FIU FLORIDA INTERNATIONAL UNIVERSITY

BIOINFORMATICS CAP5610

Phylogenetic Trees Analysis Pipeline “SemiTree”

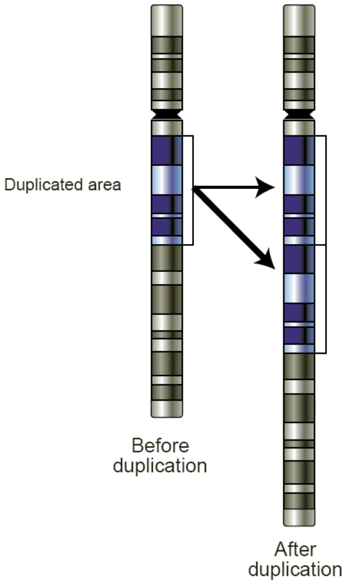
ANDRIUS BUBELIS 4957928

April 28, 2015

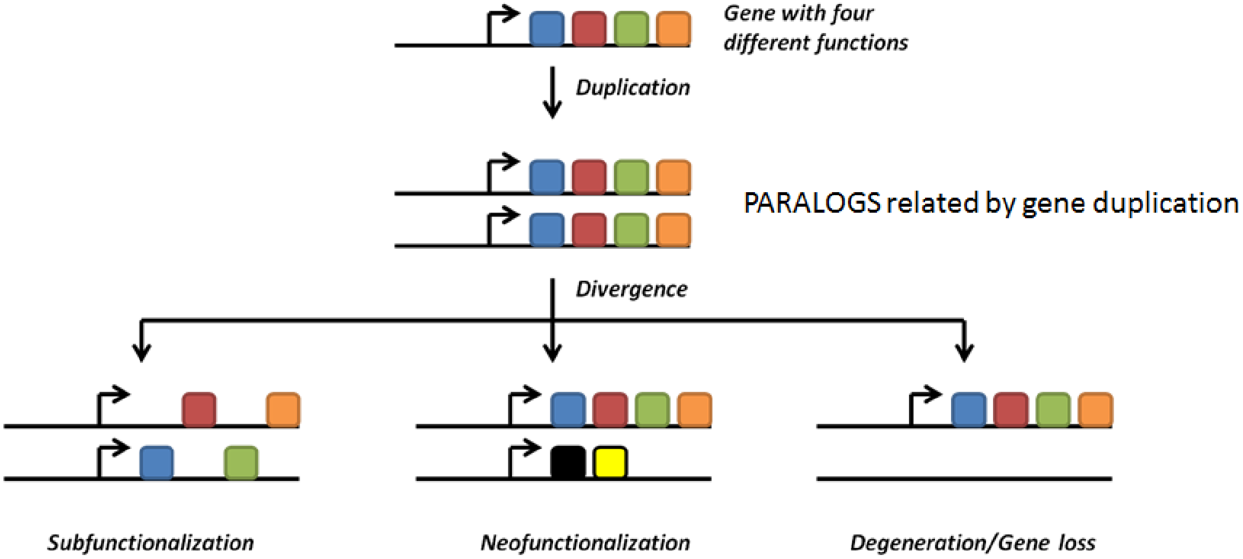
Gene Duplication is a major mechanism for evolution, and driving factor for gene **Functional Divergence** - process by which genes shift in function from an ancestral function.

Common sources of gene duplications:

- ***Ectopic Recombination*** could occur during meiosis when crossing over occurs between homologous chromosomes. There are cases when chromosomes do not align perfectly and unequal crossing-over occurs, causing one chromosome end up with the duplications of genes.
- ***Replication Slippage*** is an error in DNA replication due to DNA polymerase error. It happens when polymerase dissociates from the DNA and after reattaches back to the DNA strand, and it aligns the replicating strand to an incorrect position.
- ***Retrotransposition*** happens during cellular invasion by a replicating retroelement or retrovirus, Reverse-Transcriptase copies their genome by reverse transcribing RNA to DNA.
- ***Whole Gene Duplication*** might be result of inheritance of two copies of its genome from each parent
- ***Aneuploidy*** occurs when nondisjunction at a single chromosome results in an abnormal number of chromosomes.



