BUILDING A BRIDGE
FROM PROBLEMS OF MATHEMATICAL OLYMPIADS
TO OPEN PROBLEMS OF MATHEMATICS

by Alexander Soifer
University of Colorado at Colorado Springs
P. O. Box 7150, Colorado Springs, CO 80933, USA
asoifer@uccs.edu
http://www.uccs.edu/~asoifer/

I. Introduction: Envisioning the Bridge

The famous Russian mathematician Boris N. Delone once said, as Andrei N. Kolmogorov recalls in his introduction to [1], that "a major scientific discovery differs from a good Olympiad problem only by the fact that a solution of the Olympiad problem requires 5 hours whereas obtaining a serious scientific result requires 5,000 hours."

My books (Bibliography, 3-8) are bricks in building a bridge from problems of Mathematical Olympiads to problems of "real" mathematics. In them, I try to show that problems of competitions and research problems of mathematics stem from the same root, made of the same fabric, have no natural boundaries to separate them.

It is, therefore, natural to give "real" problems to young high school Olympians (maybe not at Mathematical Olympiads, as they do not last 5,000 hours :-). In fact, when the mid 1960s, as a high school student, I attended the Award Presentation Ceremony of the Moscow Mathematical Olympiad, the Chairman of the Olympiad's Jury and the famed mathematician Andrej Nikolaevich Kolmogorov paid us, young Olympians, an elegant compliment. "Perhaps, the only way to receive a proof of Fermat's' Last Theorem is to offer it at the Moscow Mathematical Olympiad," he said.

We ought to stop discrimination of young high school mathematicians based on their tender age. However, we can offer true research problem only to a small percent of high school mathematicians. What can we do for others, who are not yet ready for research? Here is one way. While reading or creating research mathematics, I caught myself many times thinking how beautiful, Olympiad-like certain ideas were. Consequently, some of these striking ideas gave birth to problems I created for the Olympiads. I first notice a fragment of the research, which utilizes a nice, better yet surprising idea. I then translate thus found mathematical gem into the language of secondary mathematics, and try to present it in a form of an engaging story – and a new Olympiad problem is ready!

Sometimes I imitate a "real" mathematical train of thought by offering at a Mathematical Olympiad a series of problems, increasing in difficulty, and leading to generalizations and deeper results. It is also important to realize that "real" mathematics cannot be reduced to just analytical reasoning, for to the tune of 50% mathematics is about construction of counterexamples. I try to reflect this dichotomy in our Olympiad problems, many of which require not only analytical proofs but also construction of examples.

The Bridge we are building can be walked in the opposite direction as well: it is worthwhile for professionals to take a deeper look at problems of Mathematical Olympiads. Those problems just might inspire exciting generalizations and new directions for mathematical research.

I will illustrate these ideas here in the context of a problem I have recently created for the Colorado Mathematical Olympiad.

II. An Olympiad Problem

My 1996 sabbatical leave I spent in several European countries, a good part of it at Charles University in the old part of Prague. While there, I attended a research number-theoretic talk by a young talented professor Martin Klazar on integral sequences. I enjoyed the talk, and took notes. When in early 2005 I came across these notes for the first time since taking them, I found a note to myself on the margin (yes, the margin again!): "use these ideas at the Olympiad!" Indeed, I put Martin's research ideas into the foundation of the part (b) of hardest problem 5 of the 22nd Colorado Mathematical Olympiad. Below you can see the logo of CMO-22.

Logo of the 22nd Colorado Mathematical Olympiad, April, 2005

Let us look at this problem and its solution.

2005.5. Love and Death

(a) The DNA of bacterium *bacillus anthracis* (causing anthrax) is a sequence, each term of which is one of 2005 genes. How long can the DNA be if no consecutive terms may be the same gene, and no two distinct genes can reappear in the same order? That is, if

distinct genes α, β occur in that order (with or without any number of genes in between), the order α, …, β cannot occur again.

