UNIVERSIDADE DA CORUÑA

CESGA

Modeltest-ng: parallelized selection of evolutionary models for DNA and proteins

Pedro Sánchez García

Master in Bioinformatics for Health Sciences

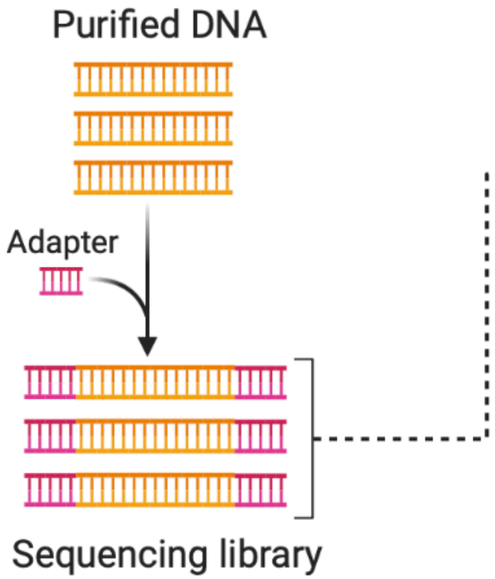
CONTENTS

INTRODUCTION

INTRODUCTION

Workflow

① **DNA purification and sequencing library preparation**

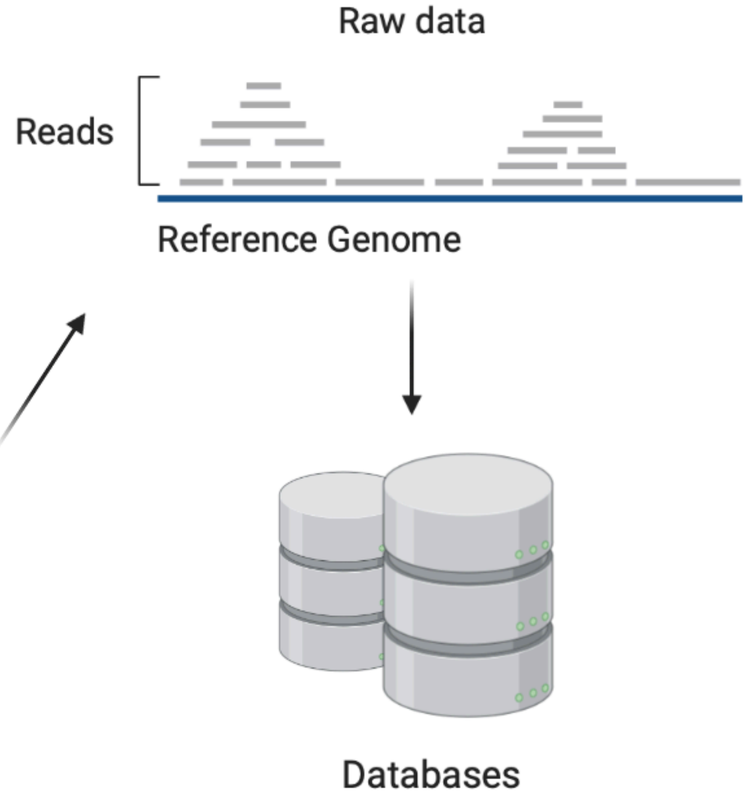


② **NGS sequencing and map assembly**

Reads:

....GTTCCTTCAGCATTTGCAGGTACCGT....

③ **Sequence analysis and motif identification**

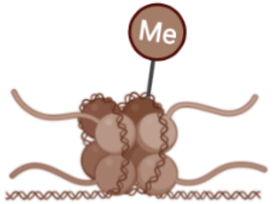


INTRODUCTION

Next Generation Sequencing Applications

Epigenetics

DNA methylation analysis



- Tissue identification
- Twins differentiation
- Prediction of age
- Drug consumption

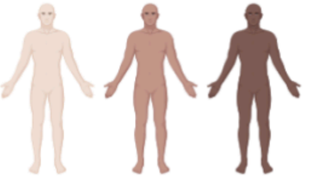
MicroRNA

RNA analysis

- Tissue identification
- Post-mortem interval
- Prediction of stain deposition time

