Simplifying the aligned NUMTs pseudogene sequence from nucleotides ‘ATCG’ to ‘0’/’1’ to test in coalescent hidden Markov Model

Yuanyuan Wu

Abstract:

In this article, the aligned NUMTs were simplified to ‘0’ for non-mutation and 1 for detected mutation. Parameters in the coal-HMM were adjusted to function with Jukes-cantor. It was also discussed that how to decode the E/A matrix in EM to the parameter for coal-HMM. Possible reasons for the failure of this model was discussed.

Introduction:

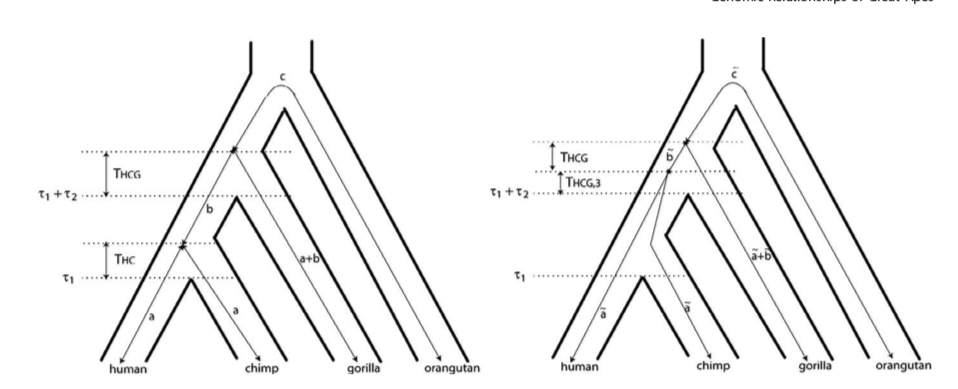
Mitochondria has transferred most of its genetic material into nucleus since its endosymbiotic event. In mammals, only 13 coding protein and a few tRNAs and rRNAs are left in the mitochondrial genome. The majority of the other proteins that are essential to the mitochondrial function is expressed in nucleus. Except for those functional gene transfer, there has been trunk of mitochondria-originated pseudogenes called NUMTs. A recent published paper stated that with the newly updated hominin chromosome sequences, it is possible to find long (~9kb) NUMTs in homolog in multiple hominin species. They were able to build phylogenic tree and found that the pseudogenes could result from interbreeding from an unknown species.

With the aligned data and coal-HMM model available, it is intriguing to test the partition of the pseudogenes to exam possible interbreeding after the pseudogenes.

Method:

The aligned sequence for the pseudogenes for human, pan, gorilla and pango were kindly gift from Dr.Konstantin Khrapko1. The pseudogenes are of length of around 9kb where gaps are included in the multiple alignment. For coalescent-HMM2, the emission probability were calculated with Jukes-cantor model, rather than the original Q-rate matrix in the original article.

In the Jukes-Cantor model, the equilibrium frequency of bases assume identical:

Where µ is mutation rate and ‘a,b’ are the nucleotide in the phylogenic tree linked by branch length t.

It is easy to get the branch length of HC1 and HC2 with given a,b,a2,b2.

In state HG,

t(human->chimp) = 2\*(tao1+tao2)+theta(HCG)\*2+1/3\*theta(HCG)\*2

while based on definition, a2+b2 = tao1+tao2+4/3\*theta(HCG)

hence, t(human->chimp) = 2\*(a2+b2)

The overall branch length rooted to human is given below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | H->C | H->G | H->O |
| HC1 | 2a | 2(a+b) | a+b+c |
| HC2 | 2a2 | 2(a2+b2) | a2+b2+c2 |
| HG | 2(a2+b2) | 2a2 | a2+b2+c2 |
| CG | 2(a2+b2) | 2(a2+b2) | a2+b2+c2 |

Since in Jukes Cantor model, the frequency of all nucleotides and their chance of mutation are considered symmetric. The aligned sequence of the four species are then simplified into the combination of 0/1 in each site. Where 0 means no mutation and 1 means mutation detected.

Compare all other 3 species to human sequences, there can possible be 8 emissions and they are:

After simplification:

The jukes cantor for the simplified version will be:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Possible emissions | 0000 | 0001 | 0011 | 0111 | 0110 | 0101 | 0100 | 0010 |
| Favored state |  |  | HC | CG | CG | HG | HG | HC |

The favored states for the specific emissions are drawn from the phylogenic tree. To further confirm the favored state from the emission probability:

According to the coal-HMM model,

So state CG is indeed favored for emission ‘0110’.

Similarly, other favored states are proved.

Also, it is confirmed in the below that the P for a state for all emissions sum up to 1.