(b) The DNA of bacterium *bacillus amoris* (causing love) is a sequence, each term of which is one of 2005 genes. No three consecutive terms may include the same gene twice, and no three distinct genes can reappear in the same order. That is, if distinct genes α, β, and γ occur in that order (with or without any number of genes in between), the order α, …, β, …, γ cannot occur again. Prove that this DNA is at most 12,032 long.

First Solution. Let us prove that in a DNA satisfying the two given conditions, there is a gene that occurs only once. Indeed, let us assume that each gene appears at least twice and for each gene select the first two appearances from the left and call them a *pair*. The first gene from the left is in the first pair. This pair must be separated, thus the pair of the second gene from the left is nestled inside the first pair. The second pair must be separated, and thus the pair of the third gene from the left must be nestled inside the second pair, etc. As there are finitely many genes, we end up with a pair of genes (nestled inside other pairs) that is not separated, a contradiction.

We will now prove by mathematical induction on the number n of genes that the DNA that satisfies the conditions and uses n genes is $2n - 1$ gene long. For $n = 1$ the statement is true, as longest DNA is $2 - 1 = 1$ gene long.

Assume that a DNA that satisfies conditions and uses n genes is at most $2n - 1$ gene long. Now let S be a DNA sequence satisfying conditions that uses $n + 1$ genes; we need to prove that it is $2(n + 1) - 1 = 2n + 1$ gene long.

By the starting paragraph, there a gene g that occurs only once in S; we throw it away. The only violation that this throwing may create is that two copies of another gene are now adjacent – if so, we throw one of them away too. We get the sequence S' that uses only n genes. By inductive assumption, S' is at most $2n - 1$ gene long. But S is at most 2 genes longer than S', i.e., S is at most $2n + 1$ gene long. The induction is complete.

All that is left is to demonstrate that the DNA length of $2n - 1$ is attainable. But this is easy: just take a sequence 1, 2, …, $n - 1$, n, $n - 1$, …, 2, 1. ■

Second Solution. We will now prove by mathematical induction on the number n of genes that the DNA that satisfies the conditions and uses n genes is $2n - 1$ gene long. For $n = 1$ the statement is true, as longest DNA is $2 - 1 = 1$ gene long.

Assume that for any positive integer k, $k < n$, a DNA that satisfies the conditions and uses k genes, is at most $2k - 1$ gene long. Now let S be the longest DNA sequence that satisfies conditions and uses n genes; we need to prove that S is at most $2n - 1$ gene long.

Let the first gene of S be 1, then the last term must be 1 as well, for otherwise we can make *S* longer by adding a 1 at the end. Indeed, assume that the added 1 has created a forbidden DNA. This means that we now have a subsequence *a*,…,1, …, *a*, …, 1 (with the added 1 at the end); but then the original DNA has already had the forbidden subsequence 1, …, *a*,…,1, …, *a*.

Case 1. If there are no more 1's in the DNA, we throw away the first 1 and the last 1, and we get a sequence S' that uses $n - 1$ genes (no more 1s). By inductive assumption, S' is at most $2n - 1$ genes long. But S is 2 genes longer than S', i.e., S is at most $2n - 1$ genes long.

Case 2. Assume now that there is a 1 between the first 1 and the last 1. The DNA then looks as follows: 1, S', 1, S", 1. Observe: if a gene *m* appears in the sequence S', it may not appear in the sequence S", for this would create the prohibited subsequence 1, …, *m*, …, 1, …, *m*. Let the sequence 1, S', 1 use *n'* genes and the sequence 1, S", 1 use *n"* genes. Obviously, $n' + n'' - 1 = n$ (we subtract 1 in the left side because we counted the gene 1 in each of two subsequences!). By inductive assumption, the length of the sequences 1, S', 1 and 1, S", 1 are at most $2n' - 1$ and $2n'' - 1$ respectively. Therefore, the length of S is $(2n' - 1) + (2n'' - 1) - 1$ (we subtract 1 because the gene 1 between S' and S" has been counted twice). But $(2n' - 1) + (2n'' - 1) - 1 = 2(n' + n'') - 3 = 2(n + 1) - 3 = 2n - 1$ as desired. The induction is complete.

