T. Koestler, A. Von Haeseler, I. Ebersbergen

REvolver: Modeling Sequence Evolution under Domain Constraints

Mol. Biol. Evol. 29(9): 2133-2145. (2012)

Introduction

- Simulation the evolution of biological sequences
  - Reduce complexity vs. Biological reality
- Seq-Gen, ROSE (indels)
- INDELible, SIMPROT, indel-Seq-Gen (manual assignment of evolutionary parameters)

Introduction

- Problems:
  - – No automatized procedure to extract meaningful constraints
  - – No standard operating procedure for inferring evolutionary constraints
  - – Structures not available
  - – Indel lengths from a single distribution

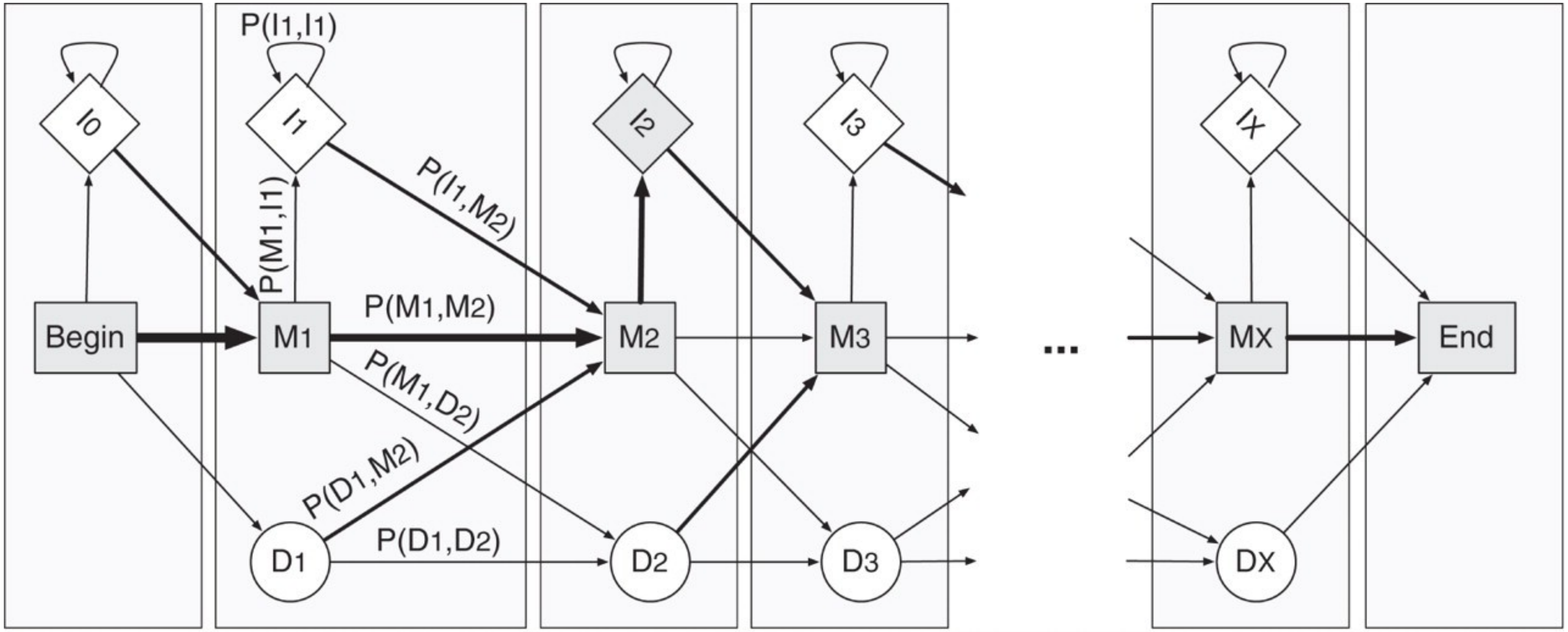
A New Approach

- Comparing homologous sequences
  - Sites that remain entirely conserved over time
  - Sites displaying only a subset of the amino acid alphabet
  - Sites that appear to be free to change
- Footprint of a constrained evolutionary process
- Profile Hidden Markov Model (pHMM)

REvolver

- Emission probabilities as site-specific AA frequnces
- Indels preferrably placed at positions where they have been observed in real instances
- No formation of repeated nested insertions
- Information about site-specific evolutionary constraints maintained throughout the simulation
- Prevents a simulated sequence from losing its identity as a domain instance

Structure of a pHMM: The pHMM comprises match states (Mx), insertion states (Ix), deletion states (Dx), a Begin state, and an End state.



