SIAM Review
Vol. 49, No. 1, pp. 3–31

Lior Pachter†
Bernd Sturmfels†

The Mathematics of Phylogenomics*

CONJECTURE 1 (the "Meaning of Life"). *The sequence of* 42 *bases*

(2.1) TTTAATTGAAAGAAGTTAATTGAATGAAAATGATCAACTAAG

was present in the genome of the ancestor of all vertebrates, and it has been completely conserved to the present time (i.e., none of the bases have been mutated, nor have there been any insertions or deletions).

The identification of such a sequence requires a highly nontrivial computation: the alignment of ten genomes (including mammalian genomes close to 3 billion bases in length) and subsequent analysis to identify conserved orthologous regions within the alignment [63]. Using the tools described in section 8, one checks that the sequence (2.1) is present in all ten genomes. For instance, in the human genome (May 2004 version), the sequence occurs on chromosome 7 in positions 156501197–156501238. By examining the alignment, one verifies that, with very high probability, the regions containing this sequence in all ten genomes are orthologous. Furthermore, the implied claim that (2.1) occurs in all present-day vertebrates can, in principle, be tested.

Identifying and analyzing sequences such as (2.1) is important because they are highly conserved yet often nongenic [7]. One of the ongoing mysteries in biology is to unravel the function of the parts of the genome that are nongenic and yet very conserved. The extent of conservation points to the possibility of critical functions within the genome. Recent studies have pointed to the association of highly conserved elements with developmental genes [48, 62].

In 2003, the sequence (2.1) appeared to be the longest completely conserved sequence among the vertebrates. We were amused to find that its length was 42. In light of [1], it was decided to name this DNA sequence "The Meaning of Life." It may be a coincidence that the segment above contains two copies of the *motif* TTAATTGAA, but this motif may also have some function (for example, it may be bound by a protein). Indeed, the identification of such elements is the first step toward understanding the complex regulatory code of the genome.

[1] D. N. ADAMS, *The Hitchhiker's Guide to the Galaxy*, Pan Books, London, 1979.