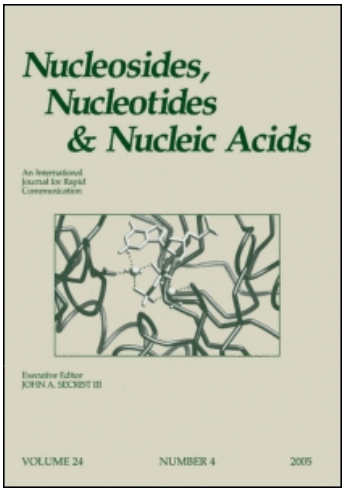

Nucleosides, Nucleotides and Nucleic Acids
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597286

The Practical and Pedagogical Advantages of an Ambigraphic Nucleic Acid Notation

David A. Rozak[a]
[a] Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland, USA

To cite this Article Rozak, David A.(2006) 'The Practical and Pedagogical Advantages of an Ambigraphic Nucleic Acid Notation', Nucleosides, Nucleotides and Nucleic Acids, 25: 7, 807 — 813
To link to this Article: DOI: 10.1080/15257770600726109
URL: http://dx.doi.org/10.1080/15257770600726109

Nucleosides, Nucleotides, and Nucleic Acids, 25:807–813, 2006

ISSN: 1525-7770 print / 1532-2335 online
DOI: 10.1080/15257770600726109

Taylor & Francis
Taylor & Francis Group

THE PRACTICAL AND PEDAGOGICAL ADVANTAGES OF AN AMBIGRAPHIC NUCLEIC ACID NOTATION

David A. Rozak □ *Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland, USA*

□ *The universally applied IUPAC notation for nucleic acids was adopted primarily to facilitate the mental association of G, A, T, C, and the related ambiguity characters with the bases they represent. However, it is possible to create a notation that offers greater support for the basic manipulations and analyses to which genetic sequences frequently are subjected. By designing a nucleic acid notation around ambigrams, it is possible to simplify the frequently applied process of reverse complementation and aid the visualization of palindromes. The ambigraphic notation presented here also uses common orthographic features such as stems and loops to highlight guanine and cytosine rich regions, support the derivation of ambiguity characters, and aid educators in teaching the fundamentals of molecular genetics.*

Keywords Nucleic acid; Ambigram; IUPAC notation; Reverse complementation; Palindrome

INTRODUCTION

In 1970 the International Union of Pure and Applied Chemistry (IUPAC) formalized a four-character code (G, A, T, and C) for representing arrangements of bases in nucleic acid sequences.[1] The organization later expanded the notation by defining 11 ambiguity characters to represent all possible combinations of the four DNA bases.[2] As illustrated in Table 1, the guiding principle in assigning letters to each of the bases and all the possible combinations thereof was to facilitate the association of each symbol with the base(s) it represents.

However, it is possible for a nucleic acid notation to do more than provide a convenient shorthand for genetic sequences. This article briefly presents an ambigraphic nucleic acid notation that facilitates common

Received 22 February 2006; accepted 27 March 2006.

This work was supported by NIH grant GM62154. I am grateful for the encouragement and advice I received from John Bodnar and Phil Bryan, particularly with respect to the notation's value in dealing with ambiguity characters and its possible role in education.

Address correspondence to David A. Rozak, The United States Army Research Institute for Infectious Diseases, 1425 Porter St., Fort Detrick, MD 21702. E-mail: david.rozak@amedd.army.mil