This proof allows us to find richer set of examples of DNAs of length of $2n - 1$ (and even describe all such examples if necessary). For example:

$$1, 2, \ldots, \mathbf{k}, k + 1, k, k + 2, k, \ldots, k, 2005, \mathbf{k}, k - 1, k - 2, \ldots, 2, 1. \blacksquare$$

Solution of problem 5(b). Assume *S* is the longest DNA string satisfying the conditions. Partition *S* into blocks of 3 terms starting from the left (the last block may be incomplete and have fewer than 3 terms, of course). We will call a block *extreme* if a number from the given set of genes {1, 2, …, 2005} appears in the block for the first or the last time. There are at most 2×2005 extreme blocks.

We claim that *there cannot be any complete (i.e., 3-gene) non-extreme blocks*.

Indeed, assume the block *B*, which consists of genes α, β, γ in *some* order, is not extreme. This means that the genes α, β, γ each appears at least once before and once after appearing in *B*. We will prove that then the DNA would contain the forbidden subsequence of the type $\alpha, \beta, \gamma\ \alpha, \beta, \gamma$. Let *A* denote the ordered triple of the first appearances of α, β, γ (these 3 genes may very well come from distinct 3-blocks). Without loss in generality we can assume that in *A* the genes α, β, γ appear in *this* order. Let *C* denote the ordered triple of the last

appearances of α, β, γ in *some* order. Let us look at the 9-term subsequence *ABC* and consider three cases, depending upon where α appears in the block *B*.

1. if α is the first gene in *B* (Fig. 5.1), then we can choose β also in *B* and γ in *C* to form α, β, γ which with α, β, γ from A gives us the forbidden $\alpha, \beta, \gamma\ \alpha, \beta, \gamma$.

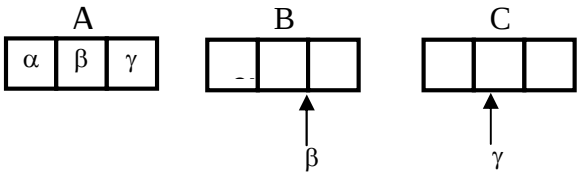

Fig. 5.1

2. let α be the second gene in *B* (Fig. 5.2). If β follows α then with a γ from *C* we get α, β, γ, which with α, β, γ from *A* produces the forbidden $\alpha, \beta, \gamma\ \alpha, \beta, \gamma$. Thus, β must precede α in *B*. If the order of the genes β, γ in *C* is β, γ, then we can combine an α from *B* with this β, γ to form α, β, γ, which with α, β, γ from *A* gives us the forbidden $\alpha, \beta, \gamma\ \alpha, \beta, \gamma$. Thus, the order in *C* must be γ, β. We can now choose α, γ from *A* followed by β, α from *B*, followed by γ, β from *C* to get $\alpha, \gamma, \beta\ \alpha, \gamma, \beta$, which is forbidden.

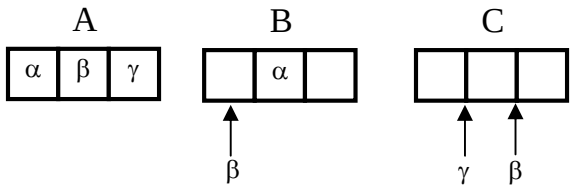

Fig. 5.2

3. let α be the third gene in *B* (Fig. 5.3), and is thus preceded by a β in *B*. If the order in *C* is β, γ, then we get α, β, γ from *A* followed by α from *B* and β, γ from *C* to get the forbidden $\alpha, \beta, \gamma\ \alpha, \beta, \gamma$. Thus, the order in C must be γ, β, and we choose α, γ from *A*, followed by β, α from *B*, and followed by γ, β from *C* to form the forbidden $\alpha, \gamma, \beta, \alpha, \gamma, \beta$.

