Papers

- Matsumoto *et al*, 2000. **Biological sequence compression algorithms.**
- Chen *et al*, 2002. **DNACompress: fast and effective DNA sequence compression.**
- Christley *et al*, 2009. **Human genomes as email attachments.**
- Baldi *et al.* **Data structures and compression algorithms for high-throughput sequencing technologies.** BMC Bioinformatics, 2010.
- Birney *et al.* **Efficient storage of high throughput sequencing data using reference-based compression.** Genome Research, 2011.

Questions

1. What are the stated applications of each tool? How broadly/narrowly focused are they?
2. What codes are used? What kind of redundancy or pattern is being compressed by these codes?
3. Can you go as far as identifying the probability distribution that these codes are (near-)optimal for?
4. Does the paper describe a proof-of-concept, a prototype implementation, or a ready-to-use software package?
5. What compression ratio is claimed (if any)? Are theoretical limits discussed?
6. Does the paper describe benchmarks? If so, what? Are other programs compared, or any standard metrics established?
7. If the codec is statistical, or otherwise involves parameterization from a training set, what corpus was used to optimize it?
8. Do the papers cite each other? Do they discuss similarities or differences between each other? Can you make any additional comparisons between the papers?
9. Can you think of any redundancies or patterns in the datasets-to-be-compressed that would be missed by these approaches?
10. Are there any other notable aspects or results of the papers?