While selection takes away variation in the population, mutation adds back. After duplication, selective pressure relaxes, and each gene will accumulate its own mutations that result in **Functional Divergence**



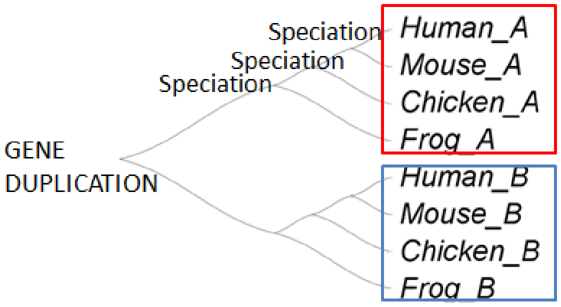
Gene Duplication → Gene Redundancy → Paralogs genes → Functional Divergence →

→ ***Subfunctionalization* -** division of ancestral function

→ ***Neofunctionality*** – newly acquired mutations will lead to new functions

→ ***Degradation*** - acquired mutations will result in gene loss by functional degradation

http://en.wikipedia.org/wiki/Gene_duplication

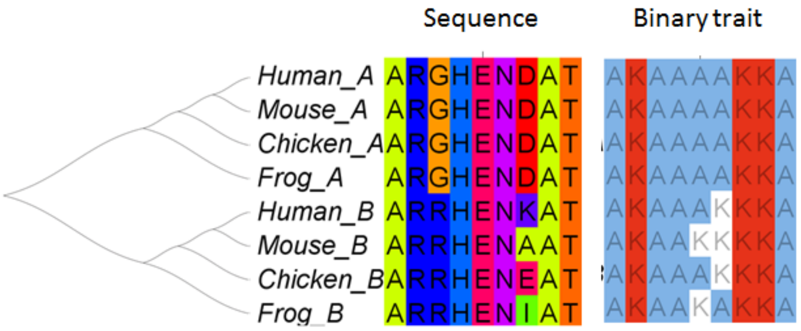


Genes related by speciation are called Orthologs

All _A's form an orthologous clade

All _B's form an orthologous clade

The main focus of my project is to create BioPython pipeline for quantitative comparative analysis between duplicated genes to help better understand functional divergence, by analyzing functional divergence between **Orthologous Clades**


Sequence
Binary trait
Human_A ARGHENDAT AKAAAAKKA
Mouse_A ARGHENDAT AKAAAAKKA
Chicken_A ARGHENDAT AKAAAAKKA
Frog_A ARGHENDAT AKAAAAKKA
Human_B ARRHENKAT AKAAAKKKA
Mouse_B ARRHENAAT AKAAKKKKA
Chicken_B ARRHENEAT AKAAAKKKA
Frog_B ARRHENIAT AKAAKAKKA

Project aim:

1. Identify sites (positions) conservation:
    1. Amino Acid
    2. physicochemical properties
    3. Trait (e.g. order/disorder by IUPred).
2. Identify conserved regions within and between clades
3. Identify sites (positions) changing rates
4. Create quantitative measure to comparing orthologous clades
5. Generate various output files for next step in the pipeline

Running SemiTree:

Project GitHub repository:

https://github.com/abube002/SemiTree---analysis-of-phylogenetic-trees

Download and install Python 2.7 https://www.python.org/downloads/

Download and install additional Libraries:

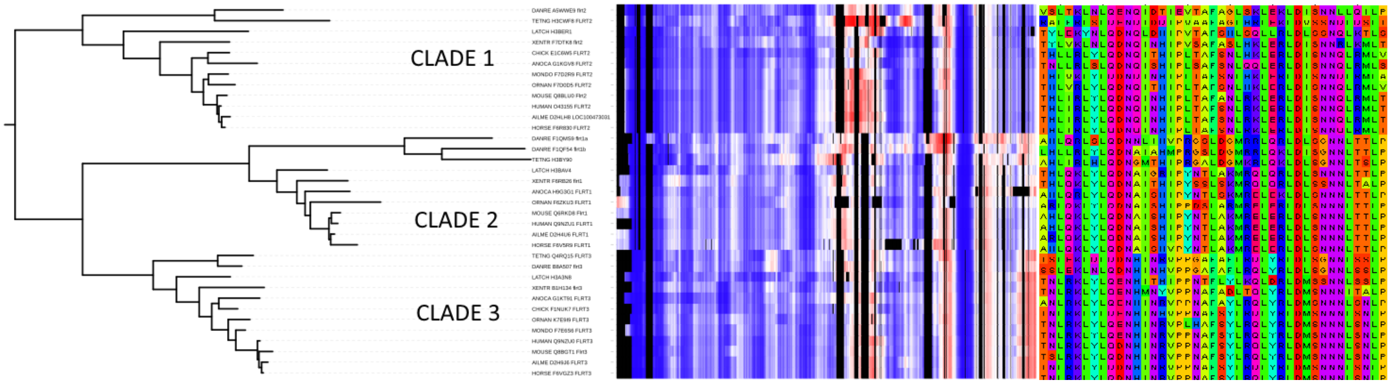
- BioPyhon www.biopython.org
- NumPy www.numpy.org

