RECONSTRUCTION OF PROTEIN AND NUCLEIC ACID SEQUENCES

II. ISOTOMERS

C. R. Merril, J. E. Mosimann, D. F. Bradley, and M. B. Shapiro

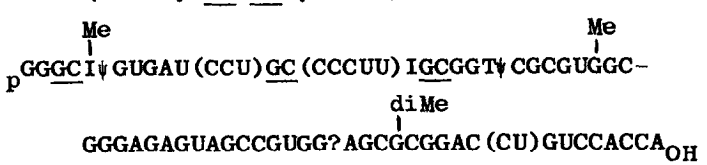
National Institutes of Health
Bethesda, Maryland

Received February 19, 1965

In the reconstruction of protein and nucleic acid sequences it is important to know whether a proposed sequence is unique or whether there are alternative sequences consistent with the data employed. We refer to such alternative sequences as isotomers. A method has been developed for testing proposed sequences and sets of fragments to determine if isotomers exist and to construct them.

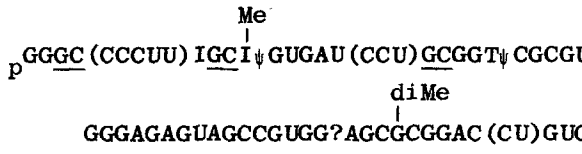
Currently the fragmentation stratagem is being used to determine the sequences of S-RNA's. The endonucleases available for this procedure are T_1 and pancreatic RNAase which cut after guanine and after the pyrimidines, respectively. If only data obtained by complete digestion with these enzymes were used in reconstructing the proposed sequences of alanine and serine S-RNA's, isotomers must exist.

The fragment data upon which the proposed alanine S-RNA sequence (Holley et al., 1964)



is based have yet to be reported. However, if one assumes that only fragments produced by complete digestion with T_1

and pancreatic RNAase were used, then other isotomers must exist, e.g.,



$$_{p}GGGC(CCCUU)IGCI\psi GUGAU(CCU)GCGGT\psi CGCGUGGC-$$

$$GGGAGAGUAGCCGUGG?AGCGCGGAC(CU)GUCCACCA_{OH}.$$

These two sequences would be indistinguishable even if all of the possible fragments were collected and their sequences determined, because they both give exactly the same fragments. Holley et al., (1964) state that their sequence "is only one of thousands of overall sequences that are equally probable with present data."

The tentative sequence for serine S-RNA

$$_{p}GUUGCCGAGAGUAUCGGGGCCUUGCAAYUGUUYAGAY-$$

$$KEGGTUUGCAAGGCCCCGAUACUCUCGGCAACCA_{OH}$$

proposed by Cantoni et al., (1963) produces the same set of fragments as

$$_{p}GUUGCCUUGCCGAGAGUAUCGGGGCAAYUGUUYAGAY-$$

$$KEGGTUUGCCCCGAUACUCUCGGCAAGGCAACCA_{OH}$$

We present these proposed sequences as examples of biopolymers whose sequences could not have been determined by complete digestion with available enzymes. To solve the sequences of such polymers would require the use of other kinds of data. Partial hydrolysis with the same enzymes could, in principle, provide such data. Cantoni actually employed additional data in his reconstruction, some of which required that the polymer exhibit a high degree of base pairing when the molecule is folded back on itself in the shape of a hairpin. The isotomer presented also satisfies the base pairing requirement.

We now demonstrate the method by which these isotomers

were constructed. In these constructions, the underlined GC pairs play a critical role. The set of elements after which an enzyme cuts will be called the "break-set" of the enzyme. Therefore G (Me-G and diMe-G) is the break-set for T_1 RNAase, and all pyrimidines (C, U, T, ψ, Y, and E) the break-set for pancreatic nuclease. We define a "sufficient sequence" as any portion of the total sequence which contains at least one element from each break-set. Thus, GC is a sufficient sequence. Any portion of the total sequence containing a sufficient sequence such as GC will be broken by either of the enzymes going to completion.

The last fifty nucleotides in the proposed alanine S-RNA sequence and its isotomer are identical. In the first twenty-five nucleotides

$$_{p}\text{GGGC}\underset{\text{Me}}{\text{I}}\psi\text{GUGAU(CCU)GC(CCCUU)IGCG},$$

the sufficient sequence GC occurs three times. The isotomer

$$_{p}\text{GGGC(CCCUU)IGC}\underset{\text{Me}}{\text{I}}\psi\text{GUGAU(CCU)GCG}$$

which gives the same fragments (Table 1) is obtained by reversing the position of the two segments, (CCCUU)I and $\underset{\text{Me}}{\text{I}}\psi$GUGAU(CCU) which occur between the GC sequences.

