Description of the input and output data for main calculation nodes of the main pipelines

Analysis of proteins evolution

1) Multiple protein alignment (Alignment node):
    - Input:
        - unaligned protein sequences in FASTA format
    - Output:
        - aligned protein sequences in FASTA format
2) Amino acid substitution matrix generating (Amino acid substitution model estimation node):
    - Input:
        - aligned protein sequences in FASTA format
    - Output:
        - custom amino acid substitution matrix (symmetric matrix containing relative rates of amino acid substitutions) in PAML format (Fig. 1)

```
0.000020
0.000020  0.000020
0.014075  0.001943  0.032909
0.003481  0.022949  0.000040  0.000758
0.000020  0.046975  0.000040  0.000024  0.000036
0.000020  0.000020  0.000040  0.081092  0.000036  0.123187
0.020578  0.008522  0.006311  0.011390  0.002851  0.000074  0.010881
0.006453  0.110002  0.049155  0.066571  0.000036  0.377067  0.000021  0.030764
0.005281  0.000020  0.004010  0.000024  0.001146  0.009836  0.000021  0.001865  0.008448
0.001751  0.000173  0.000040  0.003031  0.002572  0.000074  0.003605  0.000025  0.011188  0.021236
0.002322  0.079895  0.021091  0.000024  0.000036  0.042718  0.023816  0.009368  0.000102  0.000034
0.001529
0.000020  0.011562  0.021295  0.000024  0.000036  0.003879  0.000021  0.002582  0.016682  0.018843
0.053856  0.013311
0.000020  0.002781  0.000040  0.002455  0.018291  0.000074  0.000021  0.000025  0.000102  0.000034
0.010372  0.001051  0.020307
0.030737  0.003357  0.000040  0.000024  0.000036  0.043882  0.000021  0.003973  0.000102  0.000034
0.002636  0.000029  0.001572  0.000029
0.034785  0.010441  0.075339  0.000024  0.024069  0.000456  0.001653  0.023063  0.000102  0.000034
0.000010  0.000029  0.006416  0.003684  0.010481
0.055411  0.016406  0.042028  0.000024  0.000036  0.032223  0.007550  0.003251  0.000102  0.010248
0.004747  0.007933  0.019027  0.000029  0.007911  0.059533
0.000020  0.000020  0.000040  0.000211  0.000036  0.000074  0.000021  0.000564  0.000102  0.001449
0.009109  0.000029  0.000063  0.000029  0.001652  0.001600  0.000036
0.007481  0.000020  0.045916  0.000024  0.054378  0.000074  0.002142  0.000025  0.208757  0.000034
0.000010  0.002613  0.009735  0.124881  0.000025  0.000015  0.004529  0.032900
0.032977  0.000622  0.000040  0.000024  0.000036  0.009031  0.004701  0.006030  0.000102  0.122295
0.014349  0.000029  0.000063  0.013511  0.000025  0.002715  0.000111  0.000738  0.000046

  0.0672743055555551  0.0674913194444453  0.0336371527777787  0.0551215277777782  0.0373263888888907
0.0182291666666666  0.0653211805555558  0.0546875000000003  0.0132378472222225  0.0399305555555538
0.1304253472222177  0.0460069444444447  0.0212673611111116  0.0462239583333345  0.0536024305555547
0.0909288194444444  0.0371093749999995  0.0169270833333359  0.0295138888888900  0.0757378472222200
```

Рис. 1. Custom amino acid substitution matrix in PAML format.

