

# SCANS: System for Compression and Analysis of Nucleotide Sequences

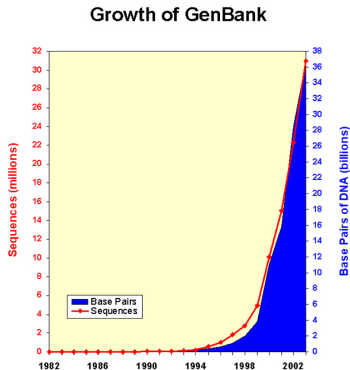
Valerio G. Aimale

SeiraD, Inc. Santa Fe, NM

Sunday, December 11th, 2005

# DNA databases—the present day

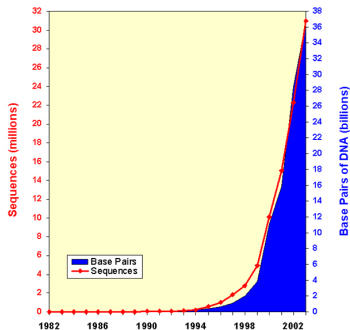
- GenBank doubling every 9-12 months
- 2003 size equal to just six human genomes



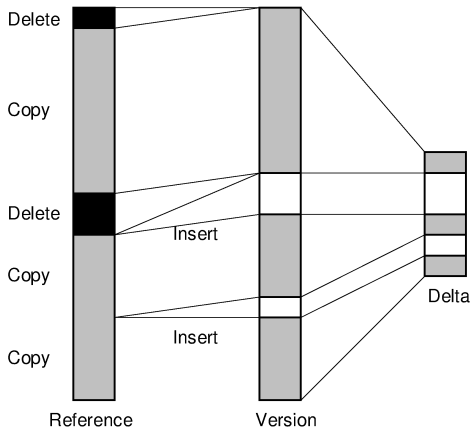
# DNA databases—the present day

- GenBank doubling every 9-12 months
- 2003 size equal to just six human genomes
- Data volume is growing fast
- Data volume still very small

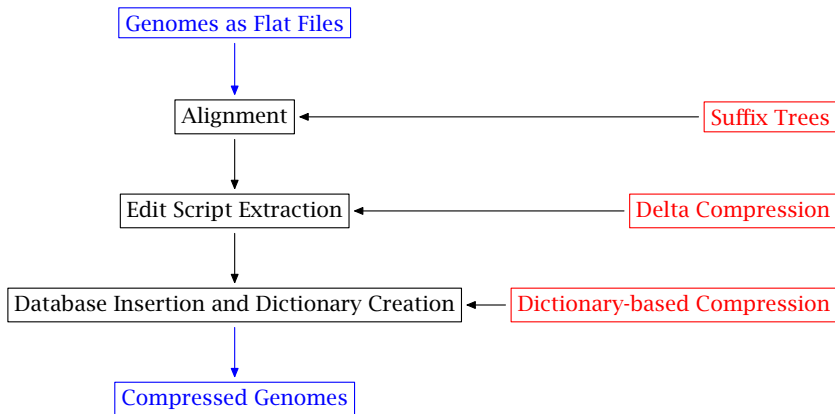
Growth of GenBank



# Delta compression exploits redundancy



# Integration of the three technologies



# Integrated technologies: do they work?

<b>Compression Approach</b>	<b>Compression Ratio (bits/base)</b>
SCANS flat edit scripts	0.0168
SCANS mySQL IC3	0.0340
Biocompress-2	1.7837
GenCompress	1.7434
CTW+LZ	1.7389
DNACompress	1.7254