Neutral versus adaptive MSA residual score calculator

Description

Compare divergence estimates from a reference between adaptively and neutrally evolving sequences to infer sequences that are more similar to the reference frame in the adaptive MSA than expected based on neutral evolution; identify the amino acid sites defining that similarity.

Usage

```
divergence_residual_scores(neut_fas,adap_fas,seed,n,ref,model_test=F,write=F)
```

Arguments

`neut_fas`
: a character matrix of an amino acid multiple sequence alignment for the neutrally evolving protein or protein segment.

`adap_fas`
: a character matrix of an amino acid multiple sequence alignment for the adaptively evolving protein or protein segment.

`seed`
: a character string of the amino acid sequence to seed the simulation of sequences on the neutrally evolving phylogeny.

`n`
: the number of sequence simulations to run.

`ref`
: a character name of the reference sequence from which divergence will be estimated.

`model_test`
: whether to run phangorn::modelTest to determine the best-fit subsitution model ('JTT' (default), 'WAG', 'LG') for building the neutrally evolving phylogeny.

`write`
: whether to write the residual scores for each sequence to a txt file.

Value

a list with the following components:

`neutral_tree`
: a phylo object of the tree built from the MSA of neutrally evolving sequences.

`model_fit`
: the best-fit subsitution model on the MSA of neutrally evolving sequences.

`residual_scores_across_seqs`	a dataframe of the residual scores calculated for each sequence as an estimate of the expected (i.e., neutral) versus actual (i.e., adaptive) divergence from a reference sequence.
`positive_conservation_of_sites`	a list of the percentage of sequences in the >75th quantile of residual scores with residues matched to the reference sequence.
`negative_conservation_of_sites`	a list of the percentage of sequences in the <25th quantile of residual score with residues matched to the reference sequence.
`positive_alignment`	a character matrix MSA of sequences in the top 25th quartile with positive residual scores.
`negative_alignment`	a character matrix MSA of sequences in the top 25th quartile with negative residual scores.
`sites_of_interest`	a list of amino acid sites that are more conserved among sequences in the >75th quantile (>=50%) than in the <25th quantile (<50%) of residual scores.
`positive_seq_logo`	strings of sequences in the >75th quantile of residual scores restricted to sites of interest.

Author(s)

E Lewitus

References

Lewitus, E., Bai, H., and Rolland, M. Design of a pan-betacoronavirus vaccine candidate through a phylogenetically-informed approach.

Examples

```
## calculate residual scores between the neutrally evolving S2 and adaptively evolving RBD of betaCoV

        #Spike amino acid MSA for betaCoV
        fas<-bio3d::read.fasta('betaCoV_spike_aa.fas')$ali

        #Remove gaps in the reference sequence
        ref_fas<-fas[,which(fas[ref,]%in%LETTERS)]

        #neut_fas: restrict the MSA to S2
        neut_fas<-ref_fas[,S2_coordinates]

        #adap_fas: restrict the MSA to RBD
```

```
adap_fas<-ref_fas[,RBD_coordinates]

n=1e3

#seed: define the reference sequence on the adap_fas MSA
seed<-adap_fas[ref,]

#run
run_drs<-divergence_residual_scores(neut_fas,adap_fas,seed,
n=n,ref='SARS2',model_test=F,write=T)
```