Dr Tomkins Falls Victim to Software Bug in Chimpanzee Comparison

Glenn Williamson

Sydney, Australia

Abstract

This article is in response to a study concerning the genetic similarity between humans and chimpanzees, authored by Dr Jeff Tomkins. Dr Tomkins performed a comprehensive comparison of the chimpanzee genome to the human genome and reports an overall similarity of 70%.

In this paper I carefully reproduce a subset of Dr Tomkins' results, and show clearly and unambiguously that Dr Tomkins has fallen victim to a serious bug in the software used to obtain his results. It is this bug that causes Dr Tomkins to report the erroneous figure of 70% similarity. After correcting for both the effects of this bug and some non-trivial errors in Dr Tomkins' methodology, I report an overall similarity of 96.90% with a standard error of ±0.21%. This figure includes indels, and the result is largely in line with the secular scientific consensus.

It is quite likely that Dr Tomkins' has fallen foul of this bug in many of his previously published studies. It is now incumbent on him to perform these comparisons again and restate his results.

Keywords

Human Chimpanzee, DNA, Genetic Similarity, BLAST, Tomkins.

Introduction

The genetic similarity between humans and chimpanzees is an important topic amongst the creationist scientific community. According to evolutionists, humans and chimpanzees share somewhere between 95% and 99% of their DNA, and this result is often touted as evidence of their common ancestry.

Dr Tomkins, on the other hand, claims that the secular scientific community is in error, and that a more reasonable figure for overall genetic similarity is around 70% [1]. To support his claim, Dr Tomkins has performed his own comparisons and, with his colleague Dr Jerry Bergman, has re-evaluated the secular scientific literature on the subject [2]. Dr Tomkins also cites two non-peer-reviewed sources in support of his position [3] [4].

Materials and Methods

All three experiments presented in this paper are comparisons of the chimpanzee genome to the human genome using the BLAST+ suite of programs (downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/). Version 2.2.27 was used for the first two experiments, since it is the same version as that used by Dr Tomkins in his comprehensive comparison. Version 2.2.29 was used for the third experiment.

Batch vs. Serial Experiment

The first experiment was designed to replicate Dr Tomkins' methodology as closely as possible in order to reproduce his results. This meant using human genome assemblies from February 2009 (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/) and chimpanzee genome assemblies from March 2006 (ftp://hgdownload.cse.ucsc.edu/goldenPath/panTro2/chromosomes/).

Similar to Dr Tomkins, I have sliced the individual chimpanzee chromosomes into small pieces and compared them against the corresponding human chromosome. This experiment is designed to show how different results can be obtained based on how these query sequences are submitted to the BLAST+ algorithm. The user can choose to submit many query sequences in a single file (hereafter called the "batch" method), or the user can choose to submit only one query sequence at a time - each time launching a new BLAST+ process (hereafter called the "serial" method). Submitting only one query sequence at a time is a horribly inefficient method, since each time the BLAST+ executable is launched, it will need to load the subject sequences (in this case, the corresponding human chromosome sequence) from disk into memory.

To illustrate by way of example, human chromosome 1 (and its chimpanzee counterpart) contain approximately 225Mbp of sequenced DNA. If we were to compare this DNA in slices that are 350 base pairs long using the serial method, we would need to launch the BLAST+ executable approximately 643,000 times – each time loading the 225Mbp chromosome file from disk into memory. Accordingly, processing times for the serial method are easily an order of magnitude greater than processing times for the batch method.

For this reason, I have chosen to limit the experiment to the two smallest chimpanzee chromosomes (chromosomes 21 and 22 each have approximately 34Mbp of sequenced DNA), purely to show the different results obtained using the serial method as opposed to the batch method. I have used the "*optimised slice siz*e" from Dr Tomkins' paper, and sliced chimpanzee chromosome 21 into sequences 500 base pairs long; chimpanzee chromosome 22 has been sliced into sequences 450 base pairs long.

From Dr Tomkins' paper, and from personal communications with him [5], I have been able to reconstruct the BLAST command used to perform the alignments reported in his paper, detailed below:

```
blastn -task blastn
-query query.22.fasta -db human.22 -out batch.22.csv \
-outfmt '10 pident nident length qlen' \
-word_size 11 \
-evalue 10 \
-max_target_seqs 1 \
-dust no \
-soft_masking false \
-ungapped \
-num_threads 8
```

Of particular interest here is the `outfmt` parameter. This parameter controls which fields are returned from a BLAST query [6], and for his study, Dr Tomkins has chosen only four: `pident`, `nident`, `length`

and `qlen`. Relevantly, Dr Tomkins did not include any of the fields that could be used to identify either the query sequence or the subject sequence (`qseqid`, `qstart`, `qend`, `sseqid`, `sstart` and `send`).