Phenotype

SNPs analysis



- Skin, hair, iris color
- Baldness
- Age

Ancestry

SNPs analysis

- Ethnic difference
- Biogeographic origin
- Population analysis

Non-human identification

STRs/SNPs analysis

- Microbiome
- Animal/plant species analysis

Human identification

STRs/SNPs/MIIs analysis

- Identity
- Kindship
- Mix analysis

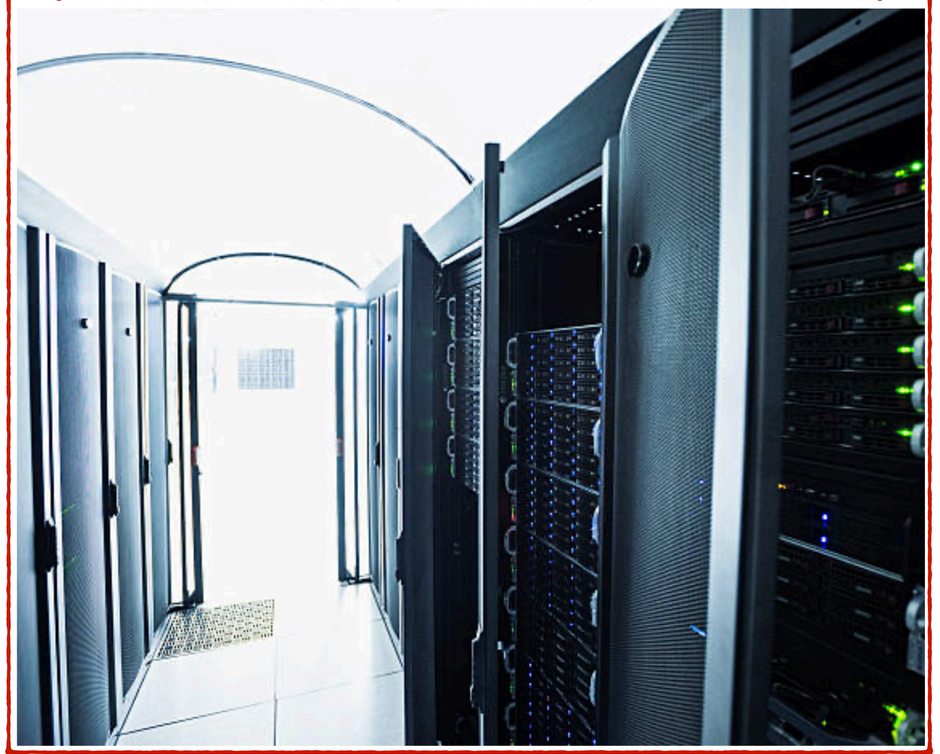
Predictiveness → **Accuracy**

INTRODUCTION

Context

Amount of data

High Performance Computing



✓ Support and performance.

INTRODUCTION

modeltest[ng]

- **Resolution of the best evolutionary model for multiple sequence alignments.**
- **Key in the first phase of phylogeny research.**

jModelTest 2

+

ProtTest 3

→

- ✓ Better performance.
- ✓ Flexibility in parameters for each type of project: *algorithms, model rates and topology.*
- ✓ *It solves problems with divergent and long sequences.*

INTRODUCTION

modeltest[ng]

Command Line Arguments

`-d`	`--datatype`	**nt,aa**	Data type is 'nt' for nucleotide (default), 'aa' for amino-acid sequences.
`-i`	`--input`	*filename*	Input MSA file in FASTA or sequential PHYLIP format. Check section 5.1
`-t`	`--topology`	*topology_type.*	Check section 5.3
		ml	maximum likelihood
		mp	maximum parsimony (default)
		fixed-ml-jc	fixed maximum likelihood (JC)
		fixed-ml-gtr	fixed maximum likelihood (GTR)
		random	random generated tree
		user	fixed user defined (requires -u argument)
`-u`	`--utree`	*filename*	User-defined tree in NEWICK format. Check section 5.3
`-q`	`--partitions`	*filename*	Partitions filename in RAxML format. Check section 5.4
`-o`	`--output`	*filename*	Pipes the output into a file
`-p`	`--processes`	*number_of_threads*	Number of concurrent threads
`-r`	`--rngseed`	*seed*	Sets the seed for the random number generator