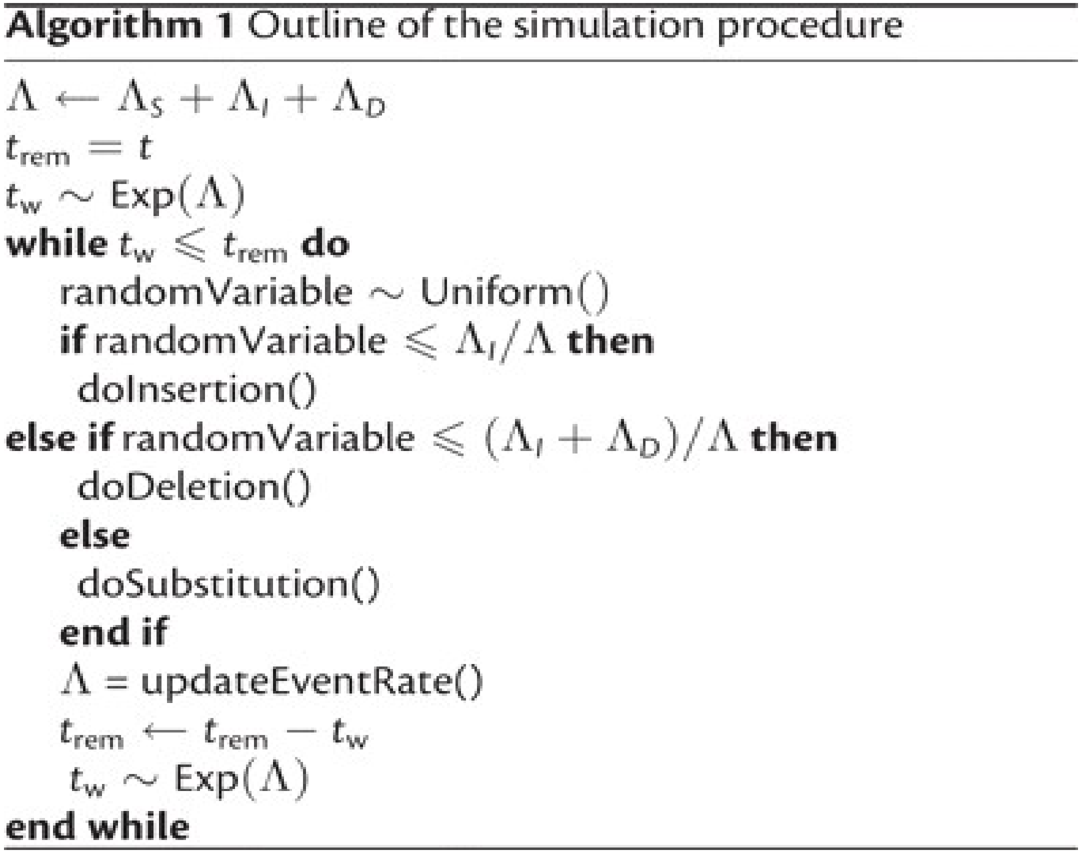
AA seq	R_1	Q_2 V_3 E_4	G_5	...	G_L
state path	M_1	M_2 I_2 I_2	M_3	...	M_X

Koestler T et al. Mol Biol Evol 2012;29:2133-2145

MBE MOLECULAR BIOLOGY AND EVOLUTION

The Simulator

Gillespie algorithm (1977)



Algorithm 1 Outline of the simulation procedure

```
$\Lambda \leftarrow \Lambda_S + \Lambda_I + \Lambda_D$
$t_{rem} = t$
$t_w \sim \text{Exp}(\Lambda)$
while $t_w \leqslant t_{rem}$ do
   randomVariable $\sim$ Uniform()
   if randomVariable $\leqslant \Lambda_I / \Lambda$ then
     doInsertion()
else if randomVariable $\leqslant (\Lambda_I + \Lambda_D)/\Lambda$ then
     doDeletion()
   else
     doSubstitution()
   end if
   $\Lambda$ = updateEventRate()
   $t_{rem} \leftarrow t_{rem} - t_w$
    $t_w \sim \text{Exp}(\Lambda)$
end while
```

