Why the shared mutations in the Hominidae exon X GULO pseudogene are not evidence for common descent

Royal Truman and Peter Borger

Appendix: The nature of bioinformatic evidence

Suppose an evolutionist has been informed that the nt *G* (Guanine) is found in all the organisms of a dataset for a particular gene, except for *Hu* (human) and *Ch* (chimpanzee), for which nt *A* (Adenine) is found at that position. Since he 'knows' the present evolutionary phylogeny for Hominidae is true, he shows figure 4 point P3 to any sceptic and presents this nucleotide information as persuasive evidence for the current model. To illustrate the trap, the reader has just been deliberately misled! From position 96 of table 2 we see that the *A* nt is unique to only *Or* (orangutan) and *Ma* (macaque), and not *Hu* and *Ch*, which makes no evolutionary sense. What seemed quite convincing a second ago, now requires a statistically improbable scenario: both neutral mutations occurred independently by chance, for only these two organisms, and this then spread throughout both populations.

Some time later you are informed that *Hu* and *Ch* uniquely share an nt *A* in a dataset, (not the same nt position as mentioned above) contrary to all other organisms which display the nt *G*. How convincing is this evidence now for a *Hu*/*Ch* common ancestor? Obviously far less, given your preceding experience. Your caution would be justified. In fact, we have deliberately misled the reader for a second time! The observation refers now to position 132 (table 2) and the imputed common ancestor is once again *Or* and *Ma* (and not for *Hu* and *Ch* as we just pretended).

There are two positions (75 and 95 see table 2) in which indeed *Hu* and *Ch* only show the same nt in the dataset.

Our evolutionist presents us next with a statistically greater challenge. We are told that *Hu*, *Ch* and *Or* all share the nt *T* (Thymine) and all the other organisms the nt *C* (Cytosine). Arguing for two such coincidences might be difficult, but for three such coincidences at exactly the same position and in accord with evolutionary thinking you are going to have difficulties. This evidence matches evolutionary theory well (figure 4 point P2), and is essentially compelling, right? Well, not really. We have chosen to mislead the reader for a third time to drive the point home. We are referring to position 55 (table 2) and the nt *T* in common refers to *Or*, *Hu* and *Ma*. Although the *Ch* is supposed to share a common ancestor with *Hu* after the *Or* line branched off, the expected nt is not found for *Ch*. The only reasonable evolutionary answer, is that precisely at that position a back-mutation occurred to the original nt *C*. But examination of table 2 implies very few mutations have occurred at all, and such coincidences demand arguing against the facts. This aspect of coincidences will be discussed further below. Incidentally, the pattern at position 140 (table 2) could indeed be interpreted in a manner the evolutionist would like: here *Or*, *Hu* and *Ch* share the nt *C*, whereas all the other organisms the nt *T*.

We see that given enough data we can easily select whatever data suits our purposes and ignore or downplay the rest.

Why are intelligent researchers being so easily misled to see evolutionary evidence in patterns of nucleotide or protein sequences? There are three principles which we hope to explain in greater detail in a future paper.

It is, hardly surprising that organisms with a *similar* Bauplan and environment will indeed share many designed genetic features. This is intuitively anticipated by those believing in design. It would be unreasonable to expect elephants and *E. coli* to possess highly similar genomes. After all, we do expect genetic information to have visible morphological outcomes! Organisms in very different taxa will on average show significantly different gene sequences. By the mathematical nature of how evolutionary trees are algorithmically programmed, in which the more similar sequences are assumed to have branched off from a common ancestor, it is inevitable that apparently reasonable evolutionary trees at this very rough degree of detail will often result.

Evolutionists and creation scientists agree that there was indeed a common ancestor for dogs, for bears, for ducks, etc. Sequence analysis at this micro-level can reveal in principle true phylogenetic relationships within the original created biblical 'kinds'. (The detailed scenarios and models do differ, however. The post-Flood environments with low population sizes would permit a large number of mutations to fix quickly, whereas evolutionists believe new information arose through a long process of random mutations plus natural selection. Most creation researchers believe organisms were endowed *ab initio* with genetic possibilities which were later expressed and that genetic information did not arise by chance.)

The evolutionary framework possesses a vast number of candidate phylogenetic markers and adjustable parameters. There are virtually no real, *a priori* predictions, uniquely limited to the evolutionists, as to what genetic data to expect.

We hope to offer a detailed analysis of point (iii) in the future. Researchers overrate the strength of evidence which

seemingly supports their theory if they can immediately map data presented to an interpretation they are very comfortable with. If two or more nts in our dataset match up in a manner consistent with a reasonable common ancestor, this explanation pops immediately into the evolutionist's mind. Both of us have spent over a decade being trained in evolution-dominated secular universities. We both can immediately offer multiple evolutionary possibilities to most data presented to us. We can also quickly offer the best evolutionary 'excuse' when the data does not meet theoretical expectations. It is our hope to show our evolutionist friends that what seems apparent is a mirage.