3) Phylogram reconstruction (Build tree node):
    - Input:
        - aligned protein sequences in FASTA format
        - custom amino acid substitution matrix in PAML format or standard amino acid substitution matrix (JTT, WAG, LG, etc.)
        - (optionally) initial cladogram in NEWICK format
    - Output:

- phylogram in NEWICK format (unrooted)
- shape parameter (alpha) for the gamma distribution of amino acid substitution rates

4) Ancestral sequence reconstruction (Ancestral reconstruction node):
   - Input:
     - gapless alignment of protein sequences in FASTA format
     - custom amino acid substitution matrix in PAML format or standard amino acid substitution matrix (JTT, WAG, LG etc.)
     - rooted cladogram in NEWICK format
   - Output:
     - gapless alignment of protein sequences in FASTA format with reconstructed ancestral sequences
     - rooted cladogram in NEWICK format containing internal node labels corresponded to the reconstructed ancestral sequences
     - reconstruction probabilities of ancestral amino acids for each internal node of the tree represented in FASTA-like format
5) Searching for statistically rare (atypical) classes of amino acid substitutions (Rare changes detection node):
   - Input:
     - gapless alignment of protein sequences in FASTA format with reconstructed ancestral sequences
     - custom amino acid substitution matrix in PAML format or standard amino acid substitution matrix (JTT, WAG, LG etc.)
     - rooted cladogram in NEWICK format containing internal node labels corresponded to the reconstructed ancestral sequences
     - rooted phylogram in NEWICK format
     - shape parameter (alpha) for the gamma distribution of amino acid substitution rates
   - Output:
     - rooted cladogram in NEWICK format containing internal node labels with information about classes and positions of atypical substitutions (Fig. 2)

```
(3/1/PQ_488_||_Arabidopsis_lyrata_scaffold_503580.1_EnsemblPlants,5/1/SC_307_||_
KEGG_AT3G62980_Arabidopsis_thaliana_TIR1,(30/2/ED_7_161_244_550|GA_169_191_||_gb
_ABG46343.1_TIR1_Gossypium_hirsutum,((20/2/PR_561|TA_204_230_353_418_449_||_gb_A
CU81102.1_TIR1_Solanum_lycopersicum,18/2/PL_539|SC_297_||_gb_ACT53268.1_TIR1_Nic
otiana_tabacum)N4_||_31/3/ED_83_137_165_221_244_365_429|GA_87_204_348_523|IG_303
,((29/2/VD_418|YH_426_450_||_KEGG_RCOM_0556140_Ricinus_communis_TIR1,32/6/CR_397
_480|GR_229_517|HR_174_216|SR_53_293|TQ_18_337|YC_88_427_||_KEGG_POPTR_572746_Po
pulus_trichocarpa_FBL1)N6_||_8/5/FC_179|GA_191_275|PL_167|VS_324|YN_3,(26/6/CR_2
28|ED_8_166_244_248_345_404|FC_547|HD_142_174|TS_117_178_296_381_557|VA_95_191_2
54_273_349_||_gb_ACX31301.2_TIR1_Dimocarpus_longan,38/4/GR_55_567|PR_408_539|SM_
98|WM_102_||_KEGG_100233127_Vitis_vinifera_TIR1)N7)N5_||_3/2/GR_286|VL_245)N3_||
_10/3/KR_43_155_210_406|ME_365|YC_179)N2_||_32/5/SA_25_97_254_324_354_362_499|SC
_117_265|TN_150_225|VG_255_348|YP_300)N1;
```

Fig. 2. Cladogram in NEWICK format containing internal node labels with information about classes and positions of atypical substitutions.

6) Summarizing of amino acid substitutions (all or atypical only) from root to tips (Evolutionary changes analysis node):
    - Input:
        - gapless alignment of protein sequences in FASTA format with reconstructed ancestral sequences
        - rooted cladogram in NEWICK format containing internal node labels corresponded to the reconstructed ancestral sequences or rooted cladogram in NEWICK format containing internal node labels with information about classes and positions of atypical substitutions
    - Output:
        - tab-delimited text table containing sums of the physicochemical changes (based on calculation which takes into account atypical or all amino acid substitutions) from the root of the phylogenetic tree to its tips
7) Chronogram reconstruction (Fast chronogram building node):
    - Input:
        - aligned protein sequences in FASTA format
        - rooted phylogram in NEWICK format
        - species divergence dates (Fig. 3)
    - Output:
        - chronogram in NEWICK format

```
calibration numbers: 1 #format: branch0 branch1 max_divergence_date
min_divergence_date
Arabidopsis_lyrata_scaffold_503580.1_EnsemblPlants
KEGG_AT3G62980_Arabidopsis_thaliana_TIR1 20 10
```

