Genomic Evolutionary Rate Profiling (GERP)

Objective: to find constraint regions in genome subject to purifying selection

Procedure overview:

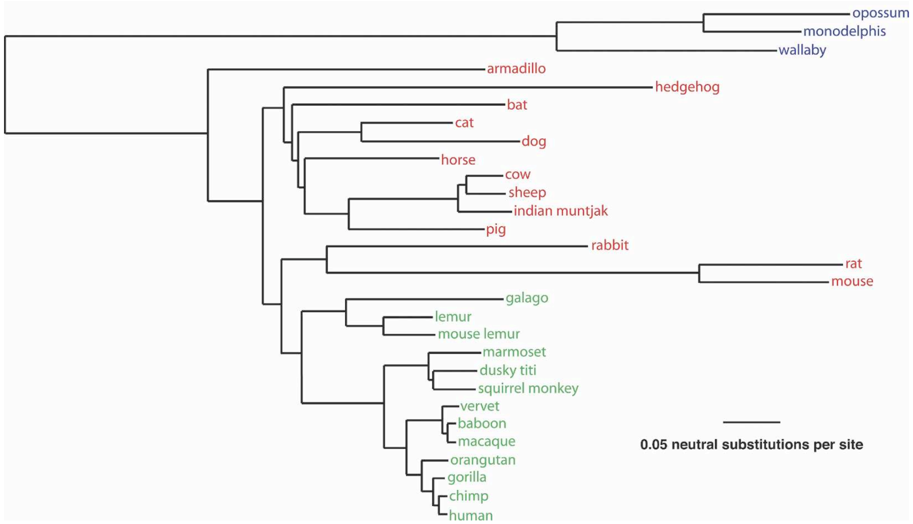
- use evolutionary tree and multiple alignments to estimate conservation scores (rejected substitution score) on a column-by-column basis.
- Constrained elements are stretches of the multiple alignment where the sequences are highly conserved according to the previous score.

Input

1. Multiple FASTA Alignment

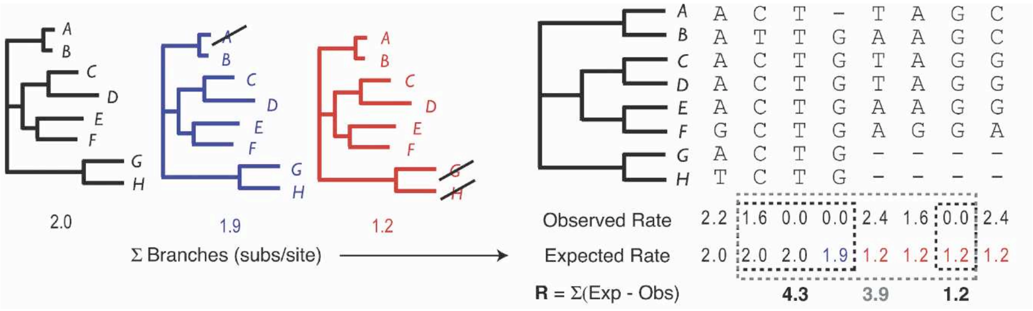
```
>human
ACTTACTTATCTAATGAAAAGTGCCCAGCATAAAAATGCAGGAGACAGACTTCCTTAGCCACCAGAGGCATCTTC
>chimp
- - TTACTTATCTAATGAAAAGTGCCCAGCATAAAAATGCAGGAGACAGACTTCCTTAGCCACCAGAGGCATCTTC
>colobus_monkey
ACTTACTTATCTAATGAAAGGTGCCCAGCATAAAAATGCAGGAGACAGACTTCCTTAGCCACCAGAGGCATCTTC
>baboon
ACTTACTTATCTAATGAAAGGTGCCCAGCATAAAAATGCAGGAGACAGACTTCCTTCGCCACCAGAGGCGTCTTC
>macaque
- - TTACTTATCTAATGAAAGGTGCCCAGCATAAAAATGCAGGAGACAGACTTCCTTCGCCACCACAGGCGTCTTC
... ...
```

