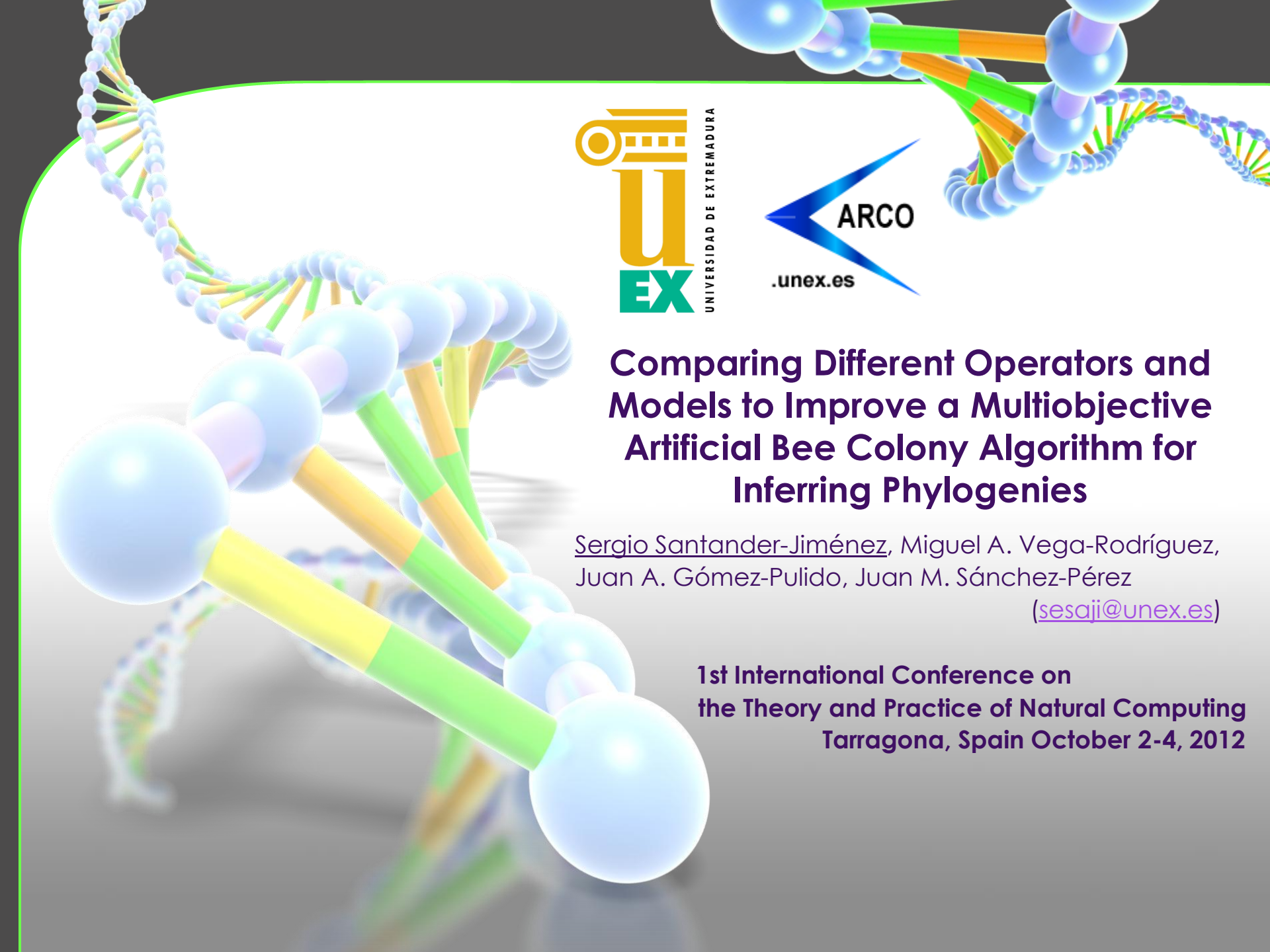


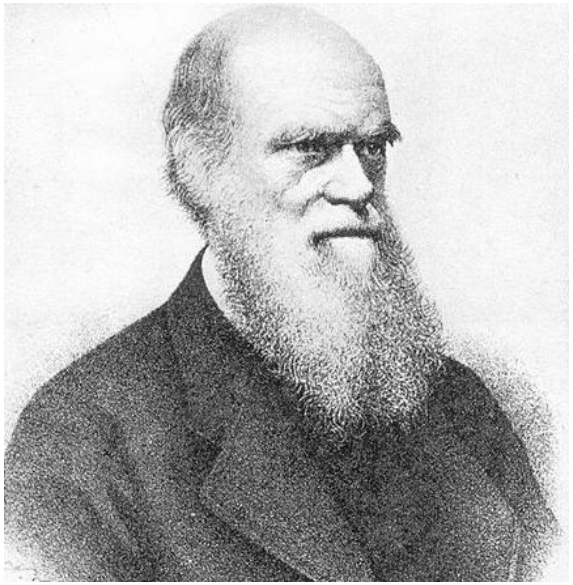
Comparing Different Operators and Models to Improve a Multiobjective Artificial Bee Colony Algorithm for Inferring Phylogenies

Sergio Santander-Jiménez, Miguel A. Vega-Rodríguez, Juan A. Gómez-Pulido, Juan M. Sánchez-Pérez
(sesaji@unex.es)

1st International Conference on the Theory and Practice of Natural Computing
Tarragona, Spain October 2-4, 2012

Contents

- In this presentation we will see:
  - An **introduction** to Phylogenetic Inference.
  - Our proposal: a **Multiobjective Artificial Bee Colony Algorithm** for Inferring Phylogenies according to two criteria:
    - Maximum Parsimony.
    - Maximum Likelihood.
  - Strategies for improving the algorithm, by considering:
    - Topological operators.
    - Evolutionary models.
  - **Experimental results**.
  - **Conclusions** and **future research work**.



On seeing the marsupials in Australia for the first time and comparing them to placental mammals: "An unbeliever might exclaim 'Surely two distinct Creators must have been at work'

~ Charles Darwin

INTRODUCTION TO PHYLOGENETIC INFERENCE

Basis of Phylogenetics

- Phylogenetic Inference encloses a wide range of estimation techniques that aim to describe ancestral evolutionary relationships among related species.