Note that evolutionary theory has not stated in advance which mutations in common would be expected to arise from which common ancestor. Intuitively, when wearing evolutionary glasses, we *accommodate* the data *post facto* into the theoretical framework. Therefore, if some organisms share a suitable pattern, and this is presented in a manner where the evolutionary explanation is immediately apparent, then too much significance is assigned to the finding. Particularly guilty are phylogenetic bifurcation trees (see figure 1), in which a common ancestor is directly claimed. Other data clustering methods merely indicate closer resemblance in a more neutral way, such as our figure 2 and figure 3, although almost all modern bioinformatic alignment algorithms are based on and calibrated on evolutionary assumptions.

The potential for coincidence is vast. Suppose our data implied a common ancestor for three out of four organisms in a dataset (or 4 out of six, or 5 out of eight …). The interpretation is then 'obvious': the shared-derived character was 'obviously' present on a common ancestor, which some lineages subsequently lost.

We can formalize this observation using decision theory. We define two statements of opinion, S1 and S2, and information fact I. Here S1 and S2 are mutually exclusive, and p(S1) + p(S2) = 1.

S1 : 'Evolutionary theory is the true explanation'
S2 : 'Evolutionary theory is not the true explanation'
I: 'A sequence pattern is found predicted by evolutionary theory'

Using Bayes's Rule,

$$P(S1 \mid I) = P(I \mid S1)/P(I) \times P(S1) \quad (1)$$

where P(S1 | I) means, 'the probability we assign to statement S1 given that we have been informed about fact I'.

Given that information *I* was in fact found, the *posterior probability* P(S1 | I), of belief statement S1, is given by the right hand side of (1). P(S1) is the *prior probability* before such data became available.

P(I | S1)/P(I) has the potential to modify a prior belief, and cannot be less than one. Now, neo-Darwinian theory has been in a state of parameter fine-tuning for over half a century. Sequence alignment weighting matrices have been optimally calibrated[1] to provide evolutionary theory the highest consistency possible. This means that the desired date of lineage divergences, according to current theory, are typically used to calculate probabilities of conversion from one nt or amino acid into another in for example PAM matrices.[2] Frequency of events such as gene duplication and mutations are also calibrated by evolutionary assumptions. When the results don't agree well,[3] the assumed evolutionary dates can be modified.[4] Discordant genes or parts of their sequences are simply stated as providing the wrong signal.[5] All these parameter fine-tunings[6] lead to a modified model which did not result from fundamental evolutionary assumptions. Fundamental theory did not predict creation of a GULO pseudogene for guinea pig and primate lineages 20 Ma ago,[7] nor was this based on any fossil or morphological data. In fact, the morphological basis for classifying the guinea pig in the order Rodentia was proposed[4] to be irrelevant only after so many gene sequence abnormalities were discovered. Dates, parameters and interpretation are constantly readjusted to optimise internal consistency, almost totally devoid of objective constraints.

These observations imply that we may find model-optimised examples in which P(I) > 0 ('A sequence pattern is found predicted by evolutionary theory'). This is hardly surprising, given the rich variety of parameters available to make evolutionary scenarios fit.[8] But are these probabilities truly lower than P(I | S1), meaning probabilities of being correct only if evolutionary theory is in fact true? Note that in the examples in which we misled the reader we cannot distinguish between P(I | S1) and P(I). Does evolutionary theory really predict the same nt for only *Hu* and *Ch*? Sometimes we find this result. This is in accord with evolutionary theory and thus reinforces the belief the theory is true. Sometimes we don't find this result. *'"So what" thinks the evolutionist. The theory never predicted this pattern anyway.'*

The creation science theoretician also has many degrees of freedom available to create scenarios which fit the available data. Different categories of gene sequences may have been created initially. Furthermore, shortly after the Flood many species were present in very low numbers. Based on radioactivity studies[9] it is possible that mutation rates may have been very high in the past.[10] The latter two factors suggest that numerous mutations may have occurred and fixed almost immediately in the entire populations very rapidly in the past. These models also have much freedom in guessing when various speciation events may have taken place. After a large number of converging interactions (i.e. model tinkering) a fine-tuned scenario would also lead to predictions of P(I) > 0, where I now mean 'A sequence pattern is found predicted by creation theory'. But once again, is it truly so, that P(I | S1) > P(I)?

