Detecting variation in evolutionary rate in MCS lineages

Introduction

The evolutionary constraint acting on MCSs (CNGs) can be investigated by comparing the shape of their evolutionary tree to the trees estimated from neutral, or nearly neutral, sequences. For the following analyses the 4D tree available at (ftp://kronos.nhgri.nih.gov/pub/outgoing/elliott/encode_tba/MAY-2005/phylo/4d-tree.nh) is used as the neutral tree.

In my analyses there are two primary questions to be asked of each MCS alignment (kindly provided by Elliott):

1) Is selection constant in the MCS?
2) How is the variation in selection distributed across branches?

The methods used to examine these questions are statistical methods, based on maximum likelihood calculations on a phylogeny, and more precise details of the methodology used are at the end of this document. The method uses rate as a proxy for selective constraint, and tests whether the ratio of branch lengths estimated from an MCS are different to those in the 4D tree. The methodology is not perfect, but should allow the initial exploration of selective constraints in MCSs across a tree.

Visualising selective variation in lineages

There are four attached .pdf files:

Explanation.pdf
ExtremeHumanENm001.pdf
ExtremePrimateENm001.pdf
ExtremeEutherianENm001.pdf
First35ENm001.pdf

These .pdf show the same style figure, demonstrating the variation in selective constraint across the encode phylogeny. To start understanding these graphs it's probably best to open, say, ExtremePrimateENm001.pdf, and Explanation.pdf. On the LHS is the tree topology used, with each branch in it mapping to one row of the'thermoplot' on the right. The 'thermoplot' on the RHS indicates the strength of variation in selective constraint in a region. The Explanation.pdf should hopefully explain what the different values mean. The deep red and blue boxes are roughly equivalent to a $P < 0.001$ ($X^2=10$ in a χ^2 distribution with 1 d.f.) for expanded and contracted branches, respectively.

The other four graphs show:

ExtremeHumanENm001.pdf – The 35 least constrained (red) MCS in human in region ENm001
ExtremePrimateENm001.pdf – The 35 least constrained (red) MCS in the lineage leading to primates in region ENm001
ExtremeEutherianENm001.pdf – The 35 least constrained (red) MCS in the lineage leading to eutherian mammals in region ENm001
First35ENm001.pdf – The first 35 MCS in region ENm001.

What do the results mean

As far as I am aware nobody has produced anything similar before, so a thorough and meaningful interpretation of results will take some time. I shall discuss some things that might be interesting in some of the attached 'thermoplots'.

The first observation, made across all plots, is that the selective constraints acting on MCSs are more variable than I originally expected. This is the most straight forward interpretation of the results, answering question 1 in the introduction with a yes.

How and why the MCSs vary in the manner they do is highly interesting and I have no general explanations yet. Looking at ExtremeEutherianENm001.pdf provides some interesting insights. The most obvious to observe is actually an artefact of the method: the high correlation between eutherian and monodelphis is a result of many of the alignments containing only eutherians + monodelphis, which results in them taking the same value (see Explanation.pdf).

A second observation is that there are lots of highly constrained branches in the mammals (dark blue), more so than in the first 35 MCSs in First35ENm001.pdf. This may indicate that MCSs that are highly variable (perhaps positively selected?) between monodelphis and the eutherians become highly constrained in most eutherians (as a result of a new function?).

There are many more potentially interesting results in the graphs and tables that I shall allow others to investigate. If any more figures are required please let me know. Unfortunately I am away from work on the Saturday afternoon and Sunday of the ENCODE get together, but I will endeavour to get them to you as soon as possible after that (or on Sat morning/Friday afternoon and evening).

Raw values for MCS variation

These are located in the Results.all file. Each line (apart from the header) describe more complete results for each MCS region in a space delimited format. The values are as follows:

- $1 = Region = ENCODE target region identifier
- $2 = Location = Location identifier (provided by Elliott)
- $3 = Name = Name of file provided by Elliott
- $4 = NoSp = Number of taxa present in the MCS alignment
- $5 = AlnLength = Length of MCS alignment (including gaps in human)
- $6 = FulllnL = log likelihood of tree when all branches estimated
- $7 = RatiolnL = log likelihood of tree when forced to adopt relative branch lengths of 4D tree
- $8 = 2Delta = Twice the likelihood improvement obtained by allowing all branches to vary (FulllnL) over enforcing 4D ratios (RatiolnL). This can be used for signifance testing on standard χ^2 tables with a number of degrees of freedom = (2*NoSp) – 2
- $9 - $51 = 1 – 43 = Signed improvement in likelihood achieved when letting this branch vary (+ = branch grows {red}; - = branch shrinks{blue})

Methodological details (light due to time constraints)

All the analysis presented is performed on home grown software I have hacked together over the last couple of days. It may potentially (inevitably…) contain some errors. From checking numerous examples I think the results in general are right, although I would warn that any single analysis could potentially contain an error (e.g. optimisation problems; data faults; &c).

All analyses are performed using the popular HKY model of evolution, with the frequencies of DNA characters counted from the data, and the transition/transversion ratio estimated using ML (see any phylogeny book/review for details). This is not ideal (it doesn't include rate variation for a start!), but I personally would not expect the results to change dramatically if a more complex model were used.

The methodology requires three basic tree models:

i) H_full: A tree with all of its branches fully optimised

ii) H_ratio: A tree with its branch ratios fixed to those of the 4D tree, with the overall length of branches allowed to vary. This is the equivalent to fixing branches to the values of the 4D tree and allowing a single universal scaling factor

iii) H_ratio_branch: As H_ratio, but with one extra degree of freedom allowing a single branch to vary (e.g. the branch leading to human).

The improvement in likelihood (model fit) this extra parameter provides is indicative of how much this branch wants to vary away from the 4D ratio. The improvements of H_ratio_branch over H_ratio in different branches of the phylogeny are the values plotted in the "thermoplots" and the values contained in $9-$51 of the results.all file.

I am also aware that there are many interesting potential directions to follow, although any suggestions are very much welcome! These results only represent what I have achieved so far.