Gapped vs Ungapped Experiment

The `ungapped` parameter determines whether to account for small indels in the comparison. If the `ungapped` parameter is used, and there is a putative single nucleotide insertion in one of the sequences, then the BLAST algorithm cannot continue the alignment.

For example, suppose we are comparing the following strings of DNA:

```
Query       GTCGTAATGATTA
            ||||||
Subject     GTCGTATGATTAC...
```

Obviously the first six nucleotides are identical, but there seems to be an extra 'A' in the query sequence which prevents the alignment from continuing any further. If the BLAST algorithm is unconstrained by the `ungapped` parameter, it is clever enough to insert a gap into the subject sequence, and that gap represents a putative insertion or deletion. So we now have:

```
Query       GTCGTAATGATTA
            |||||| ||||||
Subject     GTCGTA-TGATTAC...
```

The first example corresponds to `ungapped` behaviour, and will report only 6 identical nucleotides in a query 13 nucleotides long (46% identical). The second example allows for small insertions and deletions, and will report that 12 out of the 13 nucleotides match (92% identical). Since Dr Tomkins is critical of studies that allegedly do not take indels into account [2], it is quite peculiar that he has failed to do so in his own study.

This second experiment follows on from the first experiment and compares the results of two chromosomes (21 and 22) both with and without the `ungapped` parameter.

Statistical Sampling Experiment

After accounting for the two major factors that contributed to Dr Tomkins' erroneous result, I now perform a statistical comparison of the entire chimpanzee genome to the entire human genome. This third experiment is independent of the first two experiments in that I do not constrain myself to outdated versions of the chimpanzee and human genomes, nor do I constrain myself to an older version of the BLAST+ software. Non-repeat masked chromosomes files in FASTA format were downloaded from Ensembl:

> ftp://ftp.ensembl.org/pub/release-76/fasta/homo_sapiens/dna/
>
> ftp://ftp.ensembl.org/pub/release-76/fasta/pan_troglodytes/dna/

To accommodate the secular scientific consensus view, I have concatenated chimpanzee chromosomes 2A and 2B into a single chromosome. I have chosen a slice length of 300 base pairs irrespective of the particular chromosome being compared. I have chosen this length since it is both the median and the mode of what Dr Tomkins considers to be the "*optimized slice size*" [1] for each chromosome.

Also, rather than comparing each and every slice of 300 base pairs, I have written a custom Perl script (available on request) that uses Perl's built-in random number generator to choose 10,000 unique slices from each chromosome, ensuring that no slices contain *"gap-filling 'N's"*. I then report an average similarity with a corresponding Standard Error (at the 99% Confidence Level). The formula to determine the Standard Error is:

$$SEM = 2.5758 \times \frac{s}{\sqrt{n}}$$

Where:

- SEM Standard Error of the Mean
- 2.5758 Z-score for the 99% Confidence Level
- s Standard Deviation of the Sample
- n Sample Size.

It is important to note that the size of the sample relative to the size of the genome is not a relevant factor to the accuracy of the results. While only 2.4% of the genome is compared, it is the absolute number of samples (that is, 10,000 queries per chromosome) that determines the statistical accuracy [7].

Results

Before discussing the results of the individual experiments, a discussion on how Dr Tomkins arrives at his final result is in order. From personal communications with Dr. Tomkins [5] I have been able to ascertain how he calculates his figures of ~70% and have reproduced the relevant portion below:

```
Ave % identity align: 98.6896080481
Ave alignment length: 294.111953678
Ave query seq length: 300.0
Number of query seqs: 437716
Number of query hits: 315702
Ave % hit frequency : ~72.12484807500753913496
Ave % query identity: ~96.76306347966666666667
Overall % DNA identity : ~69.79021252743268694770
```

The results above are for chromosome 10, using sliced human query sequences against the chimpanzee genome. Dr. Tomkins gives no indication that these results are anomalous and therefore, I will presume that Dr. Tomkins' result above is similar to his results for all the autosomes. In summary, Dr. Tomkins returns a very high level of identity for those sequences that do return hits (98.7%), those alignments span a very high percentage of the query sequence (98%), but claims that only 72% of query sequences return hits at all.

Batch vs. Serial Experiment

This experiment was designed purely to show the different results obtained by submitting queries using the batch method as opposed to the serial method. The results are shown in Table 1. Using Dr Tomkins' parameters and methodology, I have been able to closely match his results for chimpanzee chromosomes 21 and 22.