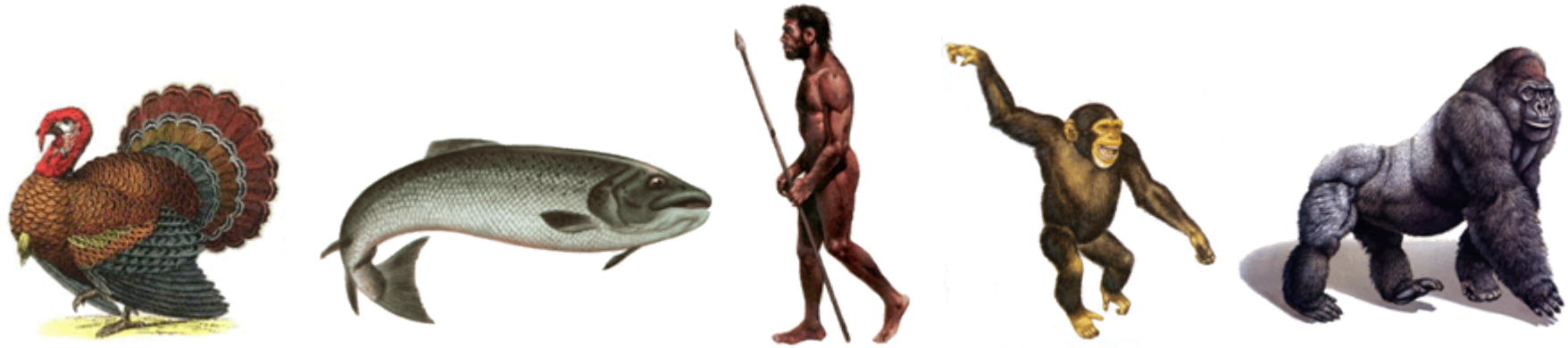
  - **Input**: a set of *n sequences of M characters (*sites), which represent molecular characteristics of the organisms.

  - **Output**: a mathematical structure that defines relationships among species by inferring hypothetical ancestors over the course of evolutionary history: *phylogenetic tree*.

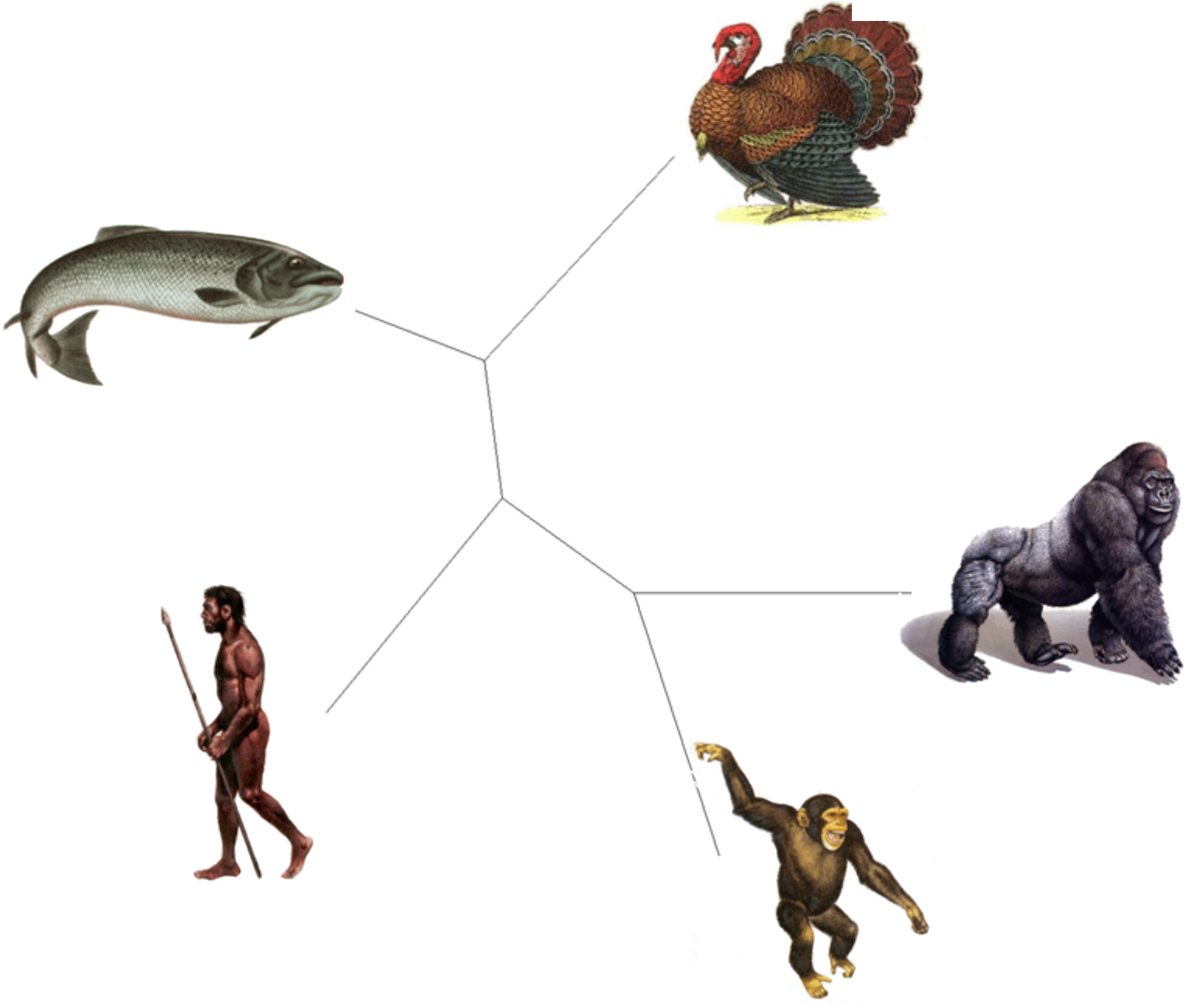
An example

5 species 42 nucleotides (DNA-based analysis)

AAGCTNGGGCATTTCAGGGTGAGCCCGGGCAATACAGGGTAT
AAGCCTTGGCAGTGCAGGGTGAGCCGTGGCCGGGCACGGTAT
ACCGGTTGGCCGTTCAGGGTACAGGTTGGCCGTTCAGGGTAA
AAACCCTTGCCGTTACGCTTAAACCGAGGCCGGGACACTCAT
AAACCCTTGCCGGTACGCTTAAACCATTGCCGGTACGCTTAA