TABLE 1 Nucleic Acid Assignments and Mnemonics

	IUPAC notation[1,2]			Ambigraphic notation	
	Symbol	Meaning	Mnemonic	Symbol	Meaning
DNA bases	G	Guanine	Guanine	b	Guanine
	T	Thymine	Thymine	u	Thymine
	A	Adenine	Adenine	n	Adenine
	C	Cytosine	Cytosine	q	Cytosine
Ambiguity sets	R	G, A	puRine	h	b, n
	Y	T, C	pYrimidine	y	u, q
	S	G, C	Strong interactions (3 H bonds)	o	b, q
	W	T, A	Weak interactions (2 H bonds)	x	u, n
	K	G, T	Keto	l	b, u
	M	A, C	aMino	j	n, q
	D	G, T, A	Not-C (D follows C in alphabet)	m	b, u, n
	H	T, A, C	Not-G (H follows G)	w	u, n, q
	B	G, T, C	Not-A (B follows A)	p	b, u, q
	V	G, A, C	Not-T or U (V follows U)	d	b, n, q
	N	G, A, T, C	aNy	z	b, u, n, q

sequence manipulations and analyses by reflecting the natural symmetries found in DNA. The notation uses ambigraphic characters contained in the Roman alphabet to reduce the frequently applied process of reverse complementation to the physical act of rotating the entire text 180°. The notation's intrinsic symmetry also supports the visualization of palindromic sequences, which can be indicative of endonuclease cleavage sites and other biologically significant genetic landmarks. Sequence analysis is further supported by the consistent use of stems and arches to indicate purines, pyrimidines, and strong guanine-cytosine binding partners, which play a key role in duplex stability. Each of these features make the proposed ambigraphic notation a practical tool for researchers and educators alike.

AMBIGRAMS AND COMPLEMENTARY DNA

Ambigrams[1] are symbols, words, or phrases that can be read in more than one orientation to convey the same or different meanings. One of the more common forms of ambigrams results from rotating a string of

[1]Douglas R. Hofstadter was apparently the first to use the term "ambigram" in print, attributing its origin to an unnamed friend. According to Hofstadter, the word describes "a single written specimen, or 'gram,' [which] has more than one reading, depending on the observer's point of view. Often the 'grams' are symmetric and read the same both ways, but this is not essential: some have two totally different readings."[3]

letters 180°. Depending on whether the text conveys the same or different meaning in each of the orientations, the ambigram is said to be ambiguous or unambiguous in nature. English-language ambigrams include the unambiguous "SWIMS" and the ambiguous "MOM" or "WOW," depending on its orientation.

Ambigrams are well suited for representing nucleic acids because they can be used to reflect an intrinsic biological symmetry of the polymers. Two strands of DNA can interact by hydrogen bonds to form a double helix if they run in opposite directions to one another and each base is paired with its complement in the other strand. Since cytosine is always paired with guanine, and adenine is matched with thymine, one strand of a double helix is generally sufficient to define the other.

When one wishes to determine the nucleic acid sequence of a complexed strand, he or she generally reverses the order of the all the letters in the sequence and replaces each IUPAC character in the original sequence with the character that represents its complement. Even though this two-step process of reverse complementation is easily performed on a computer, it is common enough in the design of short oligonucleotides that researchers frequently find it just as convenient to carry out the transformation with paper and pencil.

If each of the two complementary pairs of bases is represented by a single ambiguous ambigram (i.e. an ambigram with different meanings depending on its orientation), then a base's complement can be determined by simply rotating its symbol 180°. For example, if guanine were represented by a lower case *b* and cytosine by a lower case *q*, the complement of guanine could be determined by rotating the *b* 180° to obtain a *q* and vice versa. The same effect could be achieved for thymine and adenine by representing each of these bases with a lowercase *u* and *n,* respectively. In this way the ambigraphic symbols reflect the complementary nature of the bases they represent rather than an abstract mnemonic for their chemical names.

On the surface, the use of ambigrams to represent complementary bases may not appear to offer much of an advantage because it is easy enough to remember that the IUPAC *G* pairs with *C* and *A* with *T*. However, as illustrated in Figure 1, ambigrams greatly facilitate the reverse complementation of oligonucleotides and reduce the chance of human error by transforming the two-step process into one of simply rotating the entire text 180°. In general, rotating a string of characters by 180° produces the same results as reversing the order of all the characters in the string and independently rotating each one by 180°. Because rotating each character in the proposed ambigraphic notation is equivalent to determining its complement, the process of rotating the entire string 180° is identical to that of reverse complementation.

