meanings when viewed in different orientations (4). While the Roman alphabet contains enough ambigraphic characters to encode the 4 DNA bases and 11 ambiguity characters, derivation of the ambiguity characters is not obvious, making our original notation difficult to learn.

Here, we abandon the Roman alphabet to design a novel ambigraphic character set, which, aside from being easier to learn and write, better supports the manipulation, visualization, and analysis of genetic sequences. As shown in Table 1, all four nucleotides of the proposed AmbiScript notation are encoded by the shape and orientation of two basic single-stroke and double-stroke patterns. Because guanine and cytosine form strong complementary interactions via three hydrogen bonds, these bases are represented by a tall slightly curved diagonal stroke, which is easily recognized in sequences and is securely rooted to the typographic baseline by a horizontal stroke. Adenine and thymine, on the other hand, form weaker interactions through two hydrogen bonds, and each is rendered by a low arch whose end points touch the baseline similarly at only two points.

The AmbiScript characters are oriented above or below the baseline depending on whether the corresponding bases are purines or pyrimidines. Because the double-ringed purines have higher molecular weights than pyrimidines, their corresponding symbols hang below the baseline. By contrast, the symbols for the lighter pyrimidines rise above the baseline.

The IUPAC defined 11 ambiguity characters, which represent variations in the type of base found at a particular location (2). Unfortunately, the mnemonics used to assign Roman characters to each possible set of nucleotides in Table 1 are not always intuitive or easy to remember. More importantly, it is often difficult to recall which IUPAC ambiguity character complements another, as the symbols do not bear any conceptual relationships to their complements.

We sought to simplify the derivation of all 11 ambiguity characters by designing AmbiScript symbols that could be overlaid to form compound characters encoding positional polymorphisms. The resulting ambiguity characters, depicted in Table 1, are easy to construct and decipher. For ease of writing and visual aesthetics, the horizontal lines associated with the symbols for guanine and cytosine

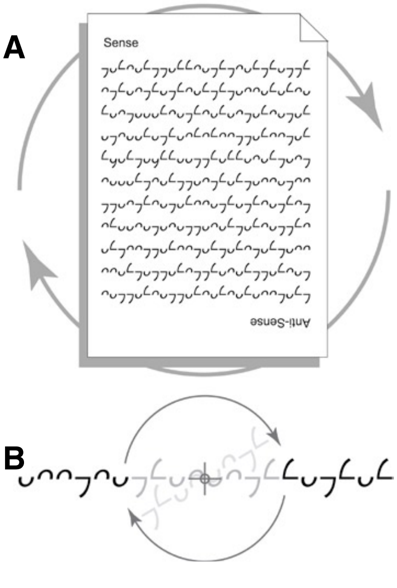


Figure 1. AmbiScript symmetries support sequence complementation and palindrome identification. (A) The proposed ambigraphic notation permits complementation of a genetic sequence by rotating the text 180°. (B) Palindromes (gray text) are identifiable as strings of nucleotides that leave the sequence unchanged when rotated 180° around their center points.

are omitted when combined with the symbols for adenine and thymine.

The symbols for the 4 bases and 11 ambiguity characters presented in Table 1 have also been designed as ambigrams, which match the symbols for their complementary bases when rotated 180°. As with our previous notation (3), this means that the complement of any AmbiScript sequence can be determined by rotating the text 180° (Figure 1A).