Similarly, in Cantoni's serine S-RNA the isotomer given was obtained by interchanging the segments CGAGAGUAUCGGG and CUU occurring between the first three GC pairs, and in addition by interchanging the segments AAG and CCCGAUACUCUCG found between the last three GC pairs.

Other isotomers can be produced for both Holley's and Cantoni's sequences by similar exchanges involving other sufficient sequences such as GU. The isotomers produced above belong to the same general class. If we let W, X, Y, Z each re-

Table 1

All possible fragments produced by complete digestion with the available endonucleases of a segment of alanine S-RNA and its isotomer

Alanine S-RNA Segment			Isotomer		
GGGCI(Me)ψGUGAU(CCU)GC(CCCUU)IGCG			GGGC(CCCUU)IGCI(Me)ψGUGAU(CCU)GCG		
With T_1 RNAase					
G	UG		G	Me	
G	AU(CCU)G		G	CI(Me)ψG	
G	C(CCCUU)IG		G	UG	
Me	CG		C(CCCUU)IG	AU(CCU)G	
CI(Me)ψG				CG	
With Pancreatic RNAase					
GGGC	C	C	GGGC	U	GAU
Me	C	C	C	IGC	C
I(Me)ψ	U	U	C	Me	C
GU	GC	U	C	I(Me)ψ	U
GAU	C	IGC	U	GU	GC
		G			G

present any number (including zero) and sequence of bases, then WGCXGCYGCZ and WGCYGCXGCZ will be isotomers. A mathematical proof of this and also of the existence of other classes of isotomers is given elsewhere (Mosimann et al., in preparation).

If one can find in any proposed sequence the same sufficient sequence three times, and if upon interchanging the segments X and Y (as above) a different total sequence results, the sequence could not have been solved using only data obtained by complete digestion with the available endonucleases.

The general result just stated shows that isotomers could exist even if four mono-base-specific enzymes were available

for the digestion of an RNA polymer. In this case, a sufficient sequence would have to contain at least one of each base, e.g., CAUG. Under such conditions, CAUGAUCAUGGCAUG and CAUGGCAUGAUCAUG would be isotomers.

The proof of the existence of general classes of isotomers provides criteria by which one can examine original fragment data to determine whether or not a unique sequence can be derived from them. If the complete hydrolysis of an S-RNA by T_1 and pancreatic RNAase yields fragments among which are three different ones ending in the same sufficient sequence, such as GU, GC, etc. the sequence cannot be solved without additional information.

The most commonly used enzymes in the investigation of protein sequences are trypsin, pepsin, chymotrypsin, papain, and subtilisin. The break-set of trypsin is arginine and lysine, while leucine and tyrosine are members of the break-sets of the remaining enzymes. If three different fragments ending in the same sufficient sequence such as arginine-leucine occur with complete digestion by one of these enzymes, isotomers must exist. Any type of partial hydrolysis such as partial acid hydrolysis could, in principle, provide the additional information needed to solve the sequence by permitting the occurrence of fragments which span the sufficient sequences.

We have previously reported a method for reconstructing polymer sequences from fragment data by digital computer (Bradley, Merril, and Shapiro, 1964). This method has been applied to hypothetical fragments produced from nearly 800 random polymer sequences (Merril, Shapiro, Bradley, and Mosimann, in preparation). In all cases where a unique sequence was not obtained, it has been possible to show that the fail-

ure of the computer to reconstruct the original sequence was due to the existence of isotomers of the sequence. Only three classes of isotomers, including the one described herein, were observed in this study.

REFERENCES

Bradley, D. F., Merril, C. R., and Shapiro, M. B., Biopolymers, 2, 415 (1964).

Cantoni, G. L., Ishikura, H., Richards, H. H., and Tanaka, K., Cold Spring Harbor Symp. Quant. Biol., 28, 123 (1963).

Holley, R. W., Everett, G. A., Madison, J. T., Marquisee, M., and Zamir, A., Abstracts (I-S5) Sixth International Congress of Biochemistry, New York, July 26-Aug. 1 (1964).

Merril, C. R., Shapiro, M. B., Bradley, D. F., and Mosimann, J. E. In preparation.

Mosimann, J. E., Shapiro, M. B., Bradley, D. F., and Merril, C. R. In preparation.