A Model of Duplication-Loss of Genes in a Multigene Family

10.12.15

The field of molecular evolution and phylogenetics was revolutionised when Felsenstein (1981) proposed inferring phylogenies as a statistical problem based on a specific model of sequence evolution. Before this, it was predominantly view as a optimisation problem trying to minimising the total number of events [parsimony] or maximising the fit of a phylogeny to data. It took 1-2 decades before this algorithm had been turned into userfriendly programs, but today it is is the foundation of testing the molecular clock and models of molecular evolution. In 1991 Thorne, Kishino and Felsenstein initiated a similar development within alignment models, that also solely were optimation based and not founded on a stochastic model of insertion-deletions. This still awaits being used as the method choice due to the large computational cost of alignment algorithms. There is a third problem that deserves also to be based on stochastic models and thus allow proper statistical inference and that is the evolution of genes in multiple contiguous copies [***multigene families***] in different species.

Most genes in higher organisms are members of a multigene family, which creates several methodological issues and makes this problem very important.

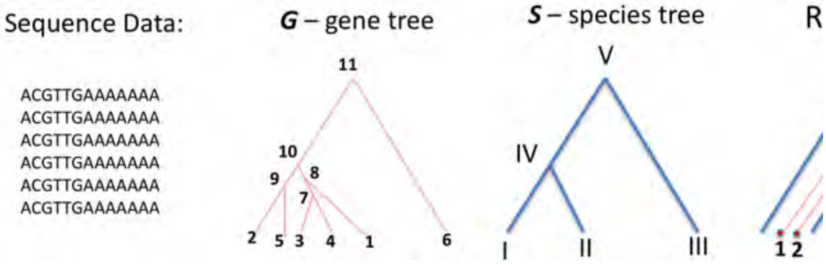
It becomes harder to make an identification between gene tree and species tree. Homologous genes from two species that have "the same" position in the multigene family and whose most recent common ancestor [MRCA] was at the speciation event at the MRCA of the two species are called ***orthologous***, otherwise ***paralogous***. This formulation ignores population aspects and what is called incomplete lineage sorting. However, determining this paralogous-orthologous relationship between genes is highly challenging and a series of methods have been developed.

Non-statistical approaches will take the set or list of genes found in the different species and propose duplications-deletions to make the gene phylogeny fit into the species tree. A statistical approach will assign birth and death rates to duplications and loss and make probablistic statements about how the gene tree is embedded in the species tree.

It seems that existing methods all assume we have a set of homologous genes in each species and not a list of genes on a chromosome or one several chromosomes. It is the goal of this project to investigate a simple stochastic model that creates multigene families on a single chromosome.

A pioneering paper is Goodman (1979), which focused on analyzing globins and used heuristic approaches. In the early 80s a series of papers [and a book by T Otha and an even earlier book by S Ohno] investigating the population genetics of recombination, gene conversion and multigenes but the perspective was rarely phylogenetic [for instance Petes].

Arvestad and Lagergren (2004-) and colleagues published a series of papers addressing this problem both in optimisation and probablistic framework. In reality a set of sequence is given, but often it is assumed that their tree (G) has been determined unambiguously and that the species tree (S) is known. The problem is then to pack G into S.



Stadler [born Gernhard](2008-) investigated a birth-death process. Much goes back to Kendall (1948) and results can also be found in Elisabeth Thompson's PhD thesis "Human Evolutionary Trees" (1975). Let l be the birth rate, m the death rate and t the time on a branch, then the probability that a gene has 0, 1, k surviving descendants are time t was calculated by Kendall as (p_0(t)), (p_1(t)), (p_n(t)), respectively:

$$p_0(t) = \frac{\mu(1 - e^{-(\lambda-\mu)t})}{\lambda - \mu e^{-(\lambda-\mu)t}},$$

$$p_1(t) = \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)t}}{(\lambda - \mu e^{-(\lambda-\mu)t})^2},$$

$$p_n(t) = (\lambda/\mu)^{n-1} p_1(t) [p_0(t)]^{n-1}$$

The probability that n genes has m surviving descendants can be calculated from the individual p-functions and is (Hein et al., 2000, but probably a known result) easily derived call P_m there.