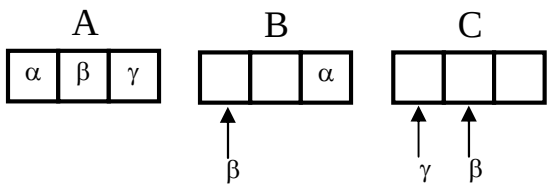

Fig. 5.3

We are done, for the DNA sequence consists of at most 2×2005 extreme 3-blocks plus perhaps an incomplete block of at most 2 genes – or 12,032 genes at the most. ■

III. Crossing the Bridge into Research

In problem 5(a) we obtained the exact result, the maximum length of the DNA of bacterium *bacillus anthracis*. One cannot do better. The problem 5(b), however, produced only an upper bound for the length of the DNA of bacterium *bacillus amoris*. Can we obtain the exact result? Can we at least reduce the upper bound?

"Yes we can!" as the most inspiring USA Presidential candidate in a generation, Barak Obama says. Let us allow the 3-gene blocks to overlap by their end terms. We can then use the same argument as in the solution of 5(b) above, and reduce the upper bound from $6n + 2$ (n is here number of available genes) to $4n + 2$. This is exactly what Martin Klazar of Charles University, Prague, presented in his 1996 talk.

Can we do better? Yes, with clever observation of the starting and ending triples Klazar was able to reduce the upper bound to $4n - 4$. In his 1996 talk, Martin even claimed the bound of $4n - 7$, proof of which required further cleverness. However, the problem of finding the exact maximum length of DNA remains open.

Open Problem. The DNA of bacterium *bacillus amoris* (causing love) is a sequence, each term of which is one of n genes. No three consecutive terms may include the same gene twice, and no three distinct genes can reappear in the same order. That is, if distinct genes α, β, and γ occur in that order (with or without any number of genes in between), the order α, …, β, …, γ cannot occur again. Find the maximum length the DNA of bacterium *bacillus amoris* may have.

Warnings.

1) While bacterium *bacillus anthracis* (causing anthrax) does exist, I claim no knowledge of its structure. It has been simply used to captivate the imagination of the Olympians.
2) The existence of bacterium *bacillus amoris* (causing love) has not been established by science.

BIBLIOGRAPHY.

1. Gal'perin G. A., and Tolpygo, A. K., *Moscow Mathematical Olympiads*, edited and with introduction by A. N. Kolmogorov, Prosvetshenie, Moscow, 1986 (Russian).

2. Soifer, A., *Mathematics as Problem Solving,* Center for Excellence in Mathematical Education, Colorado Springs, CO, 1987.

   Second edition to appear in Springer, New York, in 2008.

3. Soifer, A., *How Does One Cut a Triangle?* Center for Excellence in Mathematical Education, Colorado Springs, CO, 1990.

   Second edition to appear in Springer, New York, 2008.

4. Boltyanski, V., and Soifer, A. *Geometric Etudes in Combinatorial Mathematics,* Center for Excellence in Mathematical Education, Colorado Springs, CO, 1991.

   Second edition to appear in Springer, New York, 2009.

5. Soifer, A., *Colorado Mathematical Olympiad: The First Ten Years and Further Explorations*, Center for Excellence in Mathematical Education, Colorado Springs, 1994.

6. Soifer, A., *Colorado Mathematical Olympiad: The First Twenty Years and Further Explorations*, Springer, New York (to appear in 2009).

7. Soifer, A., *Mathematical Coloring Book: Mathematics of Coloring and the Colorful Life of Its Creators*, Springer, New York (to appear in 2008-9).

8. Erdös, P., and Soifer, A., *Problems of pgom Erdös* (to appear in 2010).