What is most striking however – and the key finding of this paper - is that when queries are submitted in batch mode, version 2.2.27 of the BLAST algorithm fails to return a significant percentage of matches. For chimpanzee chromosomes 21 and 22, BLAST returns hits for approximately 88% of the sequences. When the queries are submitted serially, a full 100% of queries find a match. This errant behavior is confirmed by NCBI, since in a subsequent release they claim to have fixed a bug that resulted in "*missing hits when running blastn with **multiple queries**, word size 7, **large evalue**, and **no low complexity filtering***" (emphasis added) [8].

The existence of this bug is doubly confirmed (but obscured) by the data in Dr Tomkins' own paper. If one looks carefully at Figure 1 in his *Comprehensive Analysis,* one can estimate the percentage similarity of chromosomes 1, 2, 3 and 4 where the slice lengths are 100, 150 and 200 base pairs. For

chromosome 4, a slice length of 200 base pairs resulted in a ~62% similarity, while a slice length of 100 base pairs resulted in a ~28% similarity. This is quite obviously a mathematical impossibility – no combination of '*Ave % identity align*', '*Ave alignment length*' and '*Ave % hit frequency*' can produce these results simply by halving the slice length. Put more simply, a 200 base slice that has - on average - 124 identical nucleotides cannot be split into two 100 base slices that each have – on average – only 28 identical nucleotides. Rather than recognizing this impossibility, Dr Tomkins dismisses the anomalous results, saying only that "*sequence slices below 200 bases produced* ***non-optimal*** *alignments*" (emphasis added) [1].

If Dr Tomkins were to replicate his experiment and he was able to identify those sequences that did not return a match at first attempt, he would be able to verify – manually or otherwise – that these sequences do indeed find a match against the corresponding human chromosome.

Gapped vs. Ungapped Experiment

The results of the *Batch vs Serial Experiment* above merely showed the *existence* of the bug in the BLAST software. There are still significant corrections to be made to Dr Tomkins' methodology before arriving at a reliable result. As discussed in the *Materials and Methods* section, Dr Tomkins' employs the `ungapped` parameter in a setting where he presumably intends to report a result that *includes* indels. Since the `ungapped` parameter returns results that exclude insertions and deletions, it should only be used to calculate the substitution rate (or mutation rate) between the two species. That is, if Dr Tomkins intended to report a similarity figure that excluded indels, he should be reporting a figure in the order of 98.5% (see *'% Identity Align'* in Table 1).

Dr Tomkins' calculation method to arrive at a final figure is also inappropriate given his use of the `ungapped` parameter. He calculates the '*Ave % query identity*' by taking into account the length of the alignment ('*Ave alignment length*') as a percentage of the length of the query ('*Ave query seq length*'). With approximately 5 million insertion and deletion events across the two genomes [9], and the BLAST algorithm intentionally constrained from continuing the alignments through those putative indels, this will obviously produce shorter average alignment lengths. In short, if Dr Tomkins wishes to factor in the length of the alignment as a percentage of the length of the query, then he must allow to BLAST algorithm to extend alignments through putative indels, and therefore it is completely inappropriate to use the `ungapped` parameter.

Dr Tomkins has chosen to exclude non-DNA letters from the analysis. I agree with this decision, but he has made an error in the method he uses to achieve this, and that error has the effect of slightly understating his results. In the unmasked chimpanzee genome, there are over one hundred thousand sections of DNA that are yet to be accurately sequenced. If the genome cannot be accurately assembled, '*gap filling 'N's'* are introduced as placeholders for where the DNA is uncertain. These gaps usually occur at the ends of chromosomes, and are often tens of thousands of base pairs long; occasionally millions of base pairs long. However, many of these gaps are much smaller (between 1 and 200 base pairs) and occur inside the euchromatic sequence.

The problem arises when Dr Tomkins' script [5] removes these characters entirely from the query sequence, but then continues to use the mangled query sequence in his comparison. This new query sequence has had, in effect, an artificial deletion event applied to it. As explained in previous paragraphs, this will lead to artificially shorter alignments due to Dr Tomkins' use of the `ungapped` parameter. This has the effect of slightly understating the overall similarity.

The results of this experiment (See Table 2) clearly show the impact of allowing the BLAST+ algorithm to extend alignments through putative indels. This is reflected by the significantly longer alignments, leading to higher overall identity.

Statistical Sampling Experiment

The method used to calculate the overall similarity is quite simple. I take the number of identical bases (represented by the `nident` parameter) and divide it by the greater of the alignment length (`length`) and the query length (`qlen`). This is quite a conservative method, since those query sequences that return shorter alignments are treated as if they aligned over the full 300 base pairs. While those query sequences that contain putative indels often return an alignment length greater than the length of the query. Those sequences that do not return a match at all are treated as if they had zero identical base pairs over the full query sequence. Some illustrative examples are given in Table 3.