$$P_m = \sum_{i=0}^{\min(m,n)} \binom{n+m-i}{m-i \quad n-i \quad i} (-a)^{n-i} d^{n-i} c^i b^{-n-m+i-1}$$

$$a = a(t) = -\frac{\lambda}{\mu-\lambda}\gamma(t)$$

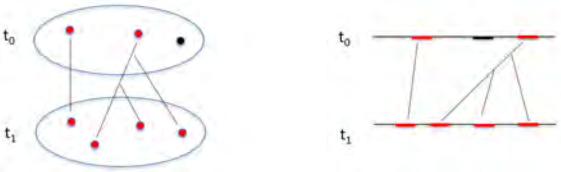
$$b = b(t) = 1 - a(t) = 1 + \frac{\lambda}{\mu-\lambda}\gamma(t)$$

$$c = c(t) = 1 - \frac{\lambda}{\mu-\lambda}\gamma(t)$$

$$d = d(t) = 1 - c(t) = \frac{\lambda}{\mu-\lambda}\gamma(t)$$

$$\gamma(t) = 1 - e^{(\lambda-\mu)t}$$

This is forward in time, but $P_{n\rightarrow m}(t)$ going backwards can be calculated in a similar fashion. This n→m includes the ancestral genes that didn't leave any descendants, but it is also possible to only calculate, $P'_{n\rightarrow m}(t)$, that only counts the genes that left at least one descendant and clearly now n<=m. In Gernhard (2008) one can also find the actual duplication times in the tree relating the m genes.



All these approaches assumed that a multigene family was a SET of genes and where here propose to consider them as a STRING of genes. It is more realistic but will lead to harder computational problems. In a set, as for instance modelled by the coalescent process ANY pair can choose to find a common ancestor [n(n-1)/2 possibilities], while for genes on a string only neighbors can choose [n-1 possibilities] can find common ancestors. Processes of gene loss and duplications on sets or strings will define a prior of on the tree relating the observed genes. If the actual sequence have been determined it will define a posterior on the trees relating the sequences with intuitive properties such as two very divergent sequences will probably have a long time back to a common ancestor. If they are placed on a string additional constraints are invoked. For instance two genes cannot find a common ancestor more recent that any of the genes between them.

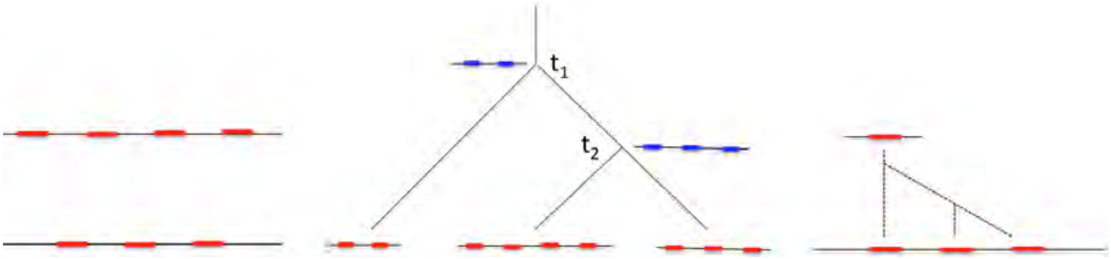
It seems the genealogies – counting and duplication times – for genes on a string is virgin territory in contrast to genealogies relating genes in a set [mainly coalescent theory] and there should be a series of interesting questions to pursue. Steel (2016) chapter 8 is a good to start for getting a better understanding of these models.

The problem is illustrated below.

Left) Two multigene families have been observed and the question is which genes in the top evolved into which genes in the bottom. This assignment could be done by minimizing some cost like total weighted sum of duplications, deletions and substitutions. Or by a probablistic treatment where the assignment would be given probabilities.

Middle) In a more realistic situation we have observed a series of multigene families at the leaves of a phylogeny. The multigene families at internal nodes are not observed (here hypothesized in blue) and again assignments of the multigene families at the leaves is deducable from all the assignments between pairs of multigene families on the edges.

Right) The key idea of this project is to create an evolutionary process on complete multigene families by defining a death-birth process on individual genes and extend by assuming independence among genes. This should give a probability distribution on any topology relating the extant genes back to a common ancestor and the duplication times (here t_1 and t_2).



Framed like this, the problem now is "Statistical Gene Alignment" with high similarity to the well studied "Statistical Sequence Alignment" problem (Lunter et al., 2008). The main and most important difference can be explained with reference to the Right in the illustration.

In sequence alignment you can find an efficient dynamic programming algorithm because you have expression for what is the probability that a nucleotide has k descendants and has itself been deleted or not. The right illustration above it has k descendants and has not been deleted.

But if have genes instead of nucleotides, it doesn't matter it if an individual gene has survived or not, only how many surviving offspring it has and which phylogeny relates them. If the phylogeny is given then the probability can be calculated using Felsenstein (1981) algorithm. However, the phylogeny is not known and the probability must be calculated by integrating over all possible phylogenies. This is very reminiscent of the problem of calculating the probability of a sample from a population where the distribution of phylogenies follows the coalescent distribution. But the tree distribution will be different for the gene duplication problem than for the allele sampling problem.