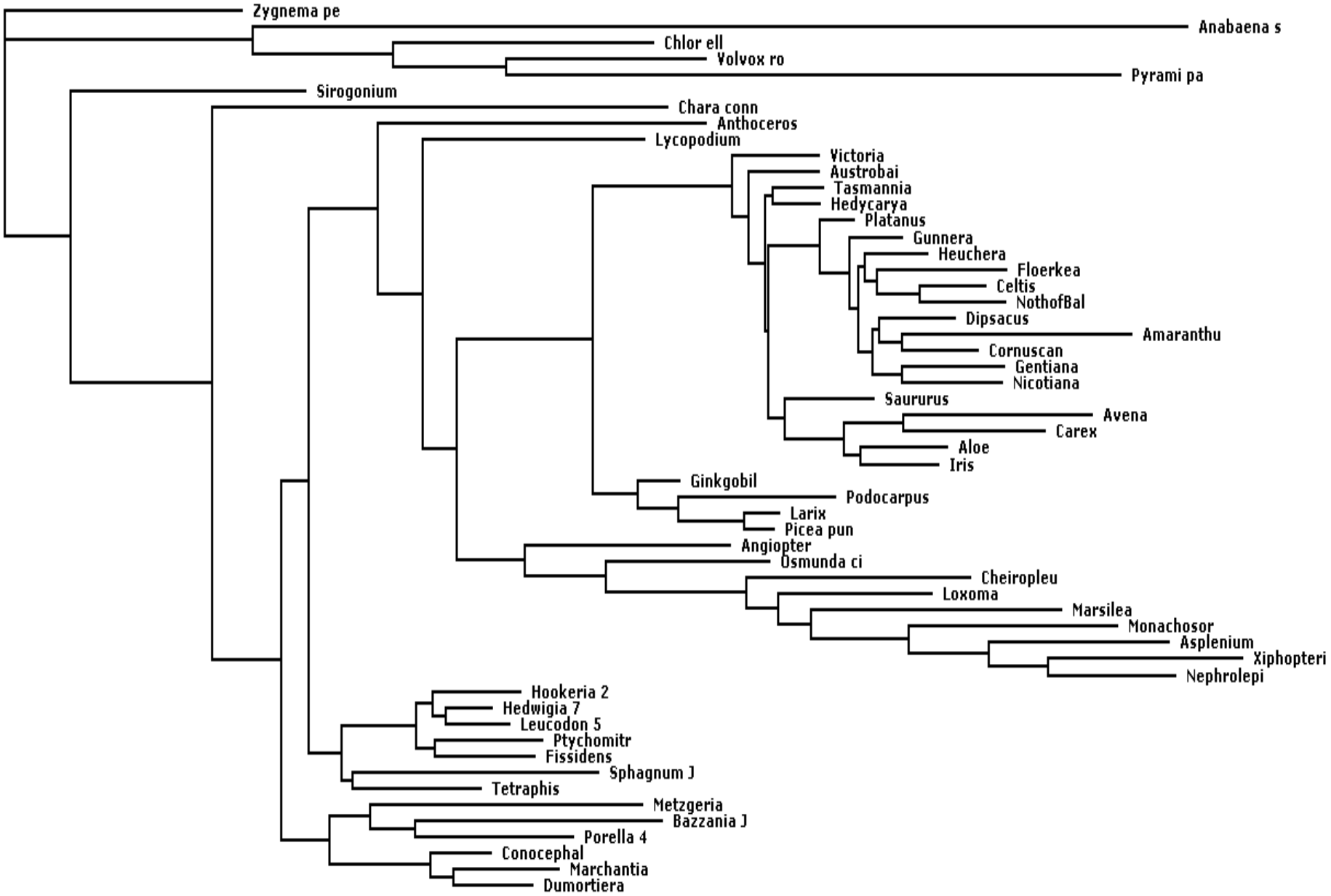


An example



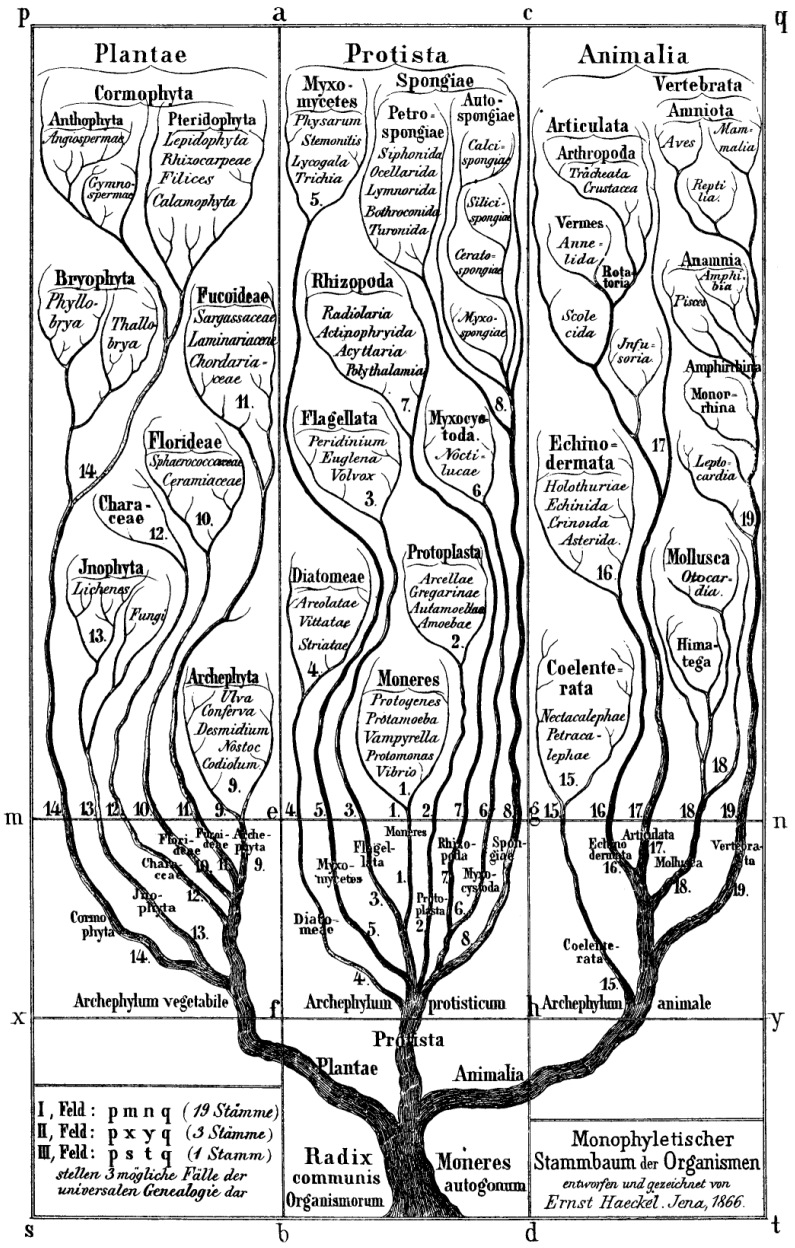
An example

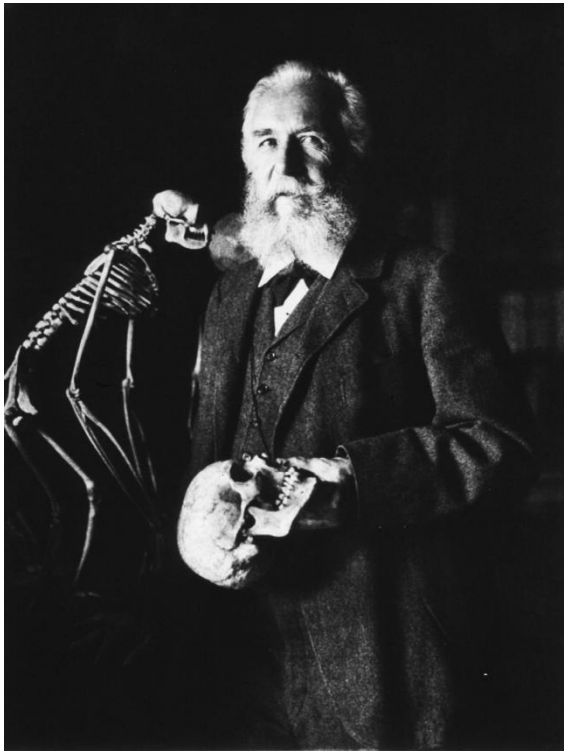
TPNC 2012



An example

TPNC 2012





Generelle Morphologie der Organismen.

Ernst Haeckel

Complexity of the problem

- Optimality criteria methods for phylogenetic reconstruction were proposed with the aim of inferring optimal phylogenetic trees according to a specific criterion.
- When dealing with large data sets, **exhaustive methods cannot be applied** due to two key factors:

$$\frac{(2n-5)!}{(n-3)!2^{n-3}}$$

1. By increasing the **number of species** in the input dataset, the search space of possible phylogenetic topologies **grows exponentially**.
2. By increasing the **number of sites** in molecular sequences, the assessment of phylogenetic trees will require **more processing times and memory consumption**.

We can tackle these issues by using Nature-inspired Computing.

Related works

- First evolutionary approaches proposed for Phylogenetic Inference:
  - Matsuda, 1995.
  - Lewis, 1998.
- Modern studies suggested different mutation operators as well as new crossover strategies to carry out topological searches.
  - Goloboff, 1999.
  - Cotta and Moscato, 2002.
- The success of model-based phylogenetic procedures depends on selecting the most accurate evolutionary models according to statistical metrics.
  - Survey by Bos and Posada, 2005.

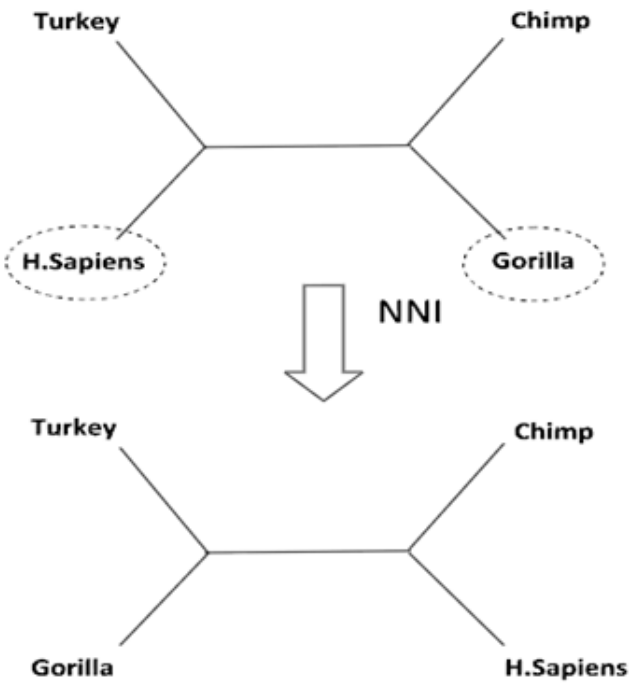
