WELLHAM PAD, FLORA, FL.

FURTHER CONCEPTS FOR PHYLOGENETICS AND WIDER BIOLOGY

6TH APRIL 2021

FURTHER CONCEPTS FOR PHYLOGENETICS AND WIDER BIOLOGY

P. A. D. Wellham[1] & F. L. Flora[2]

[1] School of Pharmacy, University of Nottingham, Nottingham United Kingdom
[2] BioFortify Research Institute, Nottingham, United Kingdom

This paper is an exploration of further concepts to apply to wider biological studies based upon phylogenies and other structures. This is a follow-on article from part one [1].

The concepts described here are:

- (i) Lineage Rationality;
- (ii) Biological Entities; and with that context the wider
- (iii) Biological Rationality; and
- (iv) Anchored Branch Factors.

(i) Lineage Rationality

Lineage rationality is an arbitrary metric which uses inputs from a given cladogram to create an infinite fraction which can describe the representation of a biological entity in a lineage (such as a taxon) in a given context. To outline this metric, an example cladogram is given below.

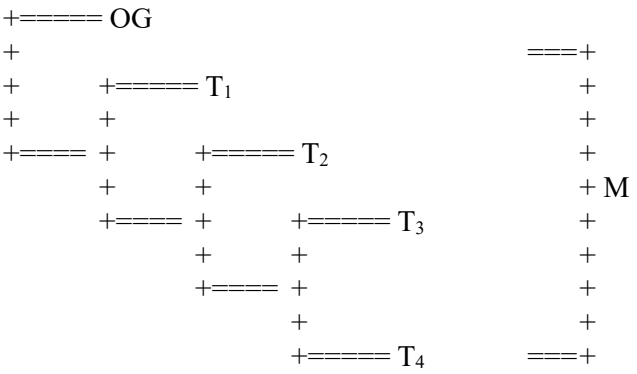

Figure 1: Example Cladogram

From the example cladogram, upwards lineage rationality for taxon T_4 with respect to lower taxon S and fellow taxa T within the higher taxon M , or $LR[T_S , T_4 \uparrow M]$ is characterised by the reciprocal of the infinite continued fraction $[N_{S(T4)} ; N_{S(T3)} , N_{S(T2)} , N_{S(T1)} , 1 , 1 , ...]$ whilst the downwards equivalent $LR[T_S , T_4 \downarrow M]$ is characterised by the reciprocal of the infinite continued fraction $[N_{S(T1)} ; N_{S(T2)} , N_{S(T3)} , N_{S(T4)} , 1 , 1 , ...]$.
$LR[T_S , T_1 \uparrow M]$ is characterised by the reciprocal of the infinite continued fraction $[N_{S(T1)} ; N_{S(T2)} + N_{S(T3)} + N_{S(T4)} , 1 , 1 , ...]$.

Where $N_{S(T)}$ is the number of taxon S in given higher taxon T.

To extend the example, we shall say $N_{S(T1)} = 4$, $N_{S(T2)} = 3$, $N_{S(T3)} = 26$, and $N_{S(T4)} = 12$.
Given this,
$LR[T_S , T_4 \uparrow M] = 1 / [12; 26, 3, 4, 1, 1, \ldots] \approx 0.083070$
$LR[T_S , T_4 \downarrow M] = 1 / [4; 3, 26, 12, 1, 1, \ldots] \approx 0.230994$
$LR[T_S , T_1 \uparrow M] = 1 / [4; 41, 1, 1, \ldots] \approx 0.248507$

These can be visually represented in “computer daisies” [2]. In crude terms a tighter spiral given by a “more irrational” continued fraction reciprocal indicates poorer representation of the relevant taxon.

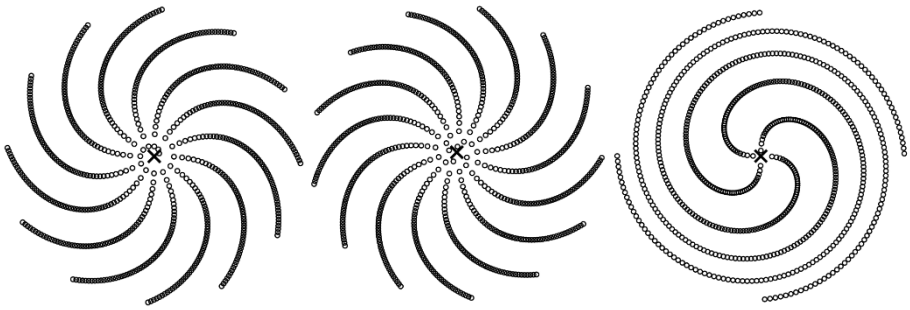

Figure 2: Computer Daisies Representing Lineage Rationalities from the Example Cladogram. Left to right - LR*[T_S, $T_4 \uparrow M$], LR[T_S, $T_4 \downarrow M$], LR[T_S, $T_1 \uparrow M$].*

(ii) Biological Entities

Entities in biology are subject to replication and transfer at different levels of organisation. They are either physically discrete objects, such as nucleotides, proteins, or metabolites, sets of discrete objects, or sets of smaller sets. There are two main types of sets: *collections* and *supercategories*. Sets can comprise a *collection* of *items* – for example a genome can be described as a collection of genes, gene clusters, or genomic regions – where items in themselves do not represent a phylogeny or lineage; or alternatively can be described as a *supercategory* of interrelated *subcategories* between which phylogenetic relationships can be drawn – for example a gene family can be described as a supercategory of genes or a genera as a subcategories of a family. Items are transferred to and between different collections by horizontal transfer; subcategories are replicated and transferred vertically within their supercategories.