After calculate emission probability form jukes cantor and branch length, it is not hard to get the emission matrix. Emissions of the same value are found in the matrix:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0000 | 0001 | 0011 | 0111 | 0110 | 0101 | 0100 | 0010 |
| HC1 |  |  |  |  |  |  |  |  |
| HC2 | X4 | X3 | X1 | X5 | X6 | X2 | X7 | X8 |
| HG | X4 | X3 | X2 | X5 | X6 | X1 | X8 | X7 |
| CG |  |  | X9 |  |  | X9 | X10 | X10 |

When estimated emission probability matrix was calculated from EM, it was then corrected with the expectation above.

To decode a,b,a2,b2 from corrected ekb in EM:

For any of the state:

Hence,

And t2, t3 can be calculated the same way.

For transition probabilities, when simplify the transition model to reduce free parameters to 3, phi(s,u,v), the transition probability matrix is as follow:

Transition probability matrix:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | HC1 | HC2 | HG | CG |
| HC1 | 1-3s | s | s | s |
| HC2 | u | 1-u-2v | v | v |
| HG | u | v | 1-u-2v | v |
| CG | u | v | v | 1-u-2v |

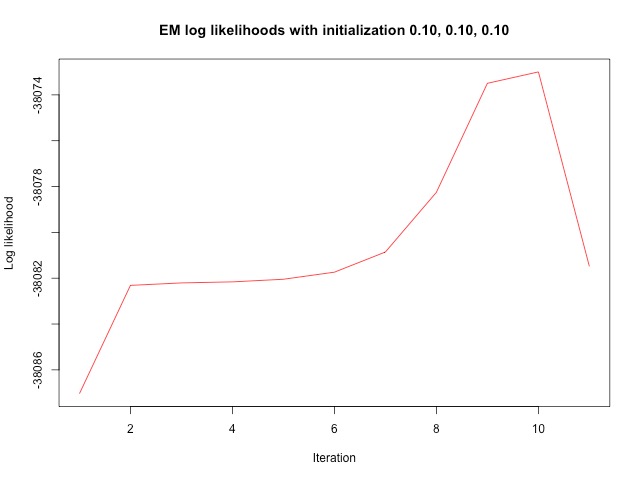
And thus:

Result and discussion:

The length of the pseudogenes for chromosome 5 are of 9159bp.And the overall emission distributions in the pseudogenes are as follow:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Emissions | 0000 | 0001 | 0011 | 0111 | 0110 | 0101 | 0100 | 0010 |
| Count | 6957 | 738 | 233 | 294 | 146 | 119 | 258 | 413 |

The overall distribution of different the sequences is ideal. However, when conducting Baum-Welch with adjusted Jukes Cantor, the statistical decay appeared.



The overall timescale(a+b+c) were adjusted in multiple tests from 36 thousand to 36 million year. (since the introgression of the mtDNA is uncertain). When larger time scale is tested, the s,u,v tend not to change and the value of the a,b,a2 are constantly optimized to around:

a: 6.97492

b: 3.03671

a2: 8.49783

despite initiations of a,b,a2.

The sum of the E and A matrix were correct at the first few rounds and suddenly lose control to a gigantic value after a few iterations.

One reason for the failed model can be the oversimplified ‘0011’ mode. The sequence of other species that does not match the human sequence are labeled as ‘1’. This lead to the change in Jukes-cantor probability and saved time and computation power. It leads to the emissions decrease from 4^4(AATC,ATCC,etc) to 8(0011,0101, etc.). It was proved in the method that the emission probability matrix for the simplified version is correct while unknown fatal mistakes happen. The aligned data is also doubted. This pseudogenes aligned quite well among the sequences and it was doubted if similar coalescent pattern I restricted it to is true in the pseudogenes. Whether recombination will occur in this 9kb long pseudogenes are doubted since the divergence of the pseudogenes is far later then the divergence of the hominin, which makes recombination in the pseudogenes less likely.



Future implements include test this model in other common sequence data where recombination is likely to happen. That is, the selected trunk of DNA sequence should occur in the common ancestor of the selected species to allow recombination and nuclear coalescence occur. This could rule out the possibility of the improper data tested for the model.

Reference:

1. Popadin K, Gunbin K, Peshkin L, et al. Mitochondrial pseudogenes suggest repeated inter-species hybridization in hominid evolution. *bioRxiv*. January 2017. http://biorxiv.org/content/early/2017/05/09/134502.abstract.

2. Hobolth A, Christensen OF, Mailund T, Schierup MH. Genomic Relationships and Speciation Times of Human, Chimpanzee, and Gorilla Inferred from a Coalescent Hidden Markov Model. *PLOS Genet*. 2007;3(2):e7. https://doi.org/10.1371/journal.pgen.0030007.