Optimality criteria

- Several criteria-based methods can be found in the literature.
- In this work we consider two of the most popular optimality criteria in Phylogenetics:
  - **Maximum parsimony** (Fitch, 1971)
    - Parsimony approaches aim to find those phylogenies that **minimize** the amount of molecular changes needed to explain the observed data.
    - Ockham's razor principle.
  - **Maximum likelihood** (Felsenstein, 1984)
    - Reconstruction of that phylogenetic tree which represents **the most likely** evolutionary history of the species.
    - Likelihood is highly related to the probabilities of mutation events, given by evolutionary models.

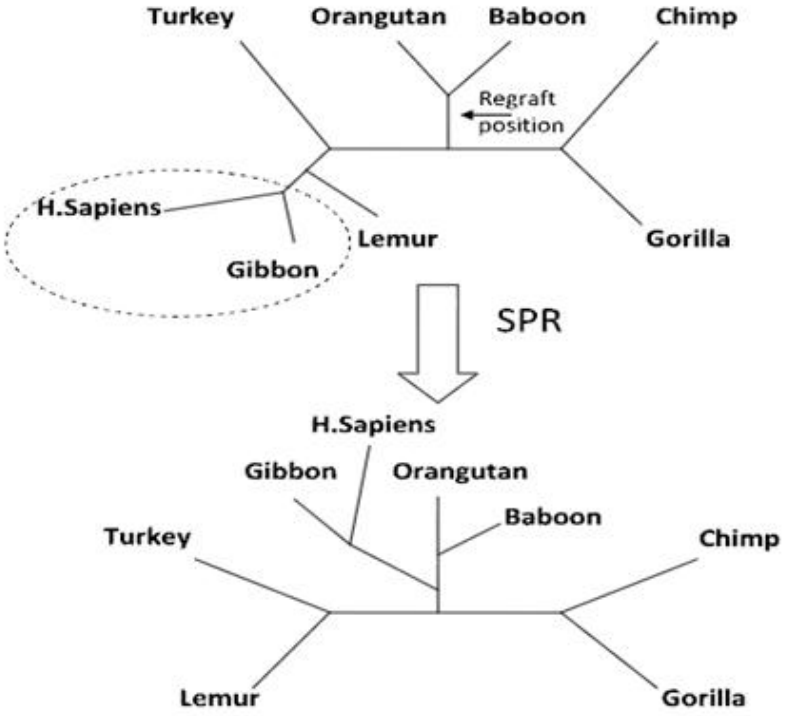
Topological operators

- In order to explore the search space for optimal phylogenetic trees, several topological operators have been proposed in literature:
  - ***Approaches based on local moves***
    - These approaches generate quickly new neighbour topologies by performing small changes in the topology. ✔
    - However, local operators do not allow us to perform an intensive search of the tree space. ✖
    - Example: **Nearest Neighbour Interchange** (NNI).
  - ***Approaches based on global moves***
    - These proposals are defined to avoid phylogenetic algorithms to be trapped on local optima, allowing intensive processing of the search space. ✔
    - Global moves imply more complexity than local moves. ✖
    - Example: **Subtree Pruning and Regrafting** (SPR).