Statistically founded guesses in the absence of any theory will generally be far better than random guesses. One need have no opinion about the origin of life to develop strictly empirical and useful statistical models. One can

collect any cellular feature one wishes, correlate with other features, cluster as one finds appropriate, and thereby permit better predictions for an unstudied organism. *I* then becomes: 'A sequence pattern is found predicted by a statistical model'. We certainly now expect $P(I) > 0$. But is $P(I \mid S1) > P(I)$ truly due to whatever story we invent to embellish the trends extracted from empirical models? Is the story of any real value, if the predictions we make simply rely on statistical observations, properly expressed mathematically?

It is our opinion that statistical analysis can indeed be fruitful in identifying patterns which provide intuition for additional research. But to fathom the true meaning behind coding and non-coding DNA sequence patterns a much deeper understanding is needed into all the kinds of coded signals[11] and Design goals needed by various cells. Superimposed are randomising mutations ('noise') which may camouflage the original intent, and these must also be studied before sequence data is to be understood.

References

1. Read revealing statements in the Materials and Methods sections of papers in which multiple parameters have been used, calibrated in a manner to optimize a desired evolutionary outcome. Here is an example: 'The mtREV-24 model of amino acid sequence evolution and the TN-93 model of nucleotide evolution were used for distance and likelihood analyses.' 'The analyses were performed under the assumptions of … plus four classes of variable sites'. From Arnason, U. *et. al.*, Mammalian mitogenomic relationships and the root of the eutherian tree, *PNAS* **99**(12):8141–8156, 2002.
2. Dayhoff, M.O., Eck, R.V. and Park, C.M., Atlas of protein sequence and structure, NBRF, Vol. 5, p. 75–84, 1972.
3. Arnason, U. *et. al.*, ref. 1, p. 8154: 'In comparison, the nucleotide trees differed considerably depending on the analytical approach used.'
4. Li, W.-H., Hide, W.A., Zharkikh, A., Ma, D.-P. and Graur, D., The molecular taxonomy and evolution of the guinea pig, *The Journal of Heredity* **83**(3):174–181, 1992.
5. Arnason, U. *et. al.*, ref. 1, p. 8155: 'Only the stem regions of the mt rRNA genes seem to carry a useful (albeit marginal) phylogenetic signal and inclusion of the other parts of these sequences may increase the background noise and promote the selection of the wrong signal.'
6. Arnason, U. *et. al.*, ref. 1, p. 8155: 'Finally, the pinniped results lend no support to the morphological hypothesis…underlining the problems associated with basing phylogenetic conclusions on anatomical features that may have strong adaptive values.'
7. Nishikimi, M., Kawai, T. and Yagi, K., Guinea pigs possess a highly mutated gene for L-gulono-gamma-lactone oxidase, the key enzyme for l-ascorbic acid biosynthesis missing in this species, *The Journal of Biological Chemistry* **267**(30):21967–21972, 1992.
8. For example, a large proportion of gene sequences were less similar between humans and chimpanzees than between humans and gorilla in some cases. A paper was published in which it was claimed that millions of years after a split and genetic isolation from the gorilla lineage cross-breeding re-occurred. This bizarre theory prompted headlines in Germany like 'Scientists show early humans had sex with monkeys'.

   This *ad hoc* rationalization is found in: Patterson, N., Genetic evidence for complex speciation of humans and chimpanzees, *Nature* **441**:1–6, 2006.
9. (a) Vardiman, L., Snelling, A.A. and Chaffin, E.D., *Radioisotopes and the Age of the Earth*, Institute for Creation Research and Creation Research Society, 2000. (b) Vardiman, L., Radioisotopes and the Age of the Earth: <www.icr.org/pdf/research/RATE_ICC_Vardiman.pdf>
10. Arnason, U. *et. al.* ref. 1, p. 8155: 'It has been argued that the molecular estimates suggesting early origin of eutherian orders are artifactual and caused by accelerated molecular evolution at the Cretaceous/Tertiary (K/T) boundary or immediately thereafter.'
11. Trifonov, E.N., Genetic sequences as product of compression by inclusive superposition of many codes, *Molecular Biology* **31**(4):647–654, 1997.

Royal Truman has bachelor's degrees in chemistry and in computer science from SUNY Buffalo, an M.B.A from the University of Michigan, a Ph.D. in organic chemistry from Michigan State University and post-graduate studies in bioinformatics from the universities of Heidelberg and Mannheim. He works for a large multinational in Europe.

Peter Borger has an M.Sc. in Biology (Hons. biochemistry and molecular genetics) and a Ph.D. in Medical Sciences from the University of Groningen, The Netherlands. He is currently working on the cellular and molecular aspects of pulmonary diseases, such as asthma and COPD, and is an expert on the molecular biology of signal transduction and gene expression.