INTRODUCTION

modeltest[ng]

Command Line Arguments: *Candidate Models*

-a	--asc-bias	algorithm[:values]	Includes ascertainment bias correction. Check section 5.5 for more details **lewis**: Lewis (2001) **felsenstein**: Felsenstein (requires number of invariant sites) **stamatakis**: Leach et al. (2015) (requires invariant sites composition)
-f	--frequencies	[ef]	Sets the candidate models frequencies **e**: Estimated - maximum likelihood (DNA) / empirical (AA) **f**: Fixed - equal (DNA) / model defined (AA)
-h	--model-het	[uigf]	Sets the candidate models rate heterogeneity **u**: Uniform **i**: Proportion of invariant sites (+I) **g**: Discrite Gamma rate categories (+G) **f**: Both +I and +G (+I+G)
-m	--models	*list* dna: protein:	Sets the candidate model matrices separated by commas JC HKY TrN TPM1 TPM2 TPM3 TIM1 TIM2 TIM3 TVM GTR DAYHOFF LG DCMUT JTT MTREV WAG RTREV CPREV VT BLOSUM62 MTMAM MTART MTZOA PMB HIVB HIVW JTTDCMUT FLU STMTREV
-s	--schemes	*number_of_schemes*	Number of DNA substitution schemes. **3**: JC, HKY, GTR **5**: JC, HKY, TrN, TPM1, GTR **7**: JC, HKY, TrN, TPM1, TIM1, TVM, GTR **11**: All models defined in Sec 5.2 **203**: All possible GTR submatrices
-T	--template	*tool* **raxml** **phyml** **mrbayes** **paup**	Sets candidate models according to a specified tool RAxML (DNA 3 schemes / AA full search) PhyML (DNA full search / 14 AA matrices) MrBayes (DNA 3 schemes / 8 AA matrices) PAUP* (DNA full search / AA full search)

INTRODUCTION

modeltest[ng]

Algorithm

1. Analysis with an initial set of models in the sequences.
2. Division of models between the specified threads.
3. Evaluation of distances, branch lengths and matrix scores for each model.
4. Continuous statistical comparisons and generation of a final sorted table.

METHODS

METHODS

GitHub

modeltest[ng]

Multithreaded version (Pthreads)

Shared memory systems

- modeltest-ng requires GNU Bison and Flex dependencies.
- Compilation with CMake 3.12.4.
- Level of parallelism:

Coarse-grained type

- 1.) An initial set of models generates optimized parameters.
- 2.) Subsequent parallel evaluation: Less synchronization.

METHODS

Performance evaluation

- **Standard selection** of best evolutionary models **between 88 posibles**.
- **Default parameters**: standard amino acid frequencies and maximum likelihood.
- **1 node** in Finis Terrae II (two 12-core Intel Xeon Haswell E5-2680V3 processors and 128 GB memory).
- Estimation of best models and reconstruction of phylogenetic tree:

1 | 2 | 4 | 8 | 12 | 16 | 24 threads

METHODS

Performance evaluation

```
#!/bin/bash
#SBATCH -N 1
#SBATCH -n 1
#SBATCH -c 24
#SBATCH -t 10:00:00
#SBATCH --error=/home/ulc/cursos/curso314/Trabajo_tutelado/env/errorenv24.txt
#SBATCH --output=/home/ulc/cursos/curso314/Trabajo_tutelado/env/outputenv24.txt

./modeltest-ng -i HIV_env.fas -d nt -p 24
```

METHODS

Performance evaluation: *Output*

- **BIC**: *Bayesian Criterion Information.*
- **K**: *Number of estimated parameters.*
- ***LnL***: *Negative log likelihod.*
- ***Delta and weight***: *dummy variables.*