Topological operators II



NNI takes an internal branch of the tree and executes a swap between the nodes in the subtrees situated at the sides of the chosen branch.



SPR consists of removing a subtree from the original topology and regraft it in a different place.

Topological operators III

- Recent studies have introduced new methodologies to combine the properties of both proposals.
- **Parametric Progressive Tree Neighbourhood**, PPN (Goëffon et al., 2008)
- PPN is defined as the set of possible SPR moves such that the distance between the pruned subtree and the regrafting position is at most *d*.
- PPN begins the search for optimal topologies using global moves, reducing progressively *d until this distance is 1, which* represents an NNI move.

In this work we will evaluate the NNI, SPR and PPN proposals.

Evolutionary models

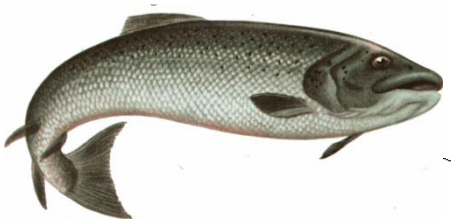
- Evolutionary models try to approximate the reality of the evolutionary process that gave rise to the observed species.
- They provide the probabilities of change from a given state to other in molecular sequences.
- Models must be chosen in accordance with the dataset to be processed.
    1. Jukes-Cantor69 (JC69)
    2. Kimura80 (K80)
    3. Hasegawa-Kishino-Yano85 (HKY85)
    4. Tamura-Nei93 (TN93)
    5. General Time Reversible model (GTR)

Conflicts

- These previous approaches only consider a single objective to be optimized.
- The inference process is carried out in agreement with the chosen criterion.
- Different criteria may lead to **discordant** phylogenetic topologies.

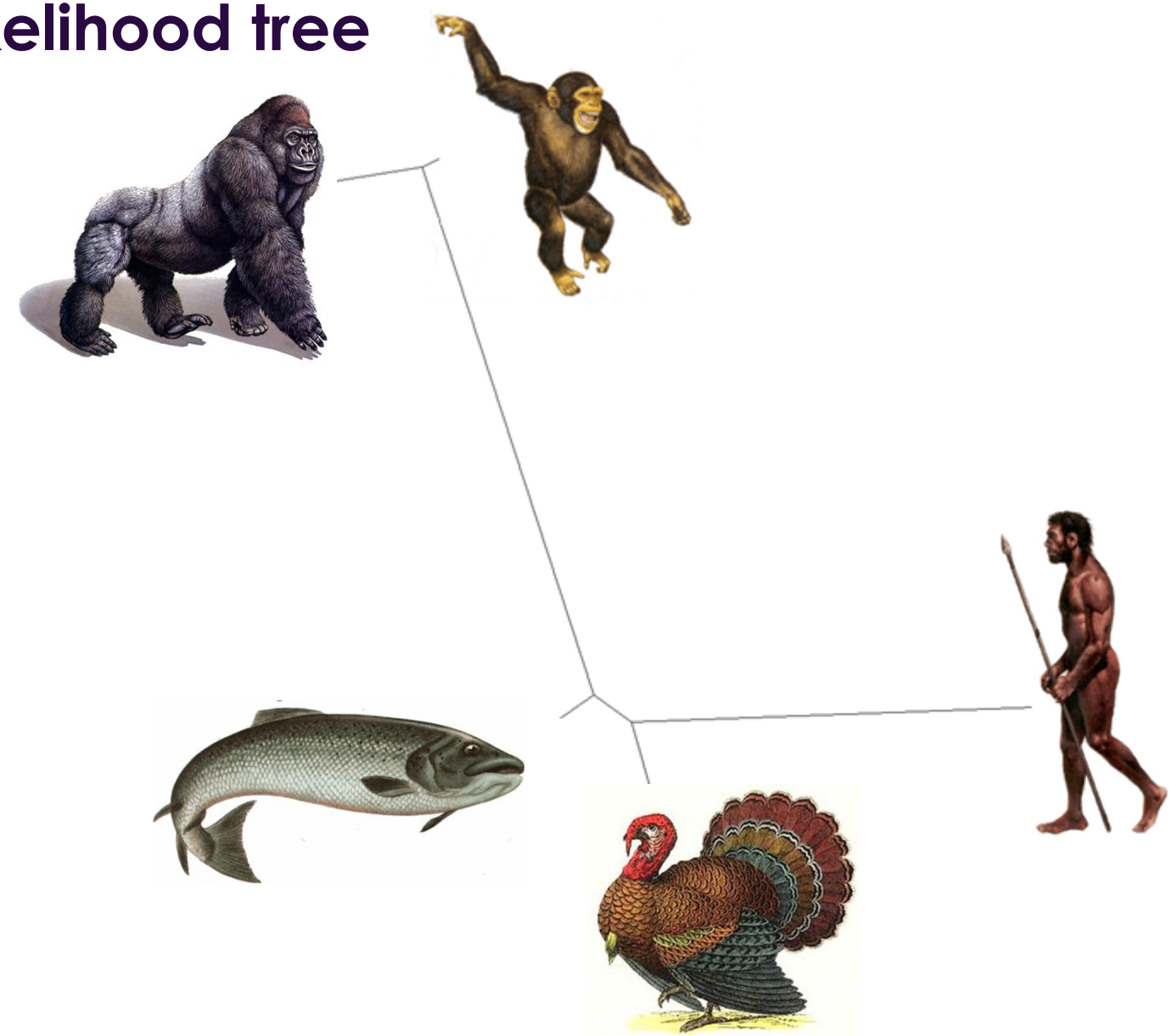
A maximum parsimony tree

TPNC 2012



A maximum likelihood tree

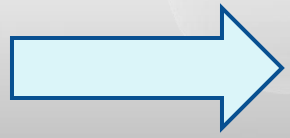
TPNC 2012



Conflicts

- These previous approaches only consider a single objective to be optimized.
- The inference process is carried out in agreement with the chosen criterion.
- Different criteria may lead to **discordant** phylogenetic topologies.