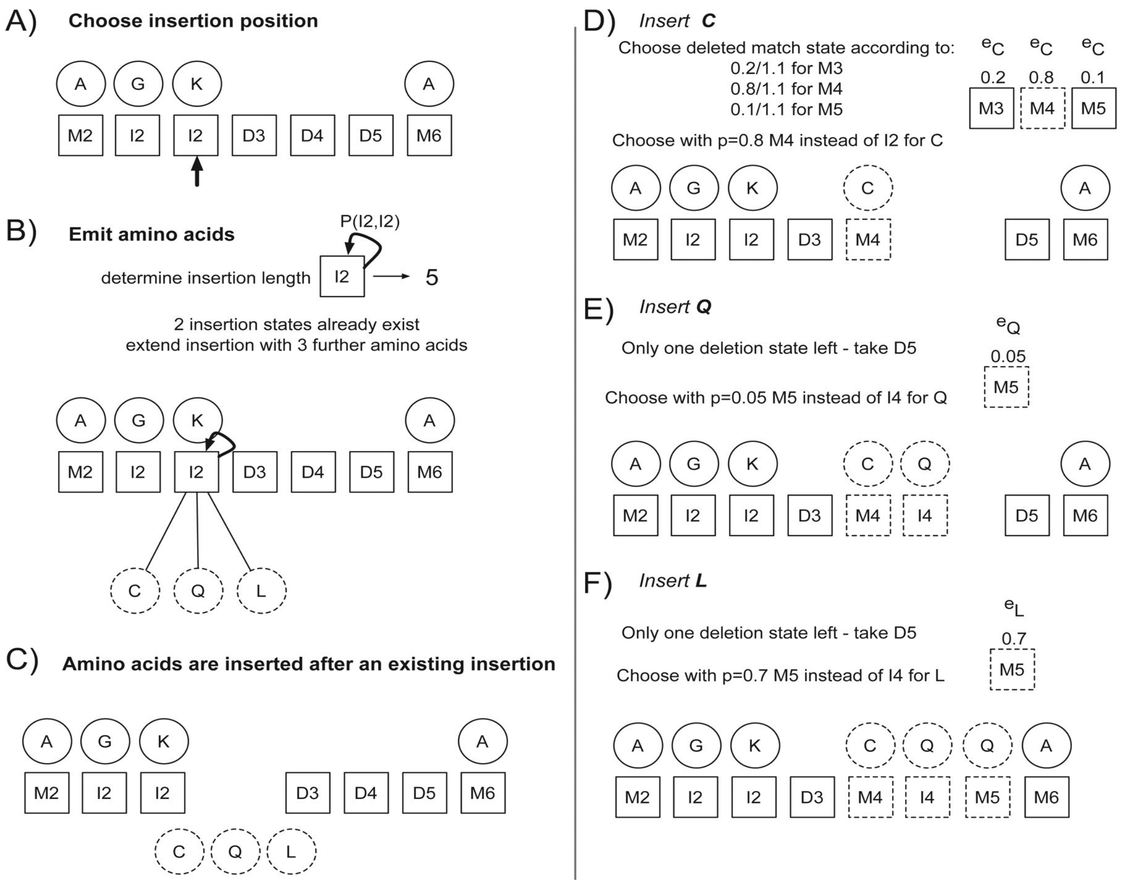
Unconstrained segments

- **Substitutions**
  - Substitution model Q
  - Scaling factor
    - Same at all sites
    - Continuous gamma distribution
    - Discrete gamma distribution
- **Insertions and Deletions**
  - **Position –** uniform distribution
  - Length - Geometric distribution
  - Length - Zipfian distribution

Constrained segments

- Substitutions
  - Each site in the domain gets assigned its own model Q
- Insertions
  - Length: geometric distribution (1-p)
  - Nested insertions
- Deletions
  - No explicit deletion length
- Resurrection of M states

A generic insertion scenario: circles represent the amino acid sequence, the corresponding state path is shown as squares.