Of the 240,000 sequences submitted, only 100 could not find a match at all; 90 of those were on the Y chromosome. In total, 72,000,000 base pairs from the chimpanzee genome were aligned to 72,144,948 base pairs from the human genome. More than 95% of the queries returned a match where 270 or more of the 300 base pairs were identical; in other words, more than 95% of the queries were *at least* 90% identical.

A percentage similarity is calculated for each chromosome (with a corresponding Standard Error) and then these figures are weighted against the amount of sequenced DNA in each chromosome to give a final figure. The headline results of this experiment are shown in Table 4. Overall, I calculate that the chimpanzee genome is 96.90% identical to the human genome, with a standard error of ±0.21%. All results are available on request.

Conclusion

My *Batch vs. Serial Experiment* conclusively shows the existence of a software bug in version 2.2.27 of the BLAST+ software, and, that *without* correcting for the effects of this bug, I report findings largely in line with Dr Tomkins.

My *Gapped vs. Ungapped Experiment* shows the effect of allowing the BLAST algorithm to extend alignments through putative indels, giving a more realistic overall figure.

After correcting for the effects of the bug in version 2.2.27 of BLAST+ and allowing for insertions and deletions, my findings are largely in line with the secular scientific consensus. Accordingly, it is incumbent on Dr. Tomkins to repeat his comparisons in such a way that he can identify and resubmit those query sequences that do not return a result at the first attempt. This would necessitate the use of the BLAST+ output format parameters outlined above that can identify the query and subject sequences. Dr Tomkins must also decide whether to report results that do or do not include indels, and adjust his methodology accordingly.

While these two errors in Dr Tomkins' paper account for the chasm between his results and the peer-reviewed secular literature, obtaining an *exact* figure for the genetic similarity between chimpanzees and humans is an immensely difficult task. There are many factors that need to be taken into account, and many of these factors may need to be given only subjective relevance when it comes to determining that final figure. To use the Chimpanzee Sequencing and Analysis Consortium as an example, there have been thirty five million single nucleotide changes and five million insertion/deletion events, yet the total impact of the insertion/deletion events dwarfs that of the single nucleotide changes (~3% versus 1.23%) and occur with only one seventh of the frequency [9]. One identified source of these deletions is due to *Alu* recombination-mediated deletion (ARMD), and have been found to cause an average loss of approximately 1,000 base pairs per event [10] [11]. It seems improper of me to suggest that a single ARMD event be considered the equivalent of 1,000 single nucleotide changes, yet in the method of calculating my results, I am effectively agreeing that to be the case. It is on this basis that I present my final figure of 96.90% (±0.21%) identity.

References

[1] J. Tomkins, “Comprehensive Analysis of Chimpanzee and Human Chromosomes Reveals Average DNA Similarity of 70%,” *Answers Research Journal,* vol. 6, pp. 63-69, 2013.

[2] J. Tomkins and J. Bergman, “Genomic monkey business – estimates of nearly identical human-chimp DNA similarity re-evaluated using omitted data,” *Journal of Creation,* vol. 26, no. 1, pp. 94-100, 2012.

[3] R. Buggs, “Chimpanzee? Reformatorisch Dagblad,” 10 10 2008. [Online]. Available: http://www.refdag.nl/chimpanzee_1_282611. [Accessed 25 10 2014].

[4] P. Cosmo, “ProgettoCosmo - An automatic Comparison of the Human and Chimpanzee Genomes,” 2012. [Online]. Available: http://progettocosmo.altervista.org/index.php?option=content&task=view&id=130. [Accessed 25 10 2014].

[5] J. Tomkins, *Personal Communications,* 2014.

[6] National Center for Biotechnology Information, “BLAST Command Line Applications User Manual,” 30 July 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK1763/. [Accessed 14 August 2014].

[7] M. Smith, “Is it the sample size of the sample as a fraction of the population that matters?,” *Journal of Statistics Education,* vol. 12, no. 2, 2004.

[8] NCBI, “BLAST+ Release Notes,” 3 1 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK131777/. [Accessed 2 9 2014].

[9] Chimpanzee Sequencing and Analysis Consortium, “Initial sequence of the chimpanzee genome and comparison with the human genome,” *Nature,* vol. 437, pp. 69-87, 2005.

[10] S. K. Sen, K. Han, J. Wang, J. Lee and H. Wang, “Human genomic deletions mediated by recombination between Alu elements,” *Am J Hum Genet,* vol. 79, pp. 41-53, 2006.

[11] K. Han, J. Lee, T. J. Meyer, J. Wang, S. K. Sen, D. Srikanta, P. Liang and M. Batzer, “Alu Recombination-Mediated Structural Deletions in the Chimpanzee Genome,” *PLoS Genet,* vol. 3, no. 10, p. 184, 2007.