Graph Theory / Algorithms-GC

Introduction

The genetic information is stored by DNA and RNA chains. An RNA chain is a sequence of bases whose length varies with the species. Each base can be one of the four: A = adenine, U = uracil, C = cytosine and G = guanine. The problem of knowing the composition of RNA chains is still quite open in general, as the sequences are extremely long.

One method of finding out the composition of an RNA chain involved fragmentating the chain into smaller pieces, and then reconstructing it back, as the chain can not be read as a whole. We are giving this method here.

We shall break/fragment the chain using two enzymes. One enzyme breaks the chain after every G. Let's call it the G-enzyme. Another enzyme, say U.C-enzyme, breaks the chain after every U and after every C. If we randomly put back together all the fragments of the chain, we obtain a number of possible solutions that grows exponentially with the number of bases in the chain. The method we are presenting here is better than that.

Reconstruction of some RNA chain (Mosimann, Hutchinson)

Given: two fragmentations of an unknown RNA chain obtained after applications of two enzymes (the order of the fragments is random):

1) G-enzyme: AUCG, AUG, G, CU, ACUAUACG.
2) U.C-enzyme: GGAC, U, AU, GAU, C, U, AC, GC, AU.

Find: initial RNA chain.

Solution:

We note that the fragment CU in 1) is "abnormal" (does not end in G) and, for this reason, it must be the last fragment in the chain.

Simulate treating each fragment in 1) with the U.C-enzyme and each fragment in 2) with the G-enzyme, which realizes a refragmentation of each fragment into **extended bases** (e.b.):

1-2) AU/C/G, AU/G, G, C/U, AC/U/AU/AC/G

2-1) G/G/AC, U, AU, G/AU, C, U, AC, G/C, AU

To find the first fragment in the RNA chain do:

A Consider all single (consisting of exactly one e.b.) fragments in 1-2) and 2-1) (these are the bold ones):

G, U, AU, C, *U*, AC, *AU*.

B Consider all interior extended bases (underlined) in 1-2) and 2-1):

C, U, AU, AC, G.

There always (why?) are two more e.b.-s in A than in B. These are exactly the beginning and the end of the RNA chain. In our case they are the e.b.-s in A. Since U is the end, AU must be the beginning.

Consider now all non-single fragments in 1-2) and 2-1) where the end fragment, in this case C/U, will be enlarged by adding to it the beginning e.b., in this case AU:

AU/C/G, AU/G, C/U*AU, AC/U/AU/AC/G, G/G/AC, G/AU, G/C.

Build a multidigraph where each fragment will become an arc from a vertex labeled with the beginning e.b. of the fragment to a vertex labeled with the end e.b. of the fragment; the interior of the fragment (if any) will be carried on the arc. Identify any vertices labeled with the same e.b. The beginning of the chain is at AU, which now labels a vertex, and the end of the chain at U which labels an arc that points to the vertex labeled AU.

There is a one-to-one correspondence between :

a) all directed circuits of the multidigraph starting at AU and ending at U, and going through each arc of the multidigraph exactly once (i.e. all Eulerian circuits starting at AU), and

b) all possible original RNA chains giving the 1) and 2) fragmentations.

The problem of reconstructing the RNA chain has been reduced to finding Eulerian circuits in a multidigraph. [see next page]

All possible original RNA chains obtained in this case are:

i) AUCGAUGGACUAUACGCU

ii) AUCGGACUAUACGAUGCU

iii) AUGAUCGGACUAUACGCU

iv) AUGGACUAUACGAUCGCU

PROJECT

Design an algorithm that reconstructs the RNA chain out of two sets of fragments:

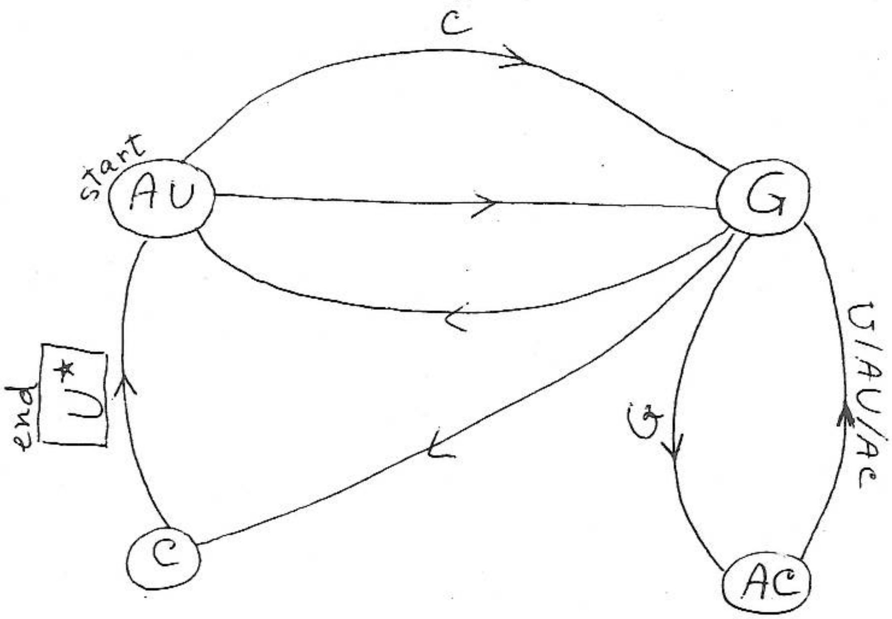
a) one set is obtained after applying the G-enzyme on one RNA chain, and

b) the other set is obtained after applying the U.C-enzyme on another, identical RNA chain.

Use the exposed method.

Presentation is of outmost importance. This includes appearance, logical thinking and persuasion.

Multigraph for given two fragmentations


C
start
AU
G
U/AU/AC
G
AC
C
end
U*