Copy All SemiTree Files into the same directory (for Windows: **semitreep.py** , **Clade.py** and **7** testing data files):

- **semitreep.py** (main)
- **Clade.py** (class *Clade* – creates orthologous clade objects)
- **Rate4site** Linux executable (modified and recompiled for Linux only)
  - ○ Source code was available only for Linux
  - ○ SemiTree checks OS, if not Linux then looks for pre-generated rate4site files

Test Data Files (*FLRT2* fibronectin leucine rich transmembrane protein):

- **F1QF54.tre** – file tree
- **F1QF54.odo** – corresponding order disorder file
- **F1QF54.fa** – corresponding fasta MSA



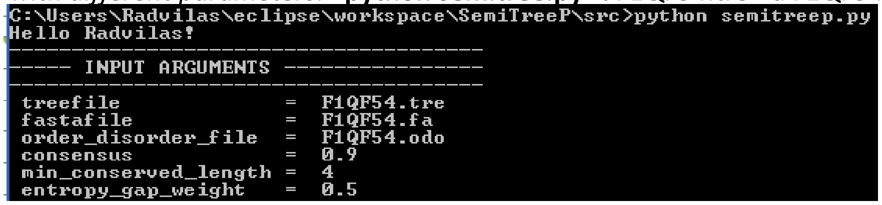
- Rate4site pre-generated data files needed to run on **Windows** (not needed on Linux)
  - ○ **F1QF54_CLADE_1.r4s**
  - ○ **F1QF54_CLADE_2.r4s**
  - ○ **F1QF54_CLADE_3.r4s**
  - ○ **F1QF54_ALL_CLADES.r4s**

Parameters:

Shor	Long	Default	Description
-t	--input_tree	F1QF54.tre	newick tree file
-a	--input_alignment	F1QF54.fa	fasta MSA file corresponding the tree
-o	--input_odo	F1QF54.odo	order/disorder file (or any other discreet trait file)
-c	-- consensus	0.9	consensus % for conservation, [0.0-1.0]
-m	--minimum_aa	4	the minimum number of amino acids for conserved region
-e	--entropy_gap_weight	0.5	entropy gap weight [0.0-1.0] e.g 0.5 => treats gaps as 50% conserved

Execute from command line: **python semitree.py**

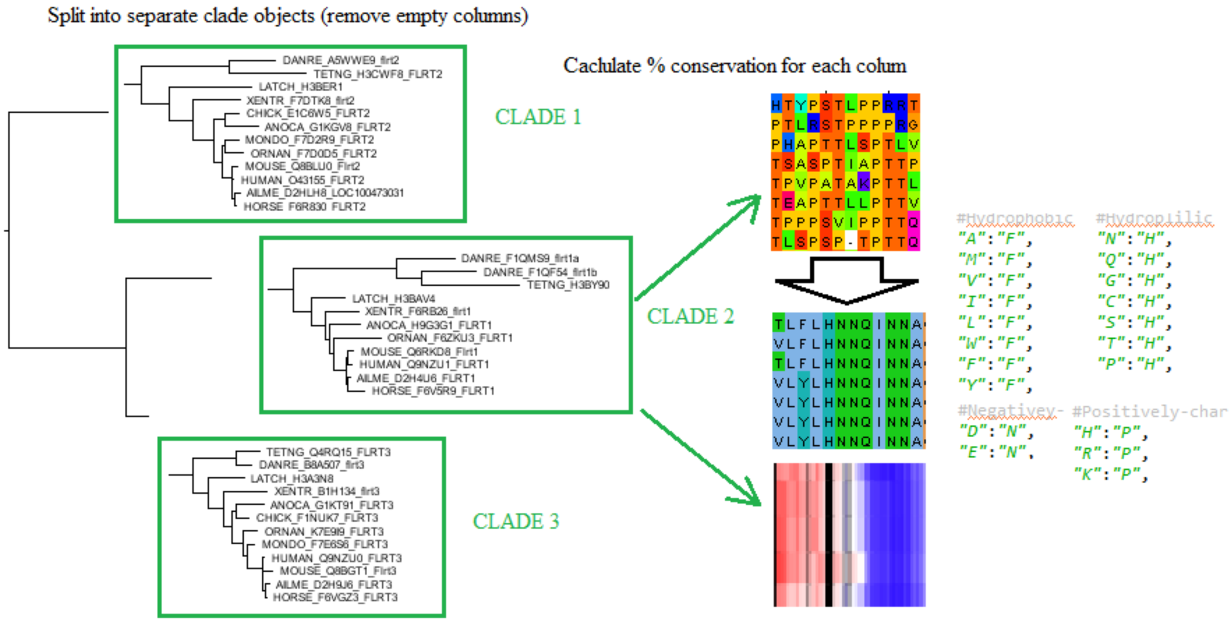
with different parameters: **python semitree.py** -t F1QF54.tre -a F1QF54.fa -o F1QF54.odo -c 0.9 -m 4 -e 0.5



```
C:\Users\Radvilas\eclipse\workspace\SemiTreeP\src>python semitreep.py
Hello Radvilas!
------------------------------------
----- INPUT ARGUMENTS ----------------
------------------------------------
 treefile                =  F1QF54.tre
 fastafile               =  F1QF54.fa
 order_disorder_file     =  F1QF54.odo
 consensus               =  0.9
 min_conserved_length =  4
 entropy_gap_weight    =  0.5
```