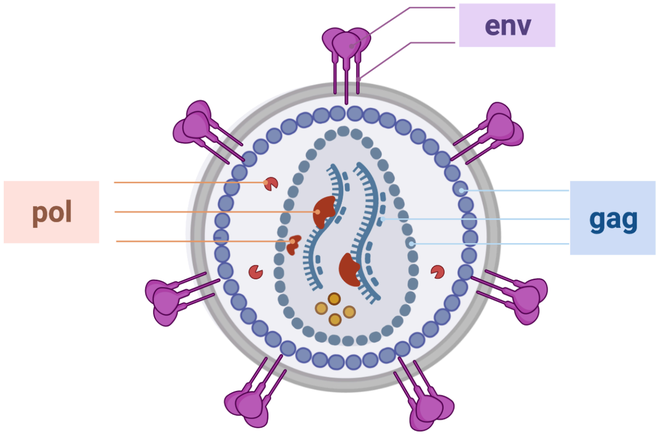
```
                                    _      _ _                 _      _   _ _____
                                   | |    | | |               | |    | \ | |/ ____|
         _ __ ___   ___   __| | ___| | |_ ___   ___| |_   |  \| | |   __
        | '_ ` _ \ / _ \ / _` |/ _ \ | __/ _ \ / __| __|  | . ` | | |_ |
        | | | | | | (_) | (_| |  __/ | ||  __/\__ \ |_   | |\  | |__| |
        |_| |_| |_|\___/ \__,_|\___|_|\__\___||___/\__|  |_| \_|\_____|
-------------------------------------------------------------------------------

Computation of likelihood scores completed. It took 0h:15:18

BIC       model              K           lnL          score          delta    weight
-------------------------------------------------------------------------------
      1   GTR+I+G4          10  -1654780.8576   3376874.4142         0.0000    1.0000
      2   TVM+I+G4           9  -1654848.4232   3377001.3932       126.9790    0.0000
      3   TPM1uf+I+G4        7  -1655178.4224   3377645.0873       770.6731    0.0000
      4   GTR+G4             9  -1655922.5333   3379149.6135      2275.1993    0.0000
      5   TPM3uf+I+G4        7  -1656048.7646   3379385.7716      2511.3574    0.0000
      6   TPM1uf+G4          6  -1656119.4196   3379518.9294      2644.5152    0.0000
      7   TVM+G4             8  -1656237.8767   3379772.1481      2897.7338    0.0000
      8   TPM2uf+I+G4        7  -1656314.7707   3379917.7838      3043.3696    0.0000
      9   HKY+I+G4           6  -1656523.9259   3380327.9420      3453.5277    0.0000
     10   TPM3uf+G4          6  -1656944.2320   3381168.5542      4294.1400    0.0000
-------------------------------------------------------------------------------
Best model according to BIC
---------------------------
Model:              GTR+I+G4
lnL:                -1654780.8576
Frequencies:        0.3999 0.1995 0.1822 0.2184
Subst. Rates:       1.5033 8.2182 0.8111 0.8099 8.8528 1.0000
Inv. sites prop:    0.0420
Gamma shape:        0.5944
Score:              3376874.4142
Weight:             1.0000
```

RESULTS

RESULTS

Datasets

HIV-1 virion



Multiple alignments of HIV-1 proteins

FASTA PHYLIP

Name	Number of sequences	Length
Pol	4125	3540
Gag	6574	2126
Env	8075	4130

RESULTS

Execution times (hh:mm:ss)

Threads	Pol	Gag	Env
1	04:00:58	01:46:29	03:32:16
2	**01:47:25**	01:00:21	02:09:18
4	00:57:45	00:35:24	01:14:33
8	00:32:38	00:22:05	00:47:20
12	00:26:08	00:19:13	00:41:49
16	00:17:23	00:17:34	00:36:40
24	00:15:18	00:12:07	**00:32:50**