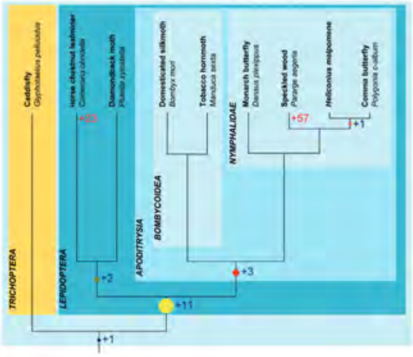
Clearly the proposed model has limitations: It only involves single genes copied-deleted in their entirety. Real evolution could involve more consecutive genes and could start-end in the middle of a gene. Additionally, such genes experience gene conversions that might be hard to model.

That genes are members of multigene families also gives opportunities. It clearly makes genomes larger and also allows duplicated to acquire new functions.

A probablistic approach to inferring the process of duplication-loss in the evolution of multigene families is plenty of work for a good Dphil. Just imagine the consequences of trying to incorporate events that include several contiguous genes or the information of the distance between genes on a chromosome.

And there is a huge amount of data to apply this problem to. So the problem would be to make a well defined and limited version of this large problem and would be an extension of Project c.

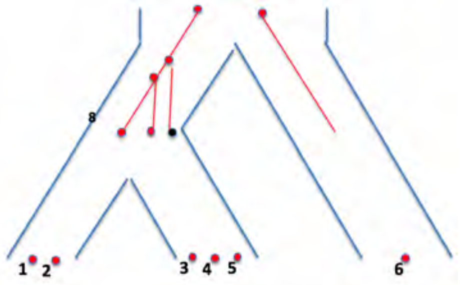
Project a: In many phylogenetic publications you can see a phylogenetic tree with a large number of genes in each multigene at the leaf and as this is traced back through the tree, then there are fewer and fewer genes assigned to the nodes. If these numbers at internal nodes are interpreted as the number of genes in the multigene family that has descendants at the leaves, then this can very well be very reasonable assertions. However, is it interpreted as the actual number in the multigene family at the ancestral node then an elementary error has been committed since genes with no descendants in the leaves will have been ignored. Committing this error is extremely commonplace and using a birth-death model of genes to infer would put this inference on a solid statistical footing. However, there is simple approach to this problem that could be useful stepping stone toward a full model: Assume you only know the NUMBER of genes in the multigene family at the leaves!! What is the distribution of the number of genes at the ancestral nodes? This and related distributions can be calculated very easily since the process of duplication-loss along an edge is Markovian and then Felsenstein's (1981) algorithm can be used. This algorithm would assign probability to any assignment of genes to ancestral nodes, but typically the combined most probably assignment to all nodes simultaneously and the marginal distribution for each node is of greatest interest and use.



A variant of the problem is not to use $P_{i,j}(t)$ the probability that ***i*** present genes descended from ***j*** genes time t ago, where some of these ***j*** genes would have no descendants among the ***i*** genes in the present, but only count the genes that had at least one descendant.

An example of such a phylogeney with less and less ancestors is here taken from Quah et al. (2015). Applying the implementation to this example would be a starting point, but much larger data sets are available.

Project b: Arvestad and Lagergren possibly made the problem difficult for themselves by taking an observed gene tree (G) and moving forward from the root toward the leaves in the species tree (S). They had to keep track of lineages that potentially could have descendants at the leaves and be consistent with their given gene tree. This project should take their key paper and reformulate the algorithm going backwards in time and implement it on a their example.



In principle this should be a simple publication. There are different versions of this:

i. Simply count the ways to embed [sometimes called reconciliations] G in S. The problem is simpler of G is rooted and will become more complicated if there is branch length information about G.

ii. Find the embedding with the most recent duplication events in G. There might be ways to select embeddings such that the subtrees in G obtained by only picking 1 gene from each leaf in S is a close to the species tree as possible.

iii. Probablistic version of ii.

Project c: The full probablistic modelling problem for a set of observed genes from a multigene family would be the ideal way to analyse real data, but would also be computationally challenging. Many, like Arvestad, assume the gene tree is given, since this simplifies the problem considerably. However, in reality the gene tree is only known via the genes and their positions and the results from Gernhard (2008) gives us a natural prior on the relationship of the genes. Thus,

$$P(D) = \int P(D|G)dG$$

where ***D*** is data (gene sequences assigned to leaves in the give species tree), ***G*** is the gene including times of internal nodes and ***dG*** is the prior on these trees. Suppressed is the specie tree and parameters of the process of molecular evolution.