Algorithm explained by test example: FLRT2 fibronectin leucine rich transmembrane protein

1. Located Tree’s Internal nodes with names CLADE_ then splits Tree into Clade objects:
2. Creates sub-trees (objects and .tre files) for each clade
3. Finds corresponding sequences (FASTA) and traits (O/D) files and maps data to Clade object
4. Removes empty columns
5. For each clade generates Physicochemical mapping matrix



```
HTYPSTLPPRRT
PTLRSTPPPPRG
PHAPTTLSPTLV
TSASPTIAPTTP
TPVPATAKPTTL
TEAPTTLLPTTV
TPPPSVIPPTTQ
TLSPSP-TPTTQ
```

```
TLFLHNNQINNA
VLFLHNNQINNA
TLFLHNNQINNA
VLYLHNNQINNA
VLYLHNNQINNA
VLYLHNNQINNA
VLYLHNNQINNA
```

```
#Hydrophobic      #Hydropilic
"A":"F",          "N":"H",
"M":"F",          "Q":"H",
"V":"F",          "G":"H",
"I":"F",          "C":"H",
"L":"F",          "S":"H",
"W":"F",          "T":"H",
"F":"F",          "P":"H",
"Y":"F",

#Negativey-  #Positively-char
"D":"N",     "H":"P",
"E":"N",     "R":"P",
             "K":"P",
```

6. Calculates conservation for each column in matrixes:
   a. Physicochemical
   b. Amino Acid
   c. O/D

   Algorithm: For each column finds most frequent character and assigns a frequency (0-1)

7. Finds conserved regions based on specified –m (min number of amino acids) and –c consensus (0-1)
   a. Physicochemically conserved regions
   b. Amino Acid conserved regions
   c. Generates conserved region summaty files for each clades:

F1QF54_CLADE_1_conserved_reg.txt

CLADE NAME: CLADE_1
MIN REGION LEN: 4
CONSENSUS: 0.9
REGION COUNT: 26

ID	Pos	Len	AA
1	76	5	CRCDR
2	82	16	FVYCNERSLTSVPLGI
3	106	12	LHNNQINNAGFP
4	128	13	TVYLYGNQLDEFP
5	142	27	NLPKNVRVLHLQENNIQTISRAALAQL
6	171	17	LEELHLDDNSISTVGVE
7	189	6	GAFREA
8	204	8	KNHLSSVP

F1QF54_CLADE_1_conserved_phychm_reg.txt

CLADE NAME: CLADE_1
MIN REGION LEN: 4
CONSENSUS: 0.9
REGION COUNT: 29

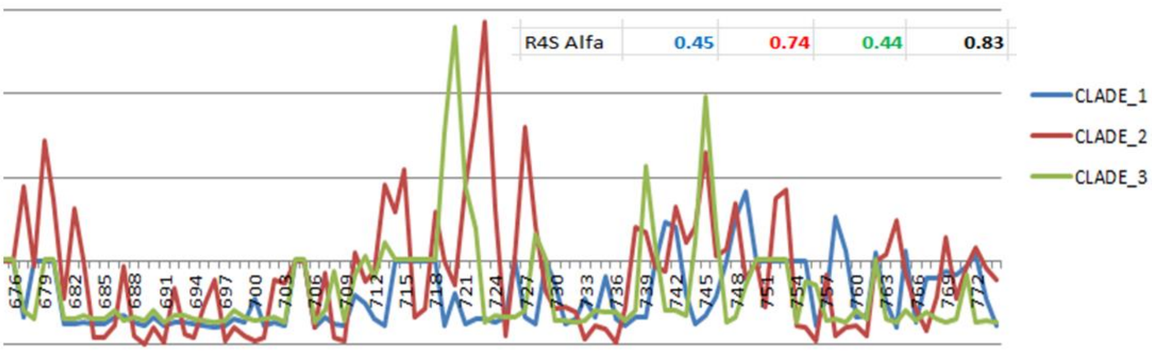
ID	Pos	Len	PhysicoChemical properties
1	76	26	HPHNPHFFFHHNPHFHHFHFHFHNHF
2	104	18	FFFPHHHFHHFHFHFNFP
3	128	41	HFFFFHHHFNNFHFHFHPHFPFFPFHNHHFHHFHPFFFFHF
4	170	47	PFNNFPFNNHHFHHFHFNNHFFPNFFHFPFFFFHPHPFHHFHFHFHF
5	220	10	NFPFNNHPFF
6	235	6	FFHHFH
7	244	9	PFFFNHHFF
8	260	4	HHFH