A) Choose insertion position

B) Emit amino acids

determine insertion length

2 insertion states already exist
extend insertion with 3 further amino acids

C) Amino acids are inserted after an existing insertion

D) *Insert **C***

Choose deleted match state according to:
0.2/1.1 for M3
0.8/1.1 for M4
0.1/1.1 for M5

Choose with p=0.8 M4 instead of I2 for C

E) *Insert **Q***

Only one deletion state left - take D5

Choose with p=0.05 M5 instead of I4 for Q

F) *Insert **L***

Only one deletion state left - take D5

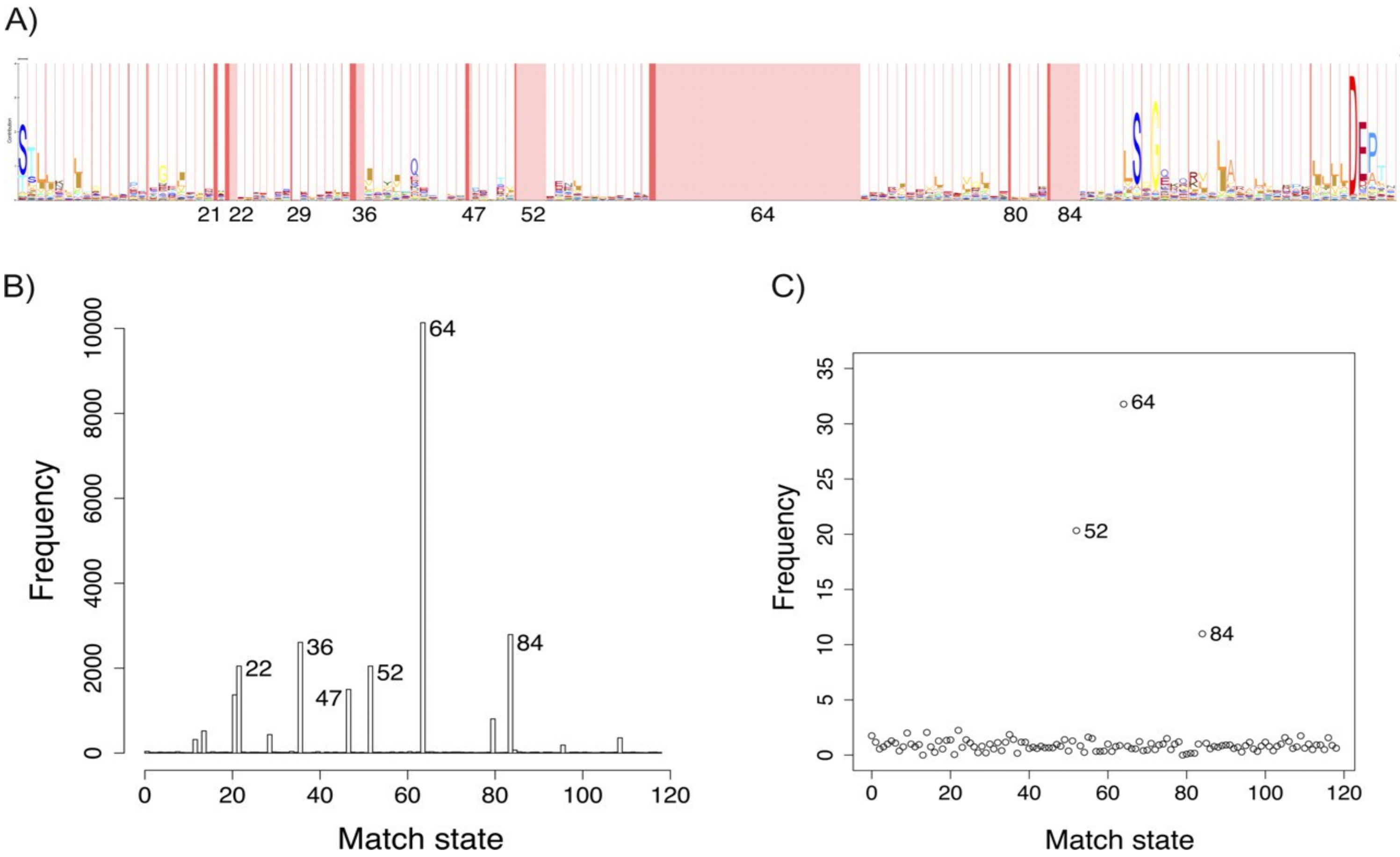
Choose with p=0.7 M5 instead of I4 for L

Koestler T et al. Mol Biol Evol 2012;29:2133-2145

Additional Features

- Input – phylogenetic tree, a root sequence
- Output – multiple alignment of simulated leaf node sequences
- Lineage-specific evolution
- Running time
- www.cibiv.at/software/revolver
  - Requires Java6 and HMMER3 software package
  - Pfam or SMART
- Verification

Positions and lengths of insertions in the ABC_tran domain.