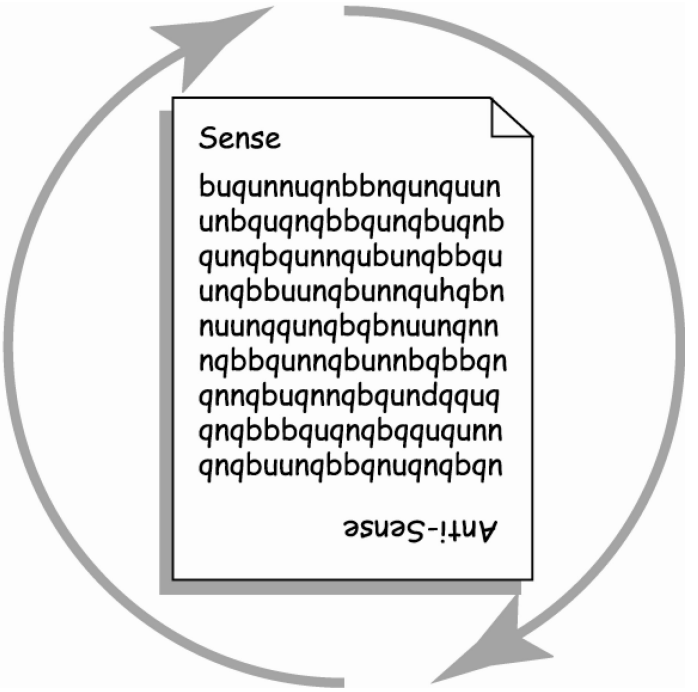

FIGURE 1 The proposed ambigraphic nucleic acid notation permits reverse complements to be determined by simply rotating the text 180°.

AMBIGRAMS AND AMBIGUOUS BASES

The ambigraphic notation can be extended to accommodate all 11 ambiguity characters as well. Unlike the set of characters that represent single bases, some ambiguity characters are actually their own complements. For example, the IUPAC *W*, which conveys the possibility of finding an adenine or thymine in the designated position, will also be represented by a *W* on the complementary strand. If ambiguity characters are to be converted to an ambigraphic notation, then *W* would have to be replaced by an unambiguous ambigram such as the lowercase *x*. The lowercase *x* is considered to be unambiguous because it resembles itself when rotated 180°. For researchers like myself, who frequently encounter ambiguity characters in the design or characterization of genetic libraries, the notation described here offers the economy of not having to memorize, derive, or look up the complements for each of the ambiguity sets encountered in a sequence transformation. Fortunately, there are enough ambiguous and unambiguous ambigraphic lowercase letters in the Roman alphabet to assign the appropriate class of character to each of the four DNA bases and the 11 ambiguity characters. These assignments are given alongside the IUPAC characters in Table 1. While some of these ambigrams such as *h* and

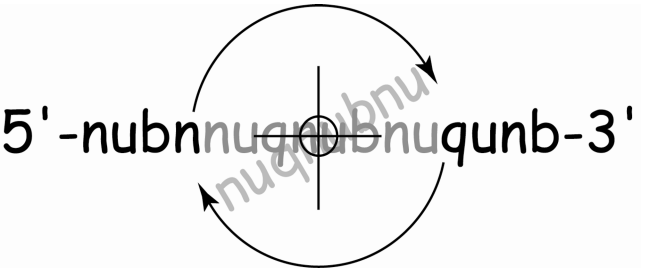

FIGURE 2 Palindromes are identifiable by scanning the text for centers of rotational symmetry around which two or more characters can be rotated 180° without changing the genetic sequence.

y are more strained in their likeness than others, these discrepancies do not impede the function of the notation and can be minimized with selected type fonts.