8. Rate4Site *http://www.tau.ac.il/~itaymay/cp/rate4site.html*
    a. Rate4Site is a program for detecting conserved amino-acid sites by computing the relative evolutionary rate for each site in the multiple sequence alignment (maximum likelihood method) given a certain phylogenetic tree.
    b. Rate4Site skips position if referenced sequence has a gap. Modification was needed to C++ code.
    c. C++ source code was modified to include all sites
    d. Modified version of Rate4site executable was compiled and integrated into pipeline.
    e. Rate4site generated data files are parsed and stored for each clade object

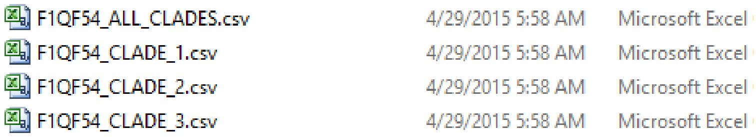
Example of Rate4Site file:

```
#POS SEQ  SCORE    QQ-INTERVAL     STD      MSA  DATA
#The alpha parameter 0.737709
#The likelihood of the data given alpha and the tree is:
#LL=-6866.52
   1     M -0.7742   [-0.904,-0.7515]   0.1947   12/12
   2     E -0.1974   [-0.678,0.1214]    0.6113   12/12
   3     F  0.9015   [-0.126, 1.721]    1.098    12/12
   4     Q   1.713   [0.1214, 4.013]    1.482    12/12
   5     T -0.4939   [-0.8081,-0.3161]  0.3533   12/12
   6     G   1.997   [0.4542, 4.013]    1.478    12/12
   7     F   3.127   [ 0.933, 4.013]    1.259    12/12
   8     W  0.2187   [-0.5841,0.4542]   0.9656   12/12
   9     N -0.09333   [-0.678,0.1214]   0.6208   12/12
```

    f. Summary file is generated to compare data from multiple clades F1QF54_r4s.txt:



9. Clade conservation comparison:
    a. Files for each clade (with gaps to maintain global original position)and global files are generated



F1QF54_ALL_CLADES.csv	4/29/2015 5:58 AM	Microsoft Excel
F1QF54_CLADE_1.csv	4/29/2015 5:58 AM	Microsoft Excel
F1QF54_CLADE_2.csv	4/29/2015 5:58 AM	Microsoft Excel
F1QF54_CLADE_3.csv	4/29/2015 5:58 AM	Microsoft Excel

    b. Aggregated values per position (site)
    Example:

```
COLUMNS:,0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
RATE4SITE:,1.215,1.357,0.8006,-0.4163,1.025,1.984,2.73,2.931,
AA_FREQ:,0.272727272727,0.272727272727,0.272727272727,0.363636363636,
PHCEM_FREQ:,0.363636363636,0.272727272727,0.363636363636,0.363636363636,
ODO_FREQ:,0.272727272727,0.272727272727,0.181818181818,0.272727272727,
AA_NEGENTROPY_NORMALIZED:,0.0630341004478,0.0630341004478,0.0630341004478,
AA_GENTROPY_Log2:,0.811278124459,0.811278124459,0.811278124459,0.0,0.811278124459,
```

c. **RATE4SITE** – parsed from rate4site output file

d. **AA_FREQ** – frequency of most frequent Amino Acid

e. **PHCEM_FREQ** - frequency of most frequent Physicochemical Property

f. **ODO_FREQ** - frequency of most Order or Disorder whichever highest

g. AA_GENTROPY_Log2

$$\text{Entropy} = -\sum_{aa} P_{aa} \log_2(P_{aa})$$

$$P = \frac{\textit{Amino Acid Count}}{\textit{number of rows}}$$

The challenge I faced was how to address the gaps?