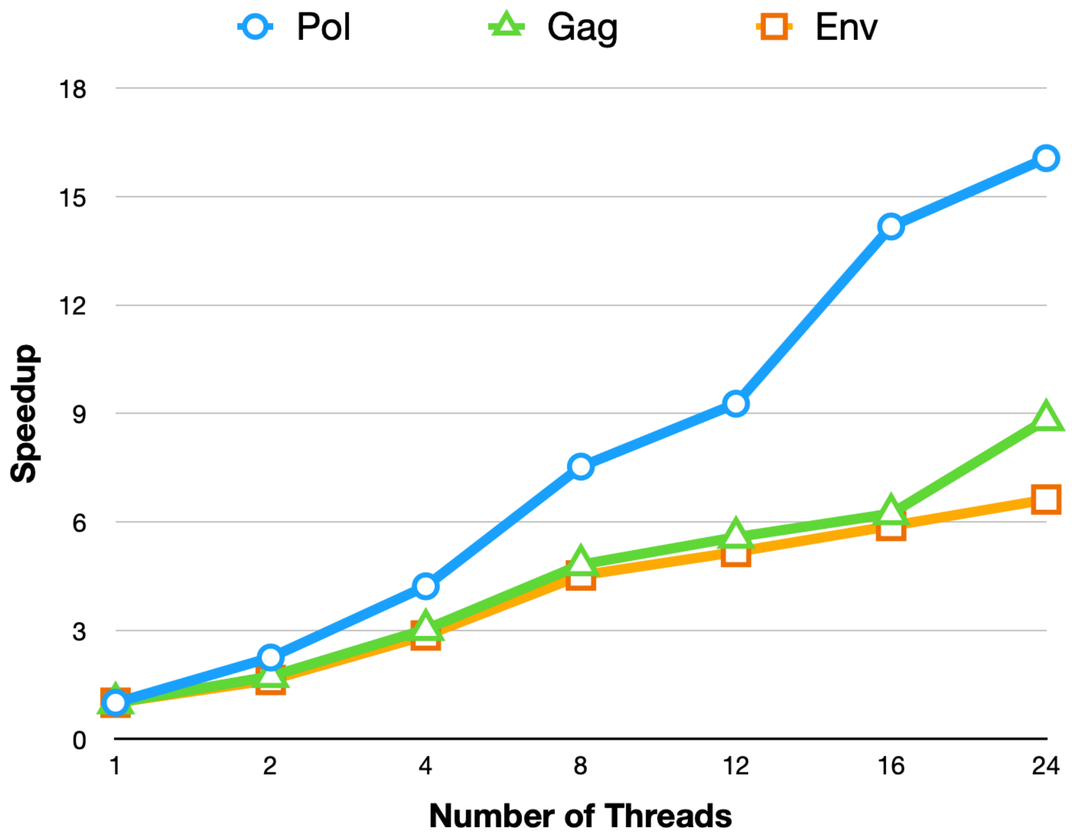
- ***Pol***: Drastic reduction in execution time with 2 threads.
- ***Gag***: Less remarkable reduction.
- **Env**: Similarity with Pol in reduction.

▽

Length of sequences and gaps

RESULTS

Speedup



$$Speedup = \frac{T_{seq}}{T_n}$$

- ***Scalability***: Linear with up to 12 threads.
- ***Gag and Env datasets***: modeltest-ng scaled well with 16 threads.
- ***Pol dataset***: similarity with 24 threads.

Number of threads

Loss of optimized workload distribution

CONCLUSIONS

CONCLUSIONS

1.) MODELTEST-NG IS 100 TIMES FASTER THAN JMODELTEST.

2.) SUPPORT TO SPEED UP RESEARCH IN PHYLOGENY.

Mitochondrial genome sequencing of marine leukaemias reveals cancer contagion between clam species in the Seas of Southern Europe

Daniel Garcia-Souto[1,2,3*†], Alicia L Bruzos[1,2†], Seila Diaz[1†], Sara Rocha[4], Ana Pequeño-Valtierra[1], Camila F Roman-Lewis[5], Juana Alonso[5,6], Rosana Rodriguez[7], Damian Costas[7], Jorge Rodriguez-Castro[1], Antonio Villanueva[7], Luis Silva[8], Jose Maria Valencia[9,10], Giovanni Annona[11], Andrea Tarallo[11], Fernando Ricardo[12], Ana Bratoš Cetinić[13], David Posada[5,6,14], Juan Jose Pasantes[14,15], Jose MC Tubio[1,2*]

" CITED BY 479

THANKS FOR YOUR ATTENTION!

UNIVERSIDADE DA CORUÑA

CESGA

Modeltest-ng: parallelized selection of evolutionary models for DNA and proteins

Pedro Sánchez García

Master in Bioinformatics for Health Sciences