Explicit symbolic integration over all possible times of nodes in the gene tree and summation over all possible topologies is not possible, so stochastic techniques such as MCMC or importance sampling would be necessary.

The method could be tried on β-globins in a set of mammal like human, mouse, cat and kangaroo.

A succesfull algorithm for ***Project c***, would also open the way for varying the species tree and thus have an efficient method for using contiguous gene families for inferring species trees.

References.

Benzaid,El-Mabrouk (2014) "Gene order alignment on trees with multiOrthoAlign " BMC Genomics, 15(Suppl 6):S5

Arvestad, L et al. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. RECOMB'04.

Höhna, Sebastian, et al. "Inferring speciation and extinction rates under different sampling schemes." Molecular biology and evolution 28.9 (2011): 2577-2589.

Lunter et al. (2005) Statistical Alignment: Recent Progress, New Applications, and Challenges Statistical Methods in Molecular Evolution pp 375-405

Stadler, Tanja. "Inferring speciation and extinction processes from extant species data." Proceedings of the National Academy of Sciences 108.39 (2011): 16145-16146.

Hein, J et al. (2000) Statistical Alignment J. Mol.Biol.

Gernhard, Tanja (2008) "The conditioned reconstructed process" J. Theor. Biol. 253. 769-78

Thorne, Jeffrey L., Hirohisa Kishino, and Joseph Felsenstein. "An evolutionary model for maximum likelihood alignment of DNA sequences." Journal of Molecular Evolution 33.2 (1991): 114-124.

Quah et al. (215) A burst of miRNA innovation in the early evolution of butterflies and moths. Mol. Biol. Evol.

Wu, Y et al. 2014. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. Genome Res 24: 475-486.

Steel, M (2016) Mathematical Phylogeny

Zhang, L. 2011. From Gene Trees to Species Trees II: Species Tree Inference by Minimizing Deep Coalescence Events. IEEE/ACM Transactions on Comp Biol & Bioinf.

Morris Goodman et al. (1979) Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences Systematic Zoology, Vol. 28, No. 2 pp. 132-163

Comments. i. This project was started at SAMSI, where Jotun Hein was on sabbatical in autumn 2014 and Jeff Thorne, Ryan Campbell, Xiang and Michael Golden discussed this at least once a week. The DTC student Guy Cooper spent some of his mini-project on this problem under the joint supervision of me and Peter Holland [Department of Zoology, Oxford].

ii. Manolis Kellis wrote papers on this topic, but it was very hard for me to see how it did something new relative to papers by Arvestad et al. a decade earlier

Oxford Work This means supervised or intiated by Jotun Hein. A lot of interesting work is going in Oxford that I will not mention here. Might be in future editions of enlarged project description. In this case some of the discussions actually took place at SAMSI, Research Triangle Park, North Carolina autumn 2014 as a part of a course given by Jotun Hein. Why this problem were chosen I don't know, but its was clear that despite having a 35 year history if we dated the start to 1979, it had in no way been solved despite hundreds of papers had been written on the topic. The group in NC consisted of Ryan Campbell, Xiang Ji, Jeff Thorne and we read 1-2 papers each week for 10+ weeks and partial clarity only gradually emerged. After returning to Oxford in December 2014, I discussed this problem with Prof. Peter Holland members of his group [Ferdinand Marletaz and Jordi Paps Montserrat] They had also collected papers on the topic and our collections were almost disjoint, hinting that the total number sampled from is real large!! There are many researchers interested in the evolution of multigene families most genes in higher organisms are part of multigene families. I have also discussed the problem with Stephen Kelly and David Emms.

Ryan Campbell

Xiang Ji

Fitting the gene tree inside the species tree is of extreme importance in phylogenetics and a long series of heuristic procedures and pipelines have been designed and papers using such methods must easily go into the thousands.

Guy Cooper

Peter Holland

Guy Cooper spent 20% of a mini-project supervised by Peter Holland discussing papers with me (and at times Mathias Cronjager). To me it was obvious that multigene evolution should be modelled by a birth-death process, but it was problem that in this context the duplication times were also important which was complicating in comparison to for instance statistical alignment. Thus we needed to know the full distribution on trees conditioned on the number of survivors. We found exactly what we need in a paper by Tanja Gernhard [now Stadler] which was a delight. I should have known the result since I was her PhD examiner in Munich!! And I did search for it, but only searched under Stadler!!