I added an input parameter –e , range [0-1] that sets the percentage of conservation assigned to the gaps. Example, if e=0.5, then gaps are treated as 50% conserved:

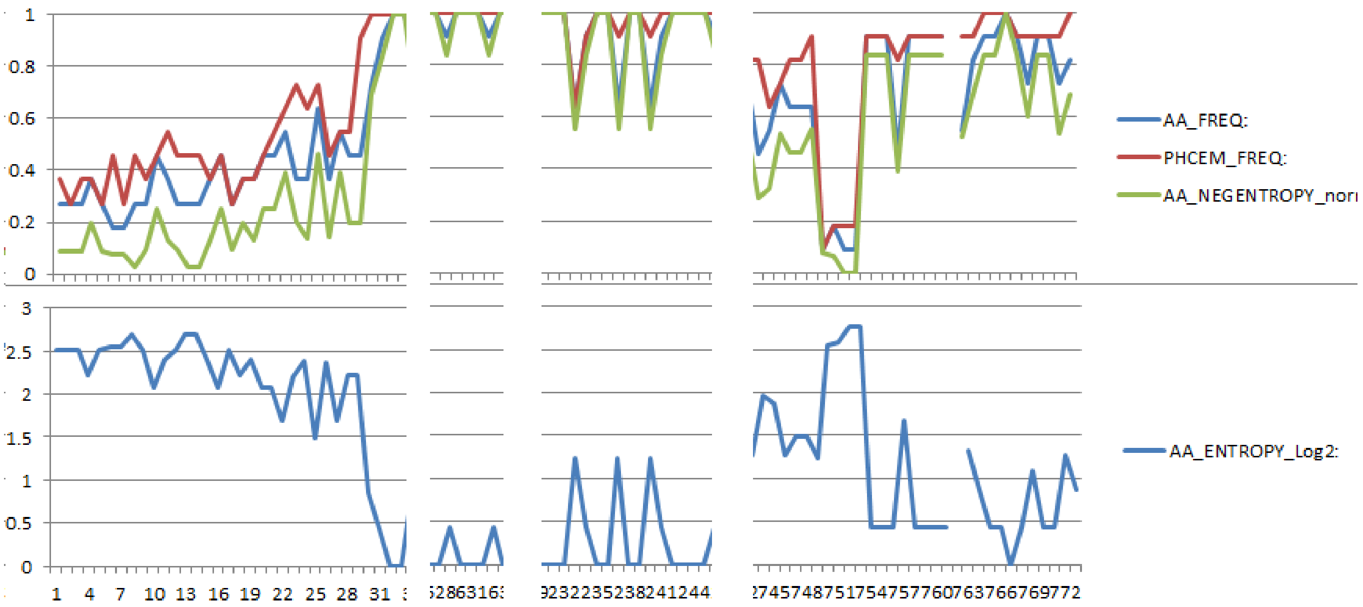
```
-    Gap = 0.4
-    A = 0.2
-    C = 0.3
-    T = 0.1
A
A    If e=0.5 then
C           Gap is split into:  Gap1 = 0.2 and Gap2 = 0.2
C
C    Calculate Entropy based on new distribution:
T    Gap1 = 0.2, Gap2 = 0.2, A = 0.2, C = 0.3, T = 0.1
```

I experimented with multiple values. The Weight of 0.5 seems to me performs the best.

So, the default for Gap weight is set to 0.5

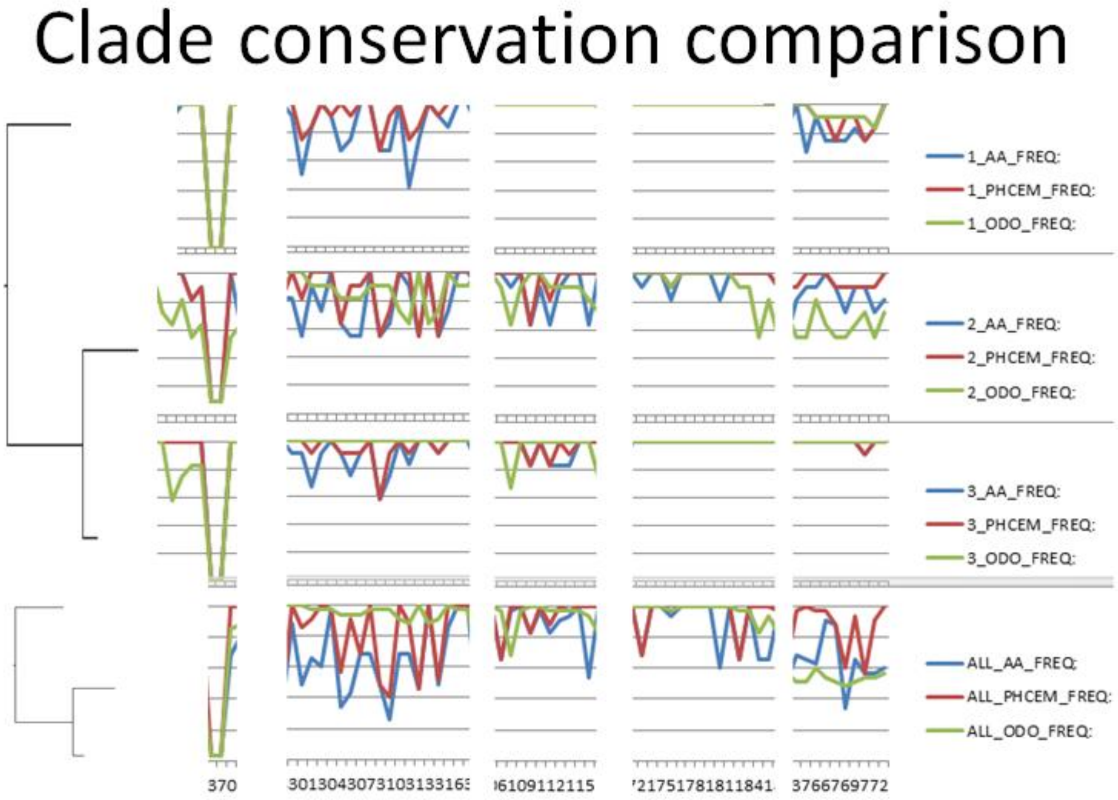
h. **AA_NEGENTROPY_NORMALIZED** = *(1 – ENTROPY)/*normalized by the highest entropy per clade. I implemented that so it could be plotted to a chart and it would be easy visually comparable to Amino Acid Conservation (range 0-1), where 1 represents 100% conservations in both functions