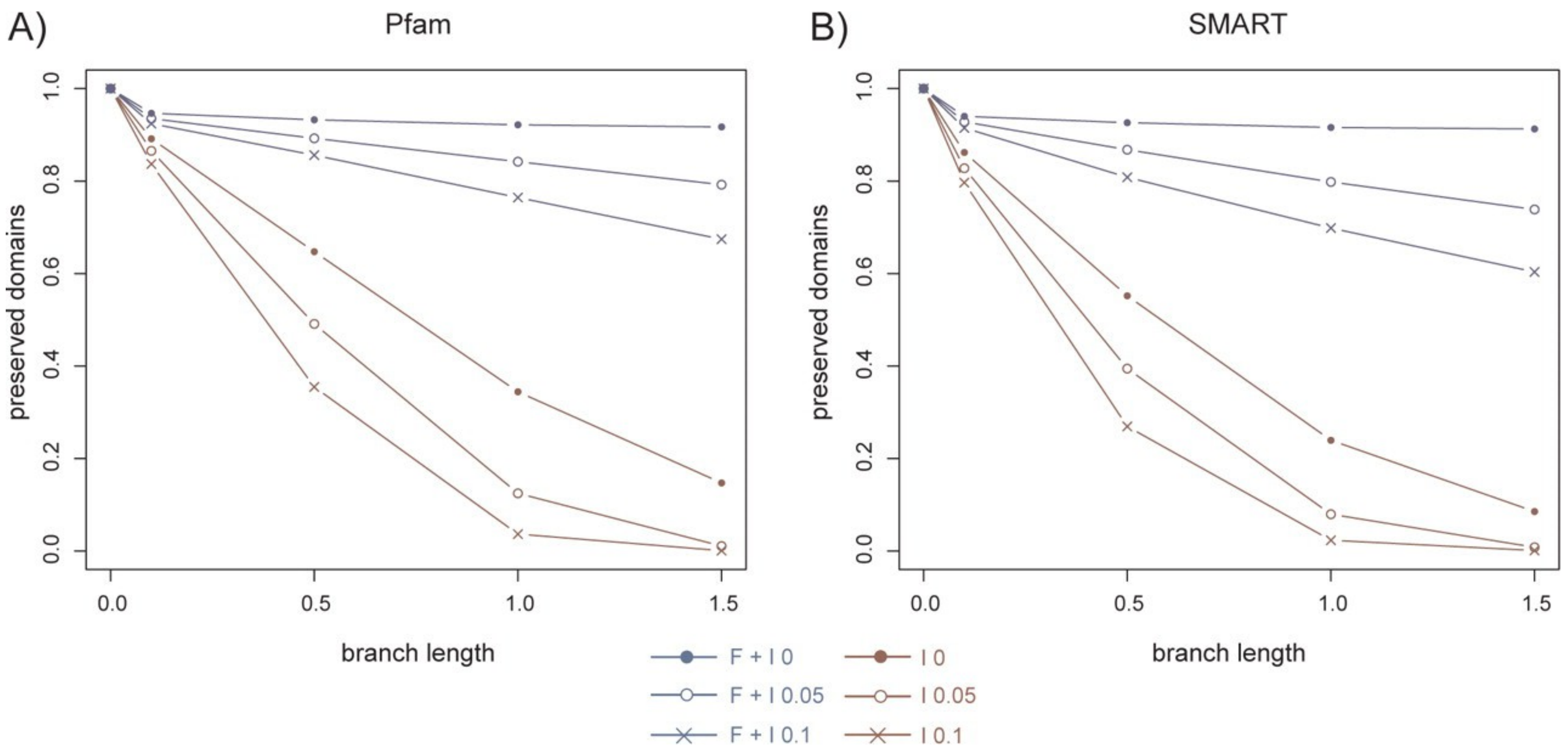
Koestler T et al. Mol Biol Evol 2012;29:2133-2145

Benchmarking and Example Applications

- Simulated evolution of G protein-coupled receptors (GPCR)