Classification of Biological Entities:

cd/aa } gm/pcs } pd } g/t/p } gc/msp/m } Gr } Gm/Tm/Pm } In) Po) S]G]F]O]C]P]K
cd/aa } gm/pcs } pd } g/t/p } gc/msp/m } ph } Phm } In) Po) S]G]F]O]C]P]K
cd/aa } gm/pcs } pd] gf/pf] gf
cd/aa } gm/pcs } pd } g/t/p } gc/msp/m] mf
cd/aa } gm/pcs } pd } g/t/p } gc/msp/m } ph } Phm: pr/mr/bh } Dev/LH/Eco
Eco { In/Po/S etc.

[Lineages and Vertical Transfers]:
Phylogenetic lineages
Biomolecule lineages

Horizontal Transfers:

cd, gm → g :	mutation
g, gc → Gr :	transposable elements
g, gc, Gr → Gr, Gm :	sexual and parasexual events
g, gc, Gr → In :	horizontal genetic transfer
Gm → In, Po, S :	(endo)symbiosis
In → Po :	migration
Po → S :	isolation, speciation
bh → In, Po :	learning, culture

Informational Transfers:
cd → aa; gm → pcs; g → t → p; gf → pf; Gm → Tm → Pm : transcription and translation
gc → msp/m : transcription and translation + protein function, enzymatic activity etc.

Key: } – "is an item of"; { – "is a collection of" ;] – "is a subcategory of"; [– "is a supercategory of";) – "is an item or subcategory of"; (– "is a collection or supercategory of"; → – "can be transferred to or between".

Abbreviations: cd – codon; aa – amino acid; gm – gene motif; pcs – protein conserved site; pd – protein domain; g – gene; t – transcript; p – protein; gf – gene family; pf – protein family; gc – gene cluster; msp – metabolite synthesis pathway; m – metabolite; Gr – genomic region; Gm – genome; Tm – transcriptome; Pm – proteome; ph – phenotype; Phm – phenome; pr – product; mr – morphological trait; bh – behavioural trait; Dev –development of an organism; LH –life history of an organism; Eco – ecology of an organism; In – individual organism; Po – population; S – species; G – genus; F – family; O – order; C – class; P – phylum; K – kingdom.

(iii) Biological Rationality

The concept of lineage rationality can be more generalised to include other biological entities, namely those which do not form taxonomic phylogenetic lineages (i.e. composed of subcategories and supercategories) but are seen in the context of collections of items.

The formula for lineage rationality (*LR*) as example

$$LR[T_S\,,\ T_n \uparrow M] = [N_{S(Tn)};\ N_{S(Tn-1)}\,,\ \ldots\,,\ N_{S(T2)}\,,\ N_{S(T1)}\,,\ 1\,,\ 1\,,\ \ldots]$$

is modified to adjusted lineage rationality (*ALR*)

$$ALR[I,\ T_S\,,\ T_n \uparrow M] = [Mean(\ N_{S(Tn)}\ );\ Mean(\ N_{S(Tn-1)}\ )\,,\ \ldots\,,\ Mean(\ N_{S(T2)}\ )\,,\ Mean(\ N_{S(T1)}\ )\,,\ 1\,,\ 1\,,\ \ldots]$$

where $I \} S] T] M$ and $Mean(\ N_{S(T)}\ )$ values are the averages of those positions in the continued fraction which correspond accordingly to all incidences of item *I* in incidences of *S* from all relevant starting positions of *T*.
In the earlier example cladogram provided, for instance, where item *I* to be present in *T* taxa T_4 and T_1, in 3 and 5 incidences of *S* respectively, the formula for $ALR[I,\ T_S\,,\ T_n \uparrow M]$ would be calculated by
$ALR = [\ (\ 3\ .\ N_{S(T4)} + 5\ .\ N_{S(T1)}\ )/8;\ (\ 3\ .\ N_{S(T3)} + 5\ .\ (N_{S(T2)} + N_{S(T3)} + N_{S(T4)})\ )/8,\ (\ 3\ .\ N_{S(T2)} + 5\ )/8,\ (\ 3\ .\ N_{S(T1)} + 5)/8,\ 1,\ 1,\ \ldots\ ]$

Here in the formula for *ALR* all the cases for all taxa *T* which contain the item *I* are considered, and all those without it are excluded. For each term in order as written in the continued fraction, the values are averaged, regardless of where in the cladogram those values are represented from.

(iv) Anchored Branch Factors

As defined in a previous work [1], a branch factor can be formulated by

$$B_{S\{T\}\hat{}M} = B_{T\hat{}M}\ .\ N_{S(T)}$$

Where $S] T] M$.

In this general case we may term the larger taxon *M* as an "anchor". In specialised cases it may be helpful to set a context of branch factors, as well as their corresponding *PhyCo* values [1] around a smaller given taxon of interest (such as an endangered species) *E* where *S* is not a subcategory of *E*. We may in these cases set *E* as an anchor. This would require a modified formula for $B_{S\{T\}E}$ in which the components would consider the common ancestor between *S* and *E*.

$B_{S(T)E} = B_{T^P} . B_{E^P} . N_{S(T)}$
Where $S \rfloor P$ and $E \rfloor P$.

As described previously [1], *PhyCo* values, as well as its extensions, can then be calculated according to the cladogram in use given that the taxon P is consistent for all branch factors.

References

1. Wellham, P. A. D. & Flora, F. L. (2021): Phylogenetic and Ecological Coefficients for Biodiversity Assessments. drwellham.com/2021-03-01-wellham.
2. Dixon, R. (1981): The Mathematical Daisy. *New Scientist.*