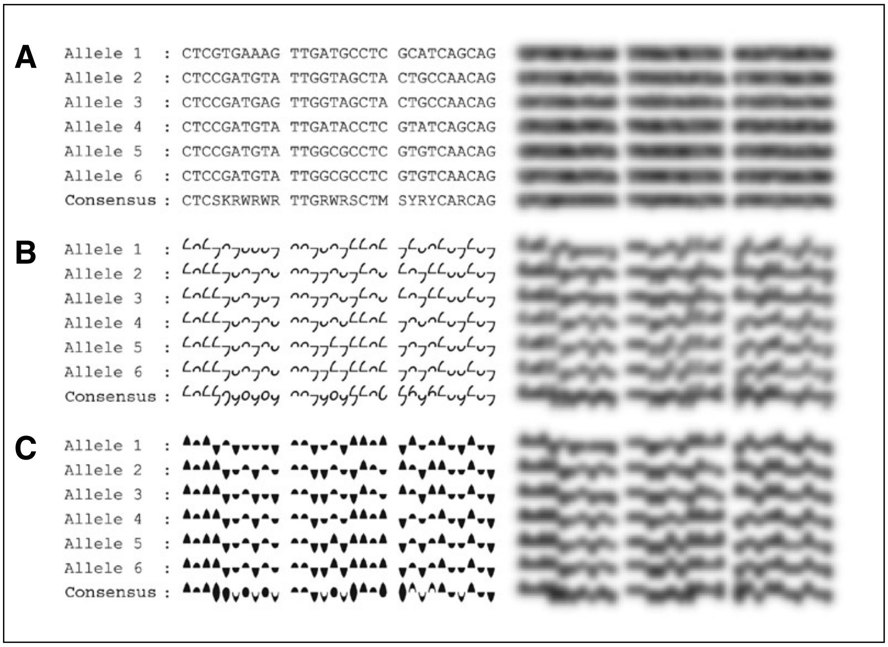
The symmetries inherent in our ambigraphic notation also facilitate the detection of palindromic sequences. Figure 1B shows how palindromes, such as those associated with endonuclease restriction sites, are readily identifiable as stretches of characters, which remain unchanged when rotated 180° around a point of symmetry. Using the AmbiScript notation, it is relatively easy to scan a genetic sequence for points of rotational symmetry around which two or more characters can be rotated without altering the sequence.

Researchers have occasionally sought to overcome visual limitations of the IUPAC symbols by proposing alternate notations that use visually distinct fonts to highlight sequence polymorphisms (5–8). Our own efforts to develop the visually distinct AmbiScript symbols were informed by earlier research on designing a phonetically consistent, concept-related set of phonograms as an alternative to the Roman alphabet (9), by Pitman shorthand, and by Miles Tinker's extensive legibility studies (10). Consistent with Tinker's findings, our notation relies heavily on ascenders and descenders to improve sequence legibility. Indeed, many of the AmbiScript symbols resemble the highly legible Caslon old-style numerals 7 and 9, which have strokes extending below the baseline while their top edges align with the font's "x" height, and exhibit symmetrical contrasts similar to those that characterize the highly legible lowercase characters d, q, b, and p. Our notation further aids visualization of sequence polymorphisms by using symbols that sit entirely above or below the typographical baseline. The blurred sequence texts in Figure 2 illustrate the extent to which the exterior shapes of

Table 1. IUPAC and Ambigraphic Nucleic Acid Notations

	IUPAC Notation[a]			AmbiScript Notation	
	Symbol	Meaning	Mnemonic	Normal	Ultra Bold
DNA Bases	G	Guanine	Guanine	ᒣ	▼
	T	Thymine	Thymine	∩	◚
	A	Adenine	Adenine	∪	◛
	C	Cytosine	Cytosine	ᒪ	▲
Ambiguity Sets	R	G + A	puRine	y	∨
	Y	T + C	pYrimidine	ʎ	∧
	S	G + C	Strong interactions (3 H bonds)	ƽ	⬮
	W	T + A	Weak interactions (2 H bonds)	o	●
	K	G + T	Keto	ʖ	⬮
	M	A + C	aMino	ʗ	⬮
	D	G + T + A	Not-C (D follows C in alphabet)	9	⬮
	H	T + A + C	Not-G (H follows G)	6	⬮
	B	G + T +C	Not-A (B follows A)	ƽ	⬮
	V	G + A + C	Not-T or U (V follows U)	ɕ	⬮
	N	G + A + T + C	aNy	ɋ	⬯

[a] See References 1 and 2.



```
Allele 1  : CTCGTGAAAG TTGATGCCTC GCATCAGCAG
Allele 2  : CTCCGATGTA TTGGTAGCTA CTGCCAACAG
Allele 3  : CTCCGATGAG TTGGTAGCTA CTGCCAACAG
Allele 4  : CTCCGATGTA TTGATACCTC GTATCAGCAG
Allele 5  : CTCCGATGTA TTGGCGCCTC GTGTCAACAG
Allele 6  : CTCCGATGTA TTGGCGCCTC GTGTCAACAG
Consensus : CTCSKRWRWR TTGRWRSCTM SYRYCARCAG
```