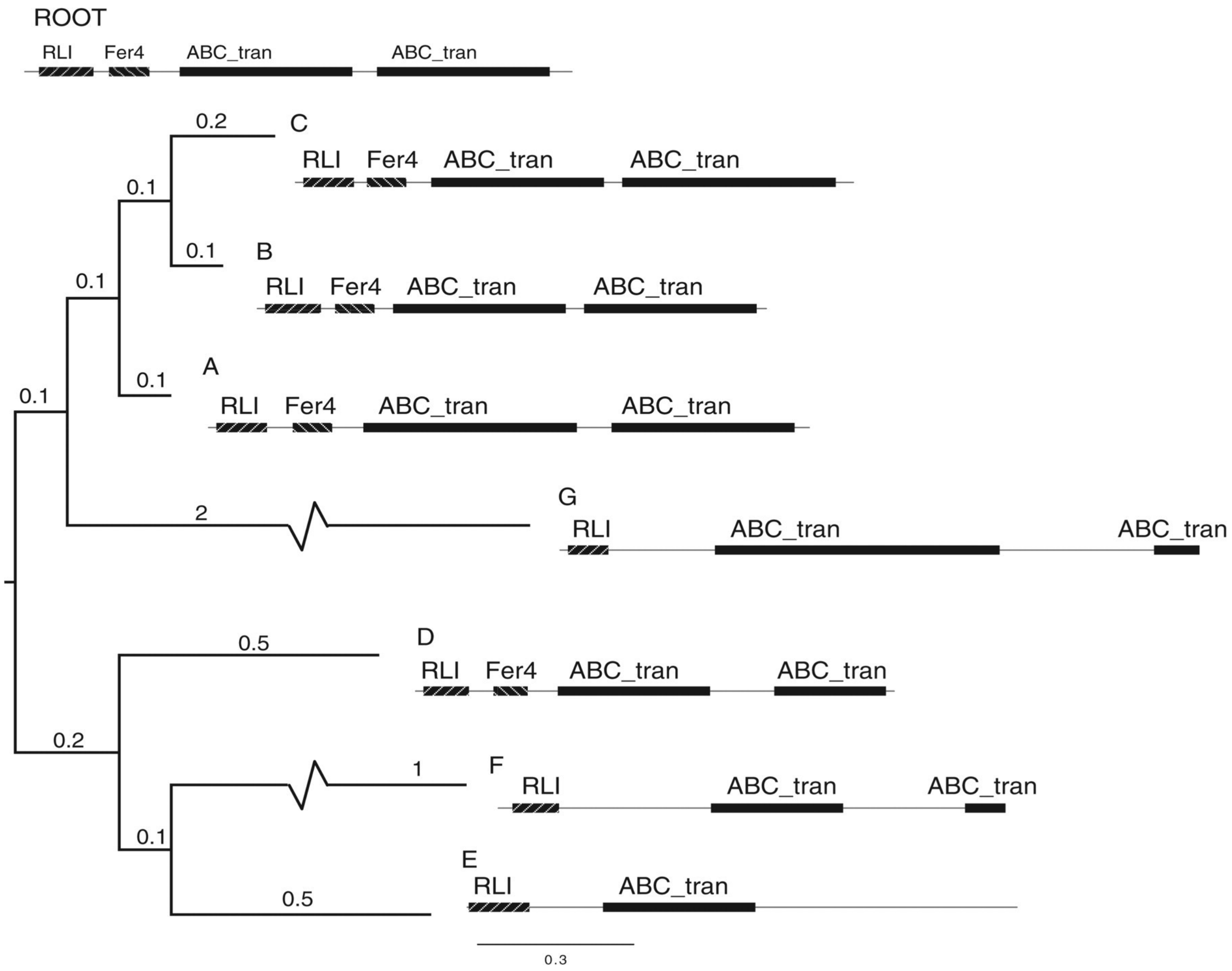
	Revolver	iSG	ROSE	SIMPROT	Seq-Gen
tm regions	6.89±0.60	7.03±0.30	5.94±1.25	0.20±0.37	6.84±0.91
Pfam bit score	102.75	−5.09	−31.47	–	−7.18
Top n BlastP hits					
25	152.0	174.0	141.1	–	196.7
100	143.6	164.7	132.7	–	183.3
250	135.5	155.9	124.4	–	177.8

Fraction of preserved Pfam (A) and SMART (B) domains.



Koestler T et al. Mol Biol Evol 2012;29:2133-2145

Domain architectures of sequences evolved with REvolver.



Koestler T et al. Mol Biol Evol 2012;29:2133-2145

Discussion

- The maintenance of protein domains in the course of evolution
- The large-scale applicability due to the automatic inference of sequence-specific evolutionary constrants