LETTER TO THE EDITOR

The Mysterious Case of the Reverse Sequences

Gabriel Gutiérrez*

Departamento de Genética, Universidad de Sevilla, Spain

I would like you to do a little experiment with me. Let’s suppose we get a bacterial SSU rRNA and we blast it at the NCBI [1]. The result we get is obvious: a large number of homologous sequences of many bacteria. Now let's do something else, we are going to reverse the same sequence. It is easy to do with the EMBOSS program *Revseq* [2]. The reverse sequence is written in the direction 3'-5'. For example 5’AGTC3’ becomes 3’CTGA5’. If we blast with the reverse sequence, what would we get? The answer is nothing. We should not get any significant result because the reverse sequence is meaningless, sequences are always written in the 5 '-3' direction in the databases. Well, have you already done the blast with the reverse sequence? You are surprised because there are significant results! How is it possible?

By this method I have detected more than 300 SSU rRNA reversed sequences in the NCBI non-redundant nucleotide database (Table **1**). I think researchers who sent the sequences were wrong and sent the sequences in 3'-5' direction, but why?

One explanation is that people make mistakes. The rRNA sequences are commonly used in identification and phylogeny of species. It is unavoidable, If you do one thing many times, sooner or later you make a mistake. This may explain the abundance but not why people makes mistakes.

I want to propose two explanations that are not mutually exclusive. The first one says that it depends on the language of the researcher who submitted the sequence to the database. Genetics was developed in western culture, where texts are read and written horizontally left-to-right (LFT), for that reason we write DNA sequences in the 5’-3’ direction, the 5’ end on the left, the 3’ end on the right. Some oriental languages are written right-to-left (RTL). Muslims and jews write horizontally in the RTL direction. The chinese and japanese write top-to-bottom (TTB) but they can both read the columns LFT or RTL. My guess is that researchers who mother-tongue is not LFT use RTL and/or TTB text editors that change the direction of the sequences. Table **1** shows the places of origin of the sequences, 57% of them belong to non-LFT countries. I have not included the sequences from India, given that Hindi or other official languages are LFT but others, like Urdu, are RTL.

It is striking that 22 eubacterial and 26 eukaryote reverse sequences come from USA and Australia respectively. I have checked these database entries and 6 out of 25 from the USA sequences were submitted to the database by Japanese researcher working in American labs. One more was submitted by a Chinese researcher working in a American lab. Moreover, 25 out of 26 Australian sequences were submitted by a Singaporean researcher working in an Australian lab. Singapore official languages are English (LTR), Malay (LTR), Tamil (LTR) and Chinese (TTB). Archea sequences from UK/Pakistan deserve special attention. These 4 sequences were the result of a scientific collaboration of English (LFT) and Pakistani (RTL) researchers. The reversing of a sequence can be produced by just one of the researchers manipulating it with a text editor adjusted for a RTL script language. If we include the sequences from India and the special cases of USA, Australia and UK/Pakistan this explanation covers the 68% of the SSU reverse sequences.

The rest of the sequences, most of them from western LFT languages, can be explained by a second hypothesis. I teach bioinformatics and many students tend to confuse reverse and complementary sequences. The EMBOSS program *Revseq* can help us demonstrate this confusion, the program gives the reverse, the complement and the reverse-complement version of a sequence. The reverse sequence of 5’AGTC3’ is 3‘CTGA’5, its complement is 3’TCAG5’, both of them are complementary and have no meaning. The only sequence with the same meaning that the original sequence is the reverse-complement one, i.e. 5’GACT3’. Many students, when using the *Revseq* program, only select the reverse or the complement options but not both options at the same time. My hypothesis is that most of the western reverse sequences sent to the databases are produced when a researcher, using any of the available sequence editors, wants to convert a sequence in its complementary and reverses the sequence but does not complement it or vice versa.

Today is difficult to know how many reverse sequences of other genes have been submitted to the databases. For example, I have searched the database with a reverse version of a dolphin *cytochrome b* and found 128 homologous reverse sequences all of them from the same UK lab, no researcher of RTL language was found among them. Regard-

*Address correspondence to this author at the Departamento de Genética, Universidad de Sevilla, Spain; Tel: +34954557112; Fax: +34954557104; E-mail: ggpozo@us.es

Table 1. Reverse Sequences of the NCBI *nr* Nucleotide Database Detected by Blastn

Domain	Country	Number of Reversed SSU RNA Sequences	Script Direction
Eubacteria	China	139	TTB,LTR or RTL
	USA	22	LTR
	Japan	21	TTB,LTR or RTL
	Switzerland	18	LTR
	The Netherlands	13	LTR
	India	18	LTR or RTL
	UK	9	LTR
	Spain	8	LTR
	Belgium	5	LTR
	Denmark	4	LTR
	Germany	2	LTR
	Pakistan	2	RTL
	New Zealand	2	LTR
	Kenya	2	LTR
	Sri Lanka	1	LTR
	Cuba	1	LTR
	Greece	1	LTR
Eukaryota			
	Australia	26	LTR
	Japan	6	TTB,LTR or RTL
	China	4	TTB,LTR or RTL
	USA	2	LTR
	Malaysia	1	LTR
	Iran	1	RTL
	New Zealand	1	LTR
Archea			
	UK/Pakistan	4	LTR/RTL
	China	2	TTB,LTR or RTL
	Germany	1	LTR
Total		316	

E value threshold 0.001. The sequences were detected using a reverse version of the SSU rRNA of *Escherichia.coli* (CP000948), yeast (J01353) and *Sulfolobus solfataricus* (X90483). Searches were made by February 2010.

less of the cause of the errors, just have to say that the curators of the databases would correct these sequences or ask the authors to do it. The percentage of reverse sequences in the database should not be too high, but correction is necessary because biological information is lost, since a homology based search can not detect them.

REFERENCES

[1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215: 403-10.

[2] Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology. Open Software Suite. Trends Genet 2000; 16: 276-7.

Received: July 13, 2011 Revised: September 22, 2011 Accepted: September 22, 2011