Statistical model selection in phylogenetic inference and its combinatorial structure

Hidetoshi Shimodaira
The Institute of Statistical Mathematics, 4–6–7 Minami-Azabu, Minato-ku, Tokyo 106–8569, JAPAN.
URL: http://www.ism.ac.jp/~shimo/ EMAIL: shimo@ism.ac.jp

Abstract: The branching order (i.e., topology) of evolutionary tree is inferred from the genetic information (i.e., molecular sequences) of contemporary species. This is an example of statistical model selection — Selection of a topology is essentially the same as the selection of a set of predictors in multiple regression; each topology is uniquely specified by a combination of *splits*, which are partitions of the species into two parts. The number of unrooted tree topologies is $M = (2m-5)!/(2^{m-3}(m-3)!)$ for m groups of species, while the number of splits is $B = 2^{m-1} - (m+1)$. Here we show the combinatorial structure of the trees and give numerical examples of real datasets[1]; we used mitochondrial protein sequences of $n = 3274$ sites for $s = 22$ species to solve the debate of the origin of tetrapods: G_1 = Tetrapods (15 species), G_2 = Lungfish, G_3 = Coelacanth, and G_4 = Ray-finned fish (4 species). The outgroup is G_5 = Lamprey. The number of groups is $m = 5$, so the number of possible unrooted topologies is $M = 15$. Also a graphical method is presented for understanding the relations among the parametric models with respect to data. The models are represented by their predictive densities, and they are drawn in Euclidean space preserving approximately the symmetrized divergence between these densities. This direct visualization of models is useful for diagnosis of the model selection, especially for nonnested models.

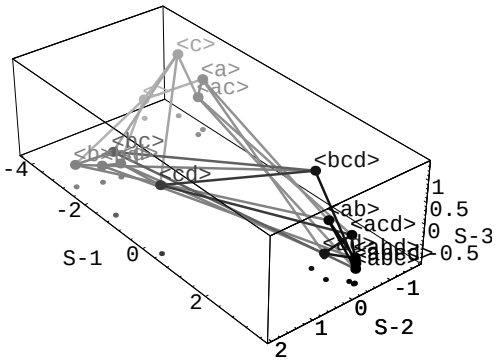

Figure 1: Example of the variable selection in multiple linear regression. The response variable (heat of cement) is expressed as a linear combination of four predictors (amounts of four major ingredients in percentage). The number of possible combinations of the predictors is $2^4 = 16$.

Table 1: Ten splits of five groups $\mathcal{G} = \{G_1, G_2, G_3, G_4, G_5\}$. Each split corresponds to the internal edge of trees which separates the groups.

a	=	$\{G_1,G_2,G_3\|G_4,G_5\}$	f	=	$\{G_3,G_4\|G_1,G_2,G_5\}$
b	=	$\{G_1,G_3,G_4\|G_2,G_5\}$	g	=	$\{G_2,G_4\|G_1,G_3,G_5\}$
c	=	$\{G_2,G_3,G_4\|G_1,G_5\}$	h	=	$\{G_1,G_3\|G_2,G_4,G_5\}$
d	=	$\{G_1,G_2,G_4\|G_3,G_5\}$	i	=	$\{G_2,G_3\|G_1,G_4,G_5\}$
e	=	$\{G_1,G_2\|G_3,G_4,G_5\}$	j	=	$\{G_1,G_4\|G_2,G_3,G_5\}$

[1] Joint work with Y. Cao and M. Hasegawa of The Institute of Statistical Mathematics.

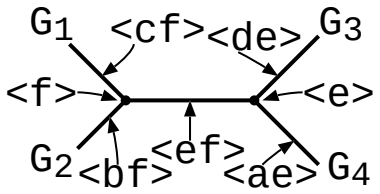

Figure 2: The five bifurcating topologies and the two multifurcating topologies, which are associated with the split $\{G_1, G_2|G_3, G_4\}$. The arrows indicate where G_5 stick to. Each bifurcating tree consists of $N = m-3 = 2$ splits. Not all the combinations of N splits out of the B splits are allowed to construct trees. Denoting $x^c = \mathcal{G} \setminus x$, for two splits $\{x|x^c\}$ and $\{y|y^c\}$ to construct a tree, one of $x \bigcap y$, $x^c \bigcap y$, $x \bigcap y^c$, or $x^c \bigcap y^c$ must be the empty set. This constraint makes the algebraic structure of tree topologies much interesting than the usual variable selection of multiple regression.

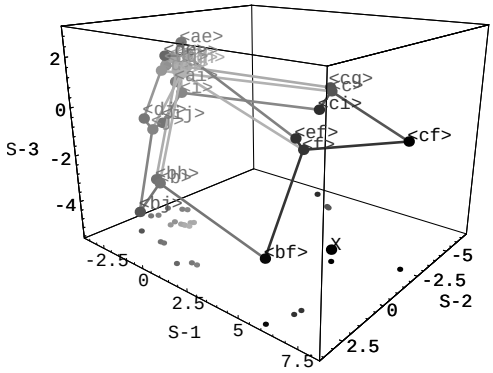

Figure 3: Model map of the vertebrates dataset drawn for the 15 bifurcating topologies and 11 multifurcating topologies. `X` denotes the full model spanned by the 10 splits around the star topology `<>`. The segments indicate the model structure.

Table 2: AIC and p-values for the bifurcating tree topologies of vertebrates. The most possible topology is `<cf>`, but the other topologies with large p-values are not rejected as well.

	α	AIC_α	$P_\alpha^{(L)}$	$P_\alpha^{(M)}$	$P_\alpha^{(B)}$
1	<cf>	104790	0.500	0.946	0.516
2	<bf>	+5.46	0.385	0.794	0.354
3	<ci>	+23.83	0.135	0.443	0.108
4	<ef>	+26.31	0.036	0.178	0.002
5	<cg>	+44.41	0.007	0.041	0.000
6	<bj>	+45.73	0.074	0.226	0.015
7	<de>	+55.19	0.015	0.103	0.002
8	<bh>	+60.76	0.018	0.032	0.000
9	<ij>	+62.76	0.022	0.145	0.001
10	<dj>	+64.17	0.020	0.132	0.001
11	<ae>	+64.40	0.004	0.036	0.000
12	<ai>	+68.56	0.012	0.078	0.000
13	<dg>	+78.18	0.003	0.027	0.000
14	<ah>	+88.12	0.001	0.008	0.000
15	<gh>	+90.82	0.000	0.004	0.000