How to resolve this issue?



Multiobjective Optimization

Other authors proposals

- First multiobjective algorithm applied to Phylogenetics.
  - Poladian and Jermiin, 2006.

- Multiobjective algorithm under the minimal evolution and mean-squared error criteria.
  - Coelho et al., 2007.

- **PhyloMOEA**, multiobjective genetic algorithm for **maximum parsimony**, and **maximum likelihood reconstruction**.
  - Cancino and Delbem, 2007.

Our Proposal

- We will tackle the phylogenetic inference problem by considering multiple objectives simultaneously.
  - Maximum parsimony.
  - Maximum likelihood.
- We have chosen the use of Bioinspired Computing.
  - Promising results achieved in a variety of optimization problems.

For so work the honey-bees, creatures that by a rule in nature teach the act of order to a peopled kingdom.
~ William Shakespeare

MULTIOBJECTIVE ARTIFICIAL BEE COLONY DESIGN

Artificial Bee Colony

- Swarm Intelligence algorithm published by D. Karaboga in 2005.
- He proposed a method to resolve classical optimization problems inspired by the collective behaviour of honey bees. We can distinguish 3 types of bees in the hive:

-**Employed bees**: These bees aim to look for and exploit food sources, by examining the neighbourhood of the known food sources.

-**Onlooker bees**: Onlooker bees will decide the sources to be exploited in accordance with the dances performed by employed bees in the dance area.

-**Scout bees**: scouts look randomly at their environment for new undiscovered food sources.

Multiobjective Optimization

- Our proposal tries to extend the ABC design by applying Multiobjective Optimization.
  - ➢ We define a **Multiobjective Optimization Problem** as the problem of finding those solutions which optimize simultaneously two or more objective functions.

$$\begin{aligned} \text{optimize } & y = \vec{f}(x) = (f_1(x), f_2(x), ..., f_n(x)), \\ \text{where } & x = (x_1, x_2, ..., x_k) \in X, \\ & y = (y_1, y_2, ..., y_n) \in Y. \end{aligned}$$

- The main goal of multiobjective algorithms is to generate a set of trade-off (***Pareto***) solutions satisfying different criteria.
- Solutions are assessed by using the ***dominance*** concept: a solution dominates other one if and only if the first solution has better or equal scores in all objectives than the second one and, at least, it is better in one of them.
- The representation of Pareto solutions in the value space of the objective functions is known as ***Pareto front***.

MOABC Scheme

1: C ⇦ Initialize and Evaluate Population (C, swarmSize)

2: ParetoFront ⇦ 0

3: **for i = 1 to maxGenerations do**

4: Exploitation step by employed bees (C, swarmSize/2)

5: Selection and exploitation step by onlooker bees (C, swarmSize/2)

6: Exploration step by scout bees (C, swarmSize, limit)

7: ParetoFront ⇦ Save Solutions(C, ParetoFront)

8: **end for**

Initialization

- Initially, the first half of the population will take the role of employed bees, and the remaining half will be onlooker bees.
- Employed bees are initialized by:
  - Assigning starter trees from a repository of 1000 phylogenetic trees generated by bootstrap analysis.
  - Configuring the parameters of the evolutionary model.
- In this work, we evaluate the performance achieved by five different evolutionary models:
  1. JC69
  2. K80
  3. HKY85
  4. TN93
  5. GTR.

Employed Bees

- Employed bees search for solutions in the neighbourhood.
- Its initial solution is compared to the result of mutating it using:
  - Nearest Neighbour Interchange (NNI) moves.
    - NNI properties are suitable to model employed bees behaviour.
  - Modifying randomly selected branch lengths using a gamma distribution.
- We calculate for each solution a MOFitness in order to save the most promising solution.

Comparing solutions

***MOFitness**(b) = Dominates(b)*
*+ isDominated(b)*swarmSize*

Onlooker Bees

- Onlooker bees will decide which solutions must be exploited in accordance with the information provided by employed bees.
- We compute a vector to define selection probabilities for each solution. These solutions are previously sorted by using two operators proposed by K. Deb (NSGA-II):
  - *Fast non dominated sort*.
  - *Crowding distance*.
- An onlooker bee will verify this vector and choose one of the current solutions, computing new neighbours and selecting the most promising ones by using MOFitness.

Scout Bees

- We define scout bees to avoid the algorithm to be trapped in local optima.

If the solution associated to a bee is not improved in ***limit*** *iterations*

→ Scout bee conversion

- Scout bee conversion begins with the assignation of a new starter topology from the initial repository.
- We introduce an optimization step to allow the new trees to compete with the solutions found by the algorithm.

Possible strategies

NNI
SPR
PPN

+

Branch lengths optimization (Gradient method)

There are three principal means of acquiring knowledge… observation of nature, reflection, and experimentation. Observation collects facts; reflection combines them; experimentation verifies the result of that combination.

~ Denis Diderot

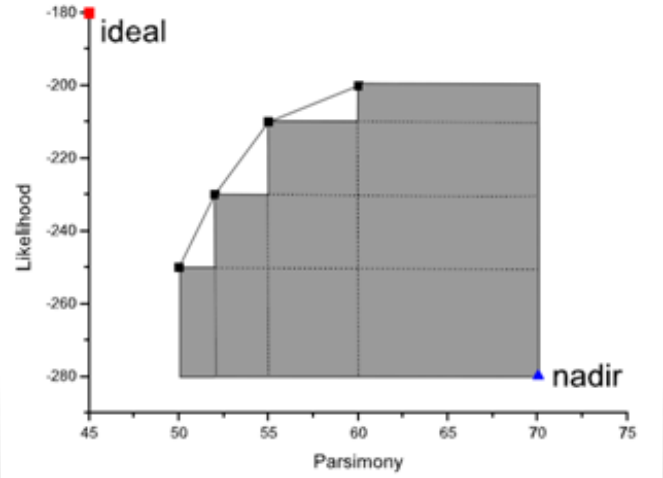
EXPERIMENTAL METHODOLOGY AND RESULTS

Experimental Methodolgy

- We have carried out a variety of experiments to assess the performance of MOABC by considering:
  - Different neighbourhood methodologies.
  - Different evolutionary models.
- Our approach has been tested on four public nucleotide data sets:
  - **rbcL_55**. 55 sequences (1314 nucleotides per sequence) of the rbcL gene from different species of green plants.
  - **mtDNA_186**. 186 sequences (16608 nucleotides) of human mitochondrial DNA.
  - **RDP-II_218**. 218 sequences (4182 nucleotides) of prokayotic RNA.
  - **ZILLA_500.** 500 sequences (759 nucleotides) from rbcL plastid gene.