Fig. 3. information about the species divergence dates

8) Evaluation of the statistical relationship between the amino acids physicochemical evolution and phenotypic characteristics of organisms (Phylogenetic comparative statistics node):
    - Input:
        - chronogram in NEWICK format

- tab-delimited text table containing sums of the physicochemical changes (based on calculation which takes into account atypical or all amino acid substitutions) from the root of the phylogenetic tree to its tips
- quantitative phenotypic characteristics of organisms

- Output:
- tab-delimited text table containing statistical estimates of the relationship between the amino acids physicochemical evolution and the phenotypic trait evolution (Fig. 4)

Amino acid index vs User data	log Likelihood	Spearman rho (additive values)	p-value (additive values)	Spearman rho (residual value)	p-value (residual value)
Hydrostatic pressure asymmetry index, PAI Di Giulio, 2005	16.30848	1	5.51E-06	-0.98333	4.96E-05
AA composition of CYT of single-spanning proteins Nakashima-Nishikawa, 1992	-4.78783	1	5.51E-06	-1	5.51E-06
Composition of amino acids in nuclear proteins percent Cedano et al., 1997	-3.96269	1	5.51E-06	-1	5.51E-06
Hydrophobic parameter Levitt, 1976	6.227831	1	5.51E-06	-0.98333	4.96E-05
AA composition of CYT2 of single-spanning proteins Nakashima-Nishikawa, 1992	-7.49489	-0.91667	0.001312	-1	5.51E-06
RF value in high salt chromatography Weber-Lacey, 1978	25.97131	0.95	0.000353	-1	5.51E-06
AA composition of CYT of multi-spanning proteins Nakashima-Nishikawa, 1992	-2.74516	1	5.51E-06	-1	5.51E-06
Normalized flexibility parameters B-values for each residue surrounded by none rigid neighbours Vihinen et al., 1994	29.61655	-0.98333	4.96E-05	-1	5.51E-06
Hydrophilicity value Hopp-Woods, 1981	6.169226	1	5.51E-06	-1	5.51E-06
AA composition of total proteins Nakashima et al., 1990	-0.92281	1	5.51E-06	-1	5.51E-06
Surface composition of amino acids in intracellular proteins of mesophiles percent Fukuchi-Nishikawa, 2001	-6.23451	1	5.51E-06	-0.98333	4.96E-05
Amino acid composition Dayhoff et al., 1978a	-3.44965	1	5.51E-06	-1	5.51E-06
Amino acid distribution Jukes et al., 1975	0.289727	1	5.51E-06	-0.98333	4.96E-05

Fig. 4. A fragment of table containing statistical estimates of the relationship between the physicochemical evolution and the phenotypic trait evolution

Analysis of protein-coding genes evolution

1) Translation of the protein-coding genes into proteins (Codons to Amino acids Translation node):
    - Input:
        - unaligned protein-coding gene sequences in FASTA format
    - Output:
        - unaligned protein sequences in FASTA format
2) Multiple protein alignment (Alignment node):
    - Input:
        - unaligned protein sequences in FASTA format
    - Output:
        - aligned protein sequences in FASTA format
3) Amino acid substitution matrix generating (Amino acid substitution model estimation node):
    - Input:
        - aligned protein sequences in FASTA format
    - Output:
        - custom amino acid substitution matrix (symmetric matrix containing relative rates of amino acid substitutions) in PAML format (Fig. 1)
4) Back translation of the protein alignment into codon alignment (Amino acids to Codons Translation node):
    - Input:
        - aligned protein sequences in FASTA format
        - unaligned protein-coding gene sequences in FASTA format
    - Output:
        - aligned codon sequences in FASTA format
5) Phylogram reconstruction (Build tree node):
    - Input:
        - aligned codon sequences in FASTA format
        - model of nucleotide substitutions (HKY, GTR, TN, etc)
        - (optionally) initial cladogram in NEWICK format
    - Output:
        - phylogram in NEWICK format (unrooted)
6) Ancestral sequence reconstruction (Ancestral reconstruction node):
    - Input:
        - gapless alignment of codon sequences in FASTA format
        - custom amino acid substitution matrix in PAML format or standard amino acid substitution matrix (JTT, WAG, LG, etc.) or model of codon evolution
        - rooted cladogram in NEWICK format
    - Output:
        - gapless alignment of codon sequences in FASTA format with reconstructed ancestral sequences