2. Evolutionary Tree



```
(((opossum:0.034160,monodelphis:0.022496):
0.032669,wallaby:0.049042):0.166848,armadillo:0.063747,
((hedgehog:0.083869,(ajbat:0.046820,(((cat:0.021296,dog:
0.035335):0.013827,(horse:0.029230,(((cow:0.007866,sheep:
0.008974):0.002377,muntjak_indian:0.011920):0.024259,pig:
0.030216):0.009758):0.001057):0.000798):0.001802):
0.004810,((rabbit:0.057760,(mouse:0.029443,rat:0.032846):
0.085321):0.011253,((galago:0.035364,(lemur:
0.010850,mouse_lemur:0.011975):0.008347):0.009622,
((marmoset:0.011613,(dusky_titi:0.011282,squirrel_monkey:
0.009659):0.000548):0.009918,((vervet:0.003591,(baboon:
0.002469,macaque:0.002468):0.001311):0.007575,(orangutan:
0.005625,(gorilla:0.002932,(chimp:0.002271,human:0.002259):
0.000476):0.003057):0.003284):0.005489):0.018117):
0.004605):0.004931):0.011979);
```

Calculating RS Score at Each Nucleotide



A	A	C	T	–	T	A	G	C
B	A	T	T	G	A	A	G	C
C	A	C	T	G	T	A	G	G
D	A	C	T	G	T	A	G	G
E	A	C	T	G	A	A	G	G
F	G	C	T	G	A	G	G	A
G	A	C	T	G	–	–	–	–
H	T	C	T	G	–	–	–	–
Observed Rate	2.2	1.6	0.0	0.0	2.4	1.6	0.0	2.4
Expected Rate	2.0	2.0	2.0	1.9	1.2	1.2	1.2	1.2
R = Σ(Exp - Obs)			**4.3**		3.9		**1.2**	

At each column of multiple sequence alignment:

1. Expected rate: sum of residue branches after elimination of gapped sequence
2. Observed rate: maximum likelihood of substitution count
3. RS = Expected - Observed

Finding Constraint Elements

Overview: Find and report a list of elements that appear constraint beyond what is likely to occur by chance.

Generate a list of candidate constrained elements that fit the following criteria

- Starts and ends on a position with positive RS score.
- Length between L_{min} and L_{max}, which are defaults to 4 and 2000, respectively.
- Score is at least (neutral_rate / q) * length ^ r. Defaults: q = 10, r = 1.15.
- No more than a pre-allowed number (default = 10) of shallow columns in the middle of a longer shallow region

P-value computation

Each candidate element of length L and score S is assigned a p-value, corresponding to the probability that a score of at least S occurs at random in a block of L positions

False discovery rate estimation

To estimate the number of false positives at a given p-value, the program randomly shuffles the positions of the alignment, apply previous steps to generate elements, and consider those elements as false positives.

Implementation

The latest version of GERP (v2.1) is implemented in C

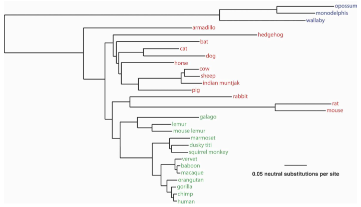
It can be downloaded from: http://mendel.stanford.edu/SidowLab/downloads/gerp/

Documents and test data are also available

Input

>human
ACTTACTTATCTAATGAAAAG
>chimp
- - TTACTTATCTAATGAAAAG
>colobus_monkey
ACTTACTTATCTAATGAAAGG
>baboon
ACTTACTTATCTAATGAAAGG
... ...

+



Intermediate

Exp.	RS
1.6	1.6
1.43	-0.927
2.3	1.15
2.47	2.47
2.85	1.67
2.85	-0.869
2.85	0.0391
2.85	1.72
2.85	1.82
2.85	2.85
2.85	1.86
2.85	1.67
2.85	2.85
2.85	0.4
2.85	2.85
2.85	1.86
2.85	0.0849
2.85	-1.79
... ...	

Output

start	end	length	RS-score	p-value
337736	337925	190	530.3	3.28225e-154
334181	334348	168	484.2	1.50556e-146
285429	285610	182	480.8	2.91752e-131
262608	262862	255	574.4	9.7306e-131
284586	284739	154	435.5	4.80595e-129
294577	294689	113	344 5.	11345e-112
281531	281670	140	377	1.01916e-105
459574	460290	717	826.2	1.51076e-99
309564	309749	186	424.4	1.55877e-98
373288	373664	377	577.6	2.38141e-89
295079	295210	132	336.9	4.77357e-89
264434	264568	135	333.5	5.42598e-85
... ...				

GERP Features

Features:

- Ignore gaps from missing data
- Rank constraint elements by score and/or p-value
- Null model of substitution at single nucleotide resolution