Experiments on Neighbourhood Methods

- We have carried out 10 MOABC runs for each dataset and neighbourhood, using the *GTR model*.



Pareto fronts have been evaluated from a multiobjective perspective by using the hypervolume metrics.

Also, we have taken the execution times required by each configuration.

Experiments on Neighbourhood Methods

Hypervolume metrics		
Neighbourhood	**Mean**	**Time(s)**
	rbcl_55	
NNI	71.506%	**1611.0**
PPN	**71.620%**	**1770.7**
SPR	**71.631%**	1918.5
	mtDNA_186	
NNI	69.888%	**12798.9**
PPN	**69.994%**	**17223.9**
SPR	**69.998%**	43369.9
	RDPII_218	
NNI	73.147%	**18462.2**
PPN	**74.022%**	**27622.0**
SPR	**74.063%**	54573.6
	ZILLA_500	
NNI	71.250%	**20595.4**
PPN	**72.345%**	**37499.4**
SPR	**72.566%**	104539.0

- As we increase the complexity of the dataset, the hypervolume values achieved by PPN and SPR overcome the values obtained by NNI.
- SPR neighbourhood achieves the best hypervolume values.
- However, execution times increase significantly with the complexity of the dataset.
- PPN proposal gets significant hypervolume values without dramatic times.
- *PPN represents a compromise between NNI and SPR.*

Experiments on Evolutionary Models

- We have performed 10 complete runs of the algorithm for each evolutionary model and dataset to evaluate the impact of models in the inference process.
  - Using the PPN neighbourhood.

➢Pareto fronts have been evaluated by using:

1. Hypervolume.
2. Statistical evaluation:
   - *Akaike Information Criterion*. Amount of information lost when a model is used to approximate the reality of the evolutionary process.
   - *Bayesian Information Criterion*. Evaluation according to Bayesian estimations.

L = likelihood
K = model parameters
N=sites in molecular sequences

$$AIC = -2L + 2K$$
$$BIC = -2L + K \log N$$

Experiments on Evolutionary Models

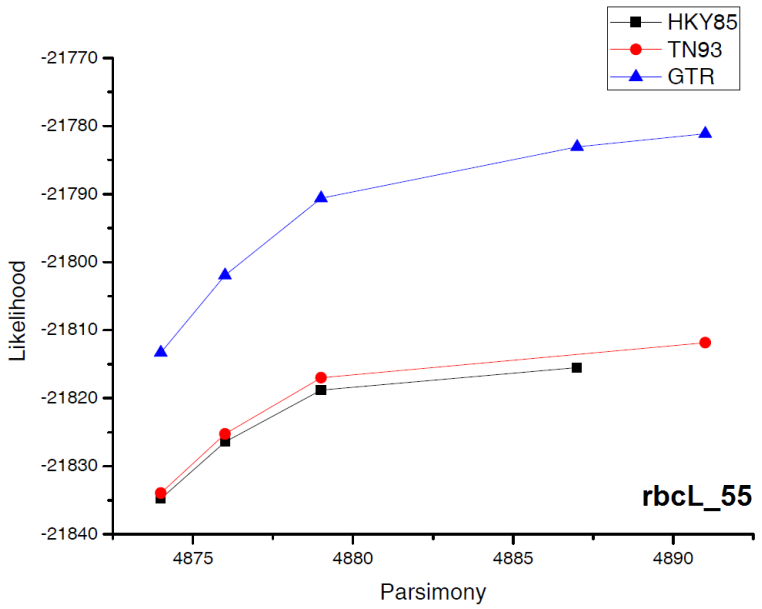
TPNC 2012

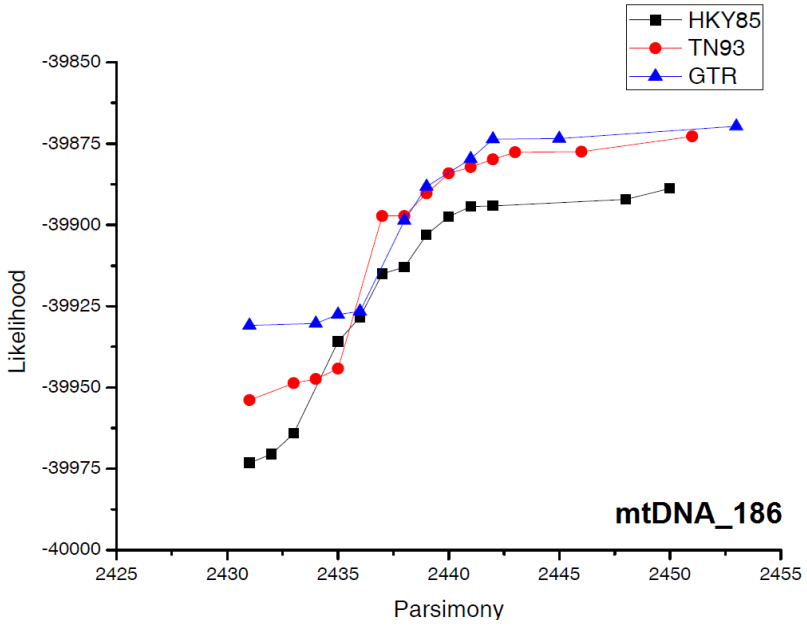
Evolutionary Model	Best parsimony tree AIC	Best parsimony tree BIC	Best likelihood tree AIC	Best likelihood tree BIC	Hypervolume Mean
		rbcL_55			
JC69	46300.39	46865.10	46294.46	46859.17	20.76%
K80	44192.64	44762.53	44184.37	44754.26	63.47%
HKY85	43895.57	44481.00	43856.99	44442.42	70.09%
TN93	43895.87	44486.48	43851.65	44442.26	70.17%
GTR	**43860.63**	**44466.79**	**43796.25**	**44402.41**	**71.62%**
		mtDNA_186			
JC69	85999.75	88862.99	85856.90	88720.14	23.52%
K80	82434.34	85305.30	82302.63	85173.60	53.88%
HKY85	80696.38	83590.49	80527.47	83421.58	69.66%
TN93	80659.81	83561.65	80497.58	**83399.42**	69.93%
GTR	**80619.80**	**83544.79**	**80497.19**	83422.18	**69.99%**
		RDPII_218			
JC69	288326.18	291083.45	275063.45	277820.72	54.66%
K80	283955.69	286719.29	270206.88	272970.49	68.73%
HKY85	273702.77	276485.39	269171.71	271954.33	73.53%
TN93	273689.13	276478.09	269132.94	271921.90	73.75%
GTR	**273525.65**	**276333.62**	**269043.31**	**271851.29**	**74.02%**
		ZILLA_500			
JC69	171299.68	175927.05	170463.43	175090.80	36.34%
K80	165320.17	169952.17	164198.31	168830.31	67.62%
HKY85	165144.06	169789.96	163937.66	168583.56	68.98%
TN93	165144.04	169794.57	163939.37	168589.90	68.94%
GTR	**164828.76**	**169493.18**	**163212.81**	**167877.24**	**72.34%**