VISUALIZING PALINDROMES

In addition to simplifying the derivation of complementary sequences, ambigraphic symbols support the visualization of short palindromic segments, which are often associated with endonuclease cleavage sites and other genetic features. Per Figure 2, palindromes are detected by scanning the string of ambigraphic characters for centers of rotational symmetry—imaginary points around which multiple letters can be rotated 180° to leave the text unchanged. Once a center of rotational symmetry has been found for two characters, it is possible to expand one's focus outward to determine the full size of the palindrome.

ASSIGNING AMBIGRAMS

As mentioned at the outset, IUPAC characters were selected largely to reinforce the mental associations of symbols and bases. As presented in Table 1, these mnemonics are effectively applied to the four basic nucleotides but become more strained in relating IUPAC symbols to some of the ambiguity sets. Unfortunately, the more limited pool of characters available for the proposed ambigraphic notation excludes those letters lacking natural ambiguous or unambiguous symmetries such as G, A, T, and C. This constraint makes it impossible to employ many of the simple abbreviations used in the IUPAC notation. Therefore, in an effort to facilitate the interpretation and analysis of the ambigrams used with the proposed notation, characters were chosen to visually reflect key biological attributes of the bases they represent. This was achieved through the consistent use of common orthographic features to represent purines, pyrimidines, and strong hydrogen-bonding partners. An attempt was also made to select

ambiguity characters that embody the dominant orthographic features of their representative sets. As a result, three basic guidelines have been established to reinforce the relationship between the proposed symbols and bases:

1. Purines and pyrimidines are represented by characters that have stems and arches facing up and down, respectively. For example, *b* and *n*, which represent the purines guanine and adenine, point upward, while their pyrimidine complements, *q* and *u*, face down.
2. Bases that strongly bind their complements are represented by characters with closed bodies—a feature that supports the visual identification of guanine- and cytosine-rich regions by permitting readers to quickly scan the text for concentrations of letters with this distinctive feature.
3. Ambiguity characters generally embody the most prevalent orthographic features of the represented set. As a result, *h* (represented by an *R* in the IUPAC notation) combines the upward stem and arch of *b* and *n* to convey the possibility of finding guanine or adenine at that position. The letter *d* (IUPAC *V*) was chosen to represent *b*, *n*, and *q* because it combines the predominantly closed bodies of *b* and *q* and upward stems of *b* and *n*.

In this way, the symbols provide visual clues to key attributes of the bases they represent rather than relying on the reader to extrapolate these features from the chemical names of the bases.

CONCLUSION

The IUPAC notation unnecessarily complicates reverse complementation by requiring researchers to carry out the common transformation on a computer or with paper and pencil—approaches that occasionally lead to transcription errors when hard copies are involved. By representing nucleic acid sequences as ambigrams it is possible to avoid these shortcomings while at the same time supporting the visualization of palindromes and identification of guanine- and cytosine-rich regions—features that benefit researchers and educators alike. Given the broad acceptance of the IUPAC notation and general use of computers in modern biology, the mechanical, visual, and pedagogical advantages of an ambigraphic notation may not warrant a change in notation. Indeed, some may argue that commonly available bioinformatics tools make the manual functions supported by an ambigraphic notation all but obsolete. However, just as word processors have failed to overcome the preference many have for reading and editing paper manuscripts, so too are computers unlikely to entirely displace the manual manipulation and interpretation of genetic code. Rather, modern computational resources can be enlisted to facilitate and even catalyze

a transition to a notation more capable of reflecting inherent genetic symmetries like the one presented here.

REFERENCES

1. IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for nucleic acids, polynucleotides and their constituents. Recommendations 1970. *Eur. J. Biochem.* **1970**, 15(2), 203–208.
2. Nomenclature Committee for the International Union of Biochemistry. Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *Journal of Biological Chemistry* **1986**, 261(1), 13–17.
3. Hofstadter, D.R. 1985. *Metamagical Themas: Questing for the Essence of Mind and Pattern*, p. 274, Basic Books, New York.