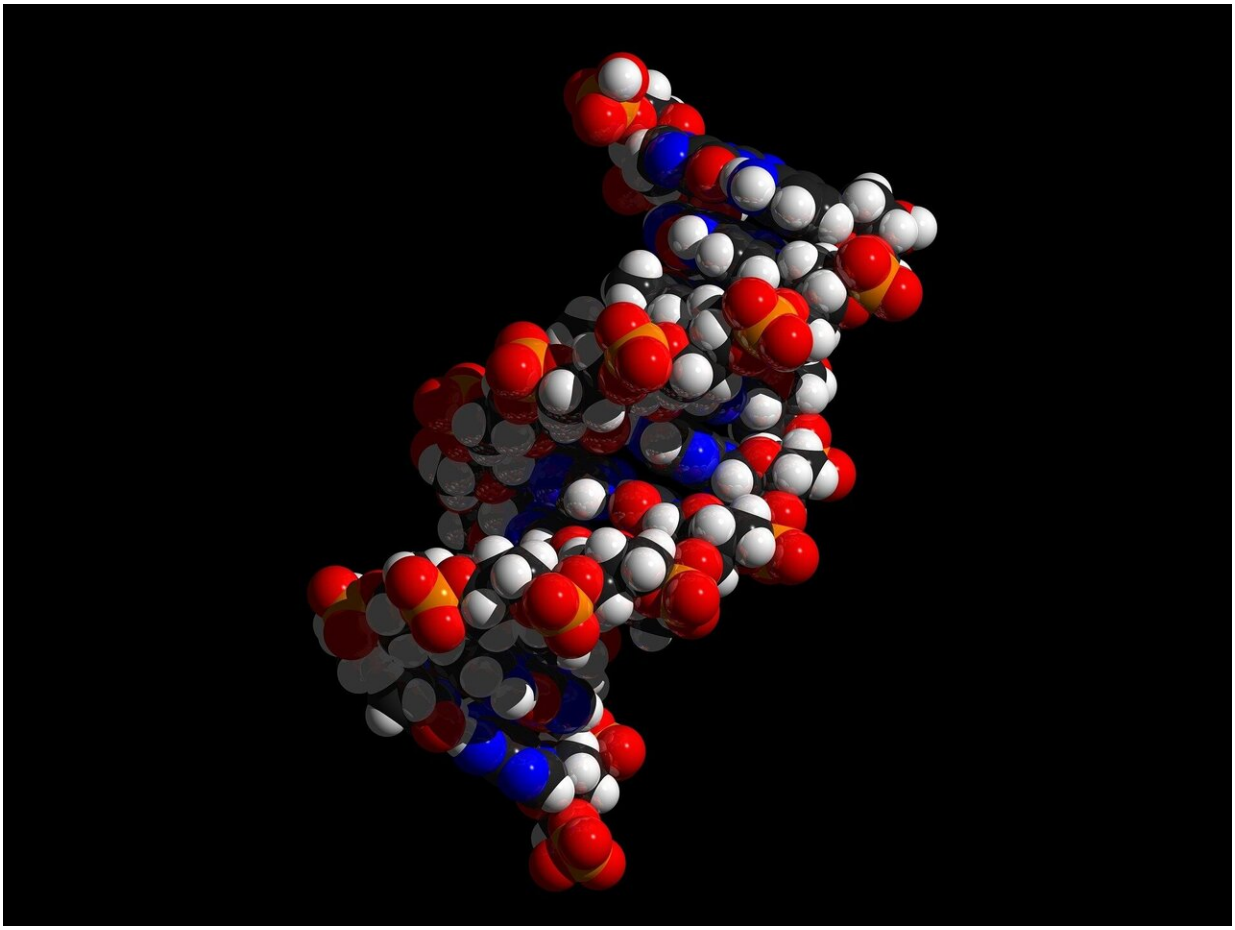
# Statistical tool finds 'gaps' in DNA data sets shouldn't be ignored

August 16 2022, by Tracey Peake



Credit: CC0 Public Domain

A simple statistical test shows that contrary to current practice, the

"gaps" within DNA protein and sequence alignments commonly used in evolutionary biology can provide important information about nucleotide and amino acid substitutions over time. The finding could be particularly relevant to those studying distantly related species. The work appears in *Proceedings of the National Academy of Sciences*.

Biologists studying evolution do so by looking at how DNA and protein sequences change over time. These changes can be sequence length changes—when specific nucleotides are deleted or added at certain positions—or substitutions, where one nucleotide type is exchanged for a different type at a given point.

"Think of the DNA sequence and its evolution as a sentence being copied by different people over time," says Jeff Thorne, professor of biological sciences and statistics at NC State and a co-corresponding author of the research. "Over time, a letter in a word will change—that's a substitution. Leaving out or adding letters or words correspond to deletions or insertions."

The first step analysts usually perform when looking at evolutionary DNA changes is to construct a sequence alignment. This means figuring out how all of the sequences correspond to one another and then aligning those corresponding positions into columns for comparison. Due to substitutions, insertions and deletions, however, nucleotide types within columns can vary among sequences, or be absent altogether. When a sequence does not have a corresponding nucleotide, a gap is placed in the alignment column for that sequence.

"Conventionally, when using sequence alignments to do analyses, the gaps within alignment columns are treated as missing data that provide no information about the substitutions," Thorne says. "Historically, the research community has assumed that gap locations are independent of the substitution process. But what if that assumption is incorrect?"

Thorne and his colleagues created a simple statistical test to assess whether gap locations are independent of the amino acid replacement process. They tested 1390 different sets of sequence alignments, and found that in roughly two-thirds of the sets, the usual assumption of independence between gap locations and amino acid replacement was rejected.

"One possibility is that gap locations provide useful information about the amino acid replacement process," Thorne says. "If so, evolutionary biologists should develop better techniques for extracting this information."

The research also illustrated how the usual approach of constructing a sequence alignment and then basing evolutionary conclusions on that single optimal alignment can be problematic. What if the alignment is wrong? Even worse, what if the alignment is biased?

For example, if substitutions occur more often than gaps, then researchers tend to repeatedly choose substitutions over gaps when building the sequence alignment and the resulting alignment can contain too few gaps overall. And while those little errors in alignments between closely related species will most likely not affect outcomes, over time—and particularly in comparisons between diverse species—that bias can create error that could affect subsequent analyses.

"Sometimes our best guesses are biased," says Tae-Kun Seo, principal research scientist at the Korea Polar Research Institute and co-corresponding author of the research. "There's no simple solution, but hopefully this study will help us be mindful about potential pitfalls. We need to be aware of the problems with conventional statistical methods and work toward fixing them."

Ben Redelings, research scientist at Duke University and the University

of Kansas, also contributed to the work.

**More information:** Correlations between alignment gaps and nucleotide substitution or amino acid replacement," *Proceedings of the National Academy of Sciences* (2022). DOI: 10.1073/pnas.2204435119

Provided by North Carolina State University