Figure 2. The proposed notation uses distinctive shapes, orientations, and vertical offsets to highlight nucleotide polymorphisms in multiple sequence alignments. (A) An IUPAC alignment of gene fragments from six *Aeromonas media* CDC0862–83 16s rRNA alleles (11) is compared with identical alignments rendered in the (B) normal and (C) ultra-bold AmbiScript fonts. Consensus sequences at the bottom of each alignment demonstrate the use of ambiguity characters. Blurred alignments illustrate the contribution of ascenders and descenders to overall sequence legibility.

characters are essential for legibility. In fact, the outer-edge shapes are generally more readable than interior details, such as the horizontal bar that distinguishes a capital "C" from a "G."

Despite its enhanced legibility, the proposed AmbiScript notation remains easy to write by hand because it avoids using solid or geometrically complex characters. In fact, if one uses the average number of pen strokes and transfers required to write a set of symbols (average strokes per character, ASPC) as a measure of writing efficiency, it becomes apparent that the AmbiScript symbols for the four DNA bases (2.50 ASPC) are easier to write than the comparable IUPAC characters (3.75 ASPC).

Where writing speed is not an issue, our notation takes full advantage of the visual contrasts afforded by solid symbols in the ultra-bold AmbiScript font for use in print and computer displays. This ultra-bold font, which is compared with the written script in Table 1 and Figure 2, fills the spaces between the strokes and arches to further emphasize nucleotide polymorphisms. Both forms of the notation are available on PCs and Macs by downloading a TrueType font, included here as Supplementary Material available online at www.BioTechniques.com.

ACKNOWLEDGMENTS

This project was supported in part by an appointment to the Internship/Research Participation Program for the United States Army Medical Research and Materiel Command, administered by the Oak Ridge Institute for Science and Education through an agreement between the United States Department of Energy and the United States Army Medical Research and Materiel Command. The views of the authors do not purport to reflect the positions of the Department of the Army or the Department of Defense.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

1. 1970. IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970. Eur. J. Biochem. *15*:203-208.
2. 1986. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Proc. Natl. Acad. Sci. USA *83*:4-8.
3. **Rozak, D.A.** 2006. The practical and pedagogical advantages of an ambigraphic nucleic acid notation. Nucleosides Nucleotides Nucleic Acids *25*:807-813.
4. **Hofstadter, D.R.** 1985. Metamagical Themas: Questioning for the Essence of Mind and Pattern. Basic Books, NY.
5. **Cowin, J.E., C.H. Jellis, and D. Rickwood.** 1986. A new method of representing DNA sequences which combines ease of visual analysis with machine readability. Nucleic Acids Res. *14*:509-515.
6. **Jarvius, J. and U. Landegren.** 2006. DNA Skyline: fonts to facilitate visual inspection of nucleic acid sequences. BioTechniques *40*:740.
7. **Schneider, T.D. and R.M. Stephens.** 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*:6097-6100.
8. **Zimmerman, P.A., M.L. Spell, J. Rawls, and T.R. Unnasch.** 1991. Transformation of DNA sequence data into geometric symbols. BioTechniques *11*:50-52.
9. **Rozak, A. J.** 1977. The inadequacies of the Roman alphabet and a proposed phonetic alphabet with concept-related phonograms. Icographic *11*:22-27.
10. **Tinker, M.A.** 1963. Legibility of Print. Iowa State University Press, Ames, IA.
11. **Morandi, A., O. Zhaxybayeva, J.P. Gogarten, and J. Graf.** 2005. Evolutionary and diagnostic implications of intragenomic heterogeneity in the 16S rRNA gene in *Aeromonas* strains. J. Bacteriol. *187*:6561-6564.

Received 23 October 2007; accepted: 11 December 2007.

Address correspondence to David A. Rozak, Bacteriology Division, United States Army Medical Research Institute of Infectious Diseases, 1425 Porter St., Fort Detrick, MD 21702, USA. e-mail: david.rozak@amedd.army.mil

To purchase reprints of this article, contact: Reprints@BioTechniques.com