- rooted cladogram in NEWICK format containing internal node labels corresponded to the reconstructed ancestral sequences
- reconstruction probabilities of ancestral amino acids for each internal node of the tree represented in FASTA-like format

7) Estimation of the K_R/K_C rate (Kr/Kc estimation node):
   - Input:
     - gapless alignment of codon sequences in FASTA format with reconstructed ancestral sequences
     - number of groups of canonical amino acids (1-5) or BLOSUM matrix (Fig. 5)
     - rooted phylogram in NEWICK format
   - Output:
     - tab-delimited text tables containing K_R, K_C, K_R/K_C values for each tree branch
     - tab-delimited text tables containing sums of K_R, K_C, K_R/K_C values from the root of the phylogenetic tree to its tips

```
BLOSUM: (C) Copyright 1992, Fred Hutchinson Cancer Research Center
Re-clustering percentage = 70
10 times log base 10 matrix:
 A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
  8
-10  10
 -4  -3  12
 -1  -7  -6   9
 -5  -4 -23 -17  16
 -5  -2   1  -2 -87  12
 -3  -1  -7   3 -87   1   8
 -1 -15  -7  -3 -35  -3 -11  10
-13   0  -6 -13 -25   9  -3 -13   0
  0  -6  -3 -17 -20  -5  -7  -4  -6   6
 -6 -11  -2 -16 -36  -9  -3 -26  -2   1   9
 -1   4  -4  -8   4   1  -1  -9   2  -5  -6   8
 -6  -1  -8 -12 -29  -5  -8  -7   0   0   1  -1  13
 -7  -4  -7 -32 -14  -5 -19 -24   0   0  -2  -2  -9  13
 -7 -87 -11  -4 -87 -20 -11  -3   5  -4  -8 -27   1 -13  12
  3  -7  -1  -3 -10   1  -3   1 -19  -3  -8  -3  -5 -24  -7   7
  1  -6  -1  -8   2  -2  -4  -7  -3  -2  -9  -3  -4 -13 -26   4   9
-87 -87 -87 -87 -87 -87 -21 -87 -87 -87 -87 -87 -87 -87 -87 -87 -87  19
 -7  -1 -12  -7  10 -11 -14 -87  10  -4 -12   0  -8   6 -28 -11  -5 -87  12
 -2  -4  -3  -8  -6  -2  -5  -3  -8   4   0  -4   1  -1 -19  -1  -5 -45  -6   6
 A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
```

Fig. 5. BLOSUM matrix.

8) Chronogram reconstruction (Fast chronogram building node):
   - Input:
     - aligned codon sequences in FASTA format
     - rooted phylogram in NEWICK format
     - species divergence dates (Fig. 3)
   - Output:
     - chronogram in NEWICK format

9) Evaluation of the statistical relationship between the gene evolution and phenotypic characteristics of organisms (Phylogenetic comparative statistics node):
    - Input:
        - chronogram in NEWICK format
        - tab-delimited text tables containing sums of K_R, K_C, K_R/K_C values from the root of the phylogenetic tree to its tips
        - quantitative phenotypic characteristics of organisms
    - Output:
        - tab-delimited text table containing statistical estimates of the relationship between the gene evolution and the phenotypic trait evolution (Fig. 4)
10) Estimation of the K_R/K_C rate under neutral evolution conditions (Neutral Kr/Kc estimation node):
    - Input:
        - gapless alignment of codon sequences in FASTA format with reconstructed ancestral sequences
        - number of groups of canonical amino acids or BLOSUM matrix
        - rooted phylogram in NEWICK format
        - number of simulations
    - Output:
        - tab-delimited text tables containing means and std. deviations for K_R, K_C, K_R/K_C values for each tree branch