F1QF54_CLADE_2 example

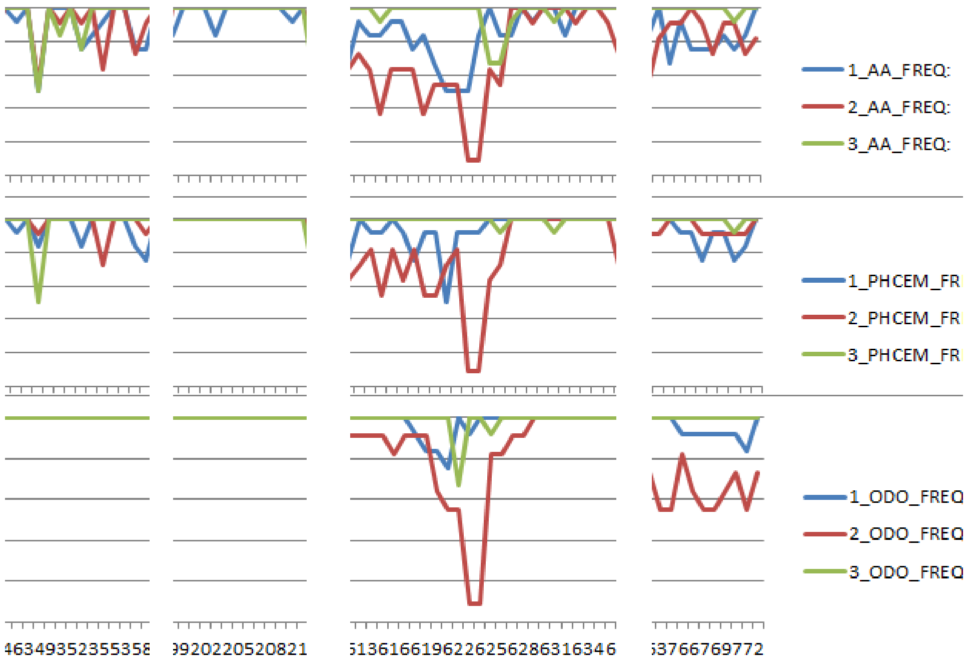


SemiTree allows differ ways to compare clades.

In chart below, we see 3 clades compared by different conservations types. Each row represents one clade and bottom (4^{th}) row is the Sum of all 3 clades. There are 4 different fragment to illustrated different divergence cases (we clearly can see Clade_2 is the most "unstable".



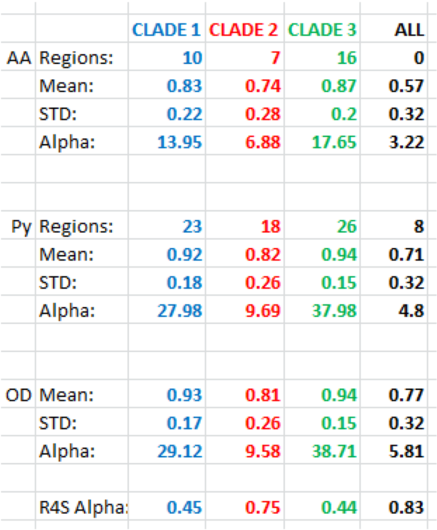
Charts below display separate conservation types and compares 3 clades within the same conservation



What did analysis find for the test data set?

CLADE 2 diverges much faster

Rate4Site is consistent with SemiTree findings (similar alpha ratio), that is a good indications that project is on the right path.



		CLADE 1	CLADE 2	CLADE 3	ALL
AA	Regions:	10	7	16	0
	Mean:	0.83	0.74	0.87	0.57
	STD:	0.22	0.28	0.2	0.32
	Alpha:	13.95	6.88	17.65	3.22
Py	Regions:	23	18	26	8
	Mean:	0.92	0.82	0.94	0.71
	STD:	0.18	0.26	0.15	0.32
	Alpha:	27.98	9.69	37.98	4.8
OD	Mean:	0.93	0.81	0.94	0.77
	STD:	0.17	0.26	0.15	0.32
	Alpha:	29.12	9.58	38.71	5.81
	R4S Alpha:	0.45	0.75	0.44	0.83

CREATED FILES :

Rate4Site global summary: F1QF54_r4s.txt
CLADE_1
tree: F1QF54_CLADE_1.tre
fasta: F1QF54_CLADE_1.fa
fasta without gaps: F1QF54_CLADE_1_nogaps.fa
amino acid conserved region summary: F1QF54_CLADE_1_conserved_reg.txt
physicochemical conserved region summary: F1QF54_CLADE_1_conserved_phychm_reg.txt
clade summary: F1QF54_CLADE_1_conserved_reg.txt
clade conservation CSV file: F1QF54_CLADE_1.csv

CLADE_2
tree: F1QF54_CLADE_2.tre
fasta: F1QF54_CLADE_2.fa
fasta without gaps: F1QF54_CLADE_2_nogaps.fa
amino acid conserved region summary: F1QF54_CLADE_2_conserved_reg.txt
physicochemical conserved region summary: F1QF54_CLADE_2_conserved_phychm_reg.txt
clade summary: F1QF54_CLADE_2_conserved_reg.txt
clade conservation CSV file: F1QF54_CLADE_2.csv

CLADE_3
tree: F1QF54_CLADE_3.tre
fasta: F1QF54_CLADE_3.fa
fasta without gaps: F1QF54_CLADE_3_nogaps.fa
amino acid conserved region summary: F1QF54_CLADE_3_conserved_reg.txt
physicochemical conserved region summary: F1QF54_CLADE_3_conserved_phychm_reg.txt
clade summary: F1QF54_CLADE_3_conserved_reg.txt
clade conservation CSV file: F1QF54_CLADE_3.csv

Example of Clade Summary file:

```
******** F1QF54 ** CLADE_2 CLADE ANALYSIS ***************
MIN REGION LEN: 4
CONSENSUS: 0.9
Shortest AA Length=628
Longest AA Length=693
CLADE Total Length=736
Rate4Site Alpha= 0.737709
--------------------------------------------------
Amino Acid Conservation, number of conserved regions:28
Amino Acid Conservation, Mean:0.736129905277
Amino Acid Conservation, STD:0.280648988251
Amino Acid Conservation, Alpha:6.87989738636
--------------------------------------------------
Amino Acid Entropy, Mean:0.936112773023
Amino Acid Entropy, STD:0.899420484857
Amino Acid Entropy, Alpha:1.08325522601
--------------------------------------------------
Physicochemical Conservation, number of conserved regions:32
Physicochemical Conservation, Mean:0.813999261902
Physicochemical Conservation, STD:0.2674570904
Physicochemical Conservation, Alpha:9.26274675976
--------------------------------------------------
O/D Conservation, Mean:0.813117588933
O/D Conservation, STD:0.262803561294
O/D Conservation, Alpha:9.57291535737
```

Goals accomplished:

- Locate specified clades within a phylogenetic protein tree
- Calculate amino acid, physicochemical and trait conservation per site and per clade
- Locating amino acid and physicochemically conserved regions given consensus
- Modified and integrate Rate4Site
- Quantitative comparative analysis between clades.
- Quick way to analyze massive trees and a massive number of trees
- Creates various data files for next level statistical analysis
- Integrated in a bigger project that is being developed in Dr.Liberles lab

Where an application could be used?

- Identifying residues that drive functional divergence or residues that are universally important for fold and function
- Show changes in conservation in protein families
- Important for understanding protein evolution
- Understanding functional divergence
- Clustering sequenced based on conservation
- As inner piece for larger pipeline implementation

Acknowledgments

- Dr. Giri Narasimhan
- Siltberg-Liberles Lab
  - Dr. Jessica Liberles
  - Dr. Helena Gomes Dos Santos
  - Joseph Ahrens
  - Janelle Nunez-Castilla
  - Itay Mayrose – for confirming where to modify the Rate4Site program