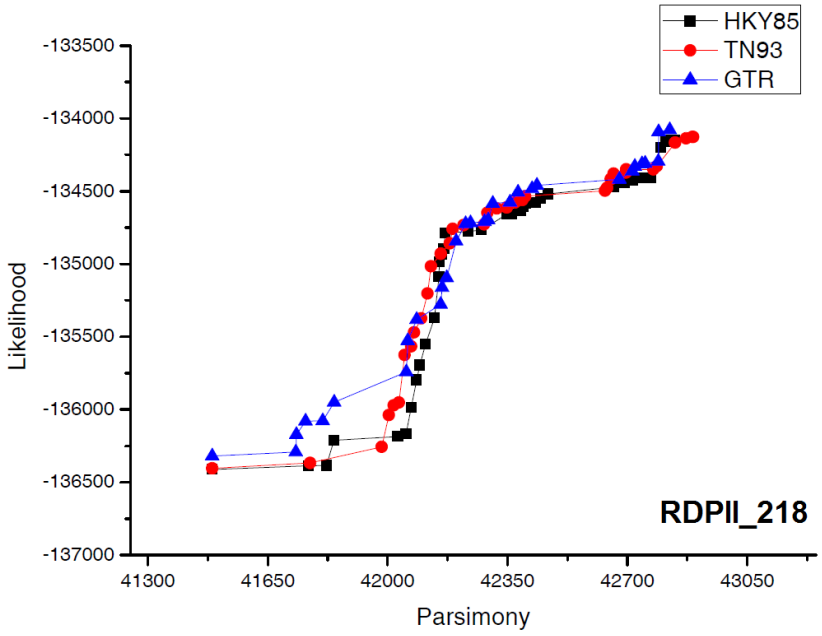
- ✓ *HKY85, TN93 and GTR* achieve the best hypervolume values for each dataset.
- ✓ AIC and BIC select *GTR and TN93 as the models which generate the best extreme* solutions for all the instances.
- ✓ HKY85, TN93 and GTR represent the most accurate models.

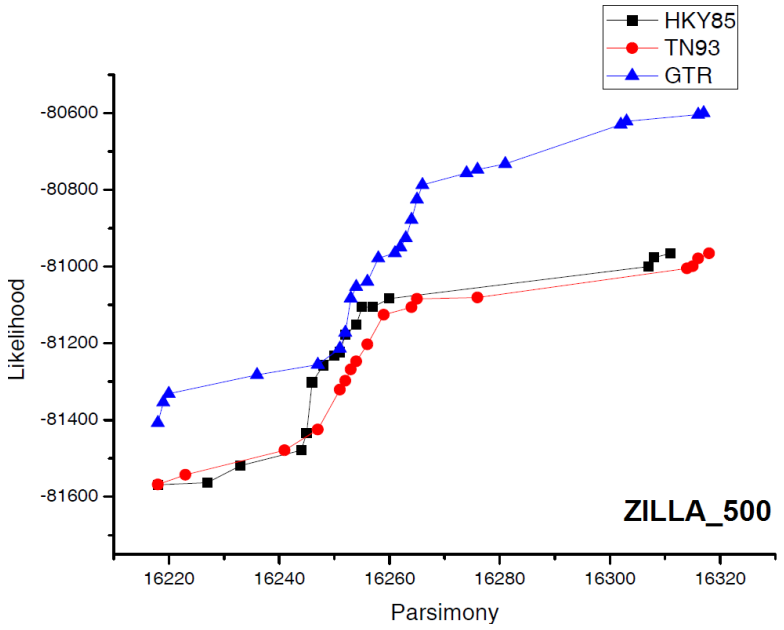
Pareto fronts

TPNC 2012









Comparisons with other authors I

- We compare our proposal with the results reported by PhyloMOEA.
  - Using HKY85 with among-site rate variation (Cancino and Delbem, 2010).

MOABC HKY85+Γ

	Best parsimony tree		Best likelihood tree	
Dataset	Parsimony	Likelihood	Parsimony	Likelihood
rbcL_55	**4874**	-21834.78	4887	**-21815.49**
mtDNA_186	**2431**	-39973.19	2450	**-39888.73**
RDPII_218	**41488**	-136412.38	42837	**-134146.85**
ZILLA_500	**16218**	-81596.03	16311	**-80965.83**

PhyloMOEA HKY85+Γ

Dataset	Best parsimony score	Best likelihood score
rbcL_55	4874	-21889.84
mtDNA_186	2437	-39896.44
RDPII_218	41534	-134696.53
ZILLA_500	16219	-81018.06

MOABC swarm intelligence approach improves the results reported by using other authors' multiobjective proposals.

Comparisons with other authors II

- Comparisons with biological methods that represent the state-of-the-art in Phylogenetics: TNT (maximum parsimony) and RAxML (maximum likelihood)

Best parsimony

Dataset	MOABC	TNT
rbcL_55	**4874**	**4874**
mtDNA_186	**2431**	**2431**
RDPII_218	**41488**	**41488**
ZILLA_500	**16218**	**16218**

Best likelihood tree

MOABC GTR

Dataset	Parsimony	Likelihood
rbcL_55	**4891**	**-21781,12**
mtDNA_186	**2453**	**-39869,59**
RDPII_218	**42824**	**-134078,65**
ZILLA_500	16317	**-80599,4**

RAxML GTR

Dataset	Parsimony	Likelihood
rbcL_55	4893	-21791,98
mtDNA_186	**2453**	-39869,63
RDPII_218	42894	-134079,42
ZILLA_500	**16305**	-80623,5

MOABC achieves the reference scores provided by one of the most powerful parsimony tools.

Likelihood trees improve at least one of the considered criteria when comparing with RAxML maximum likelihood trees.

We must use time wisely and forever realize that the time is always ripe to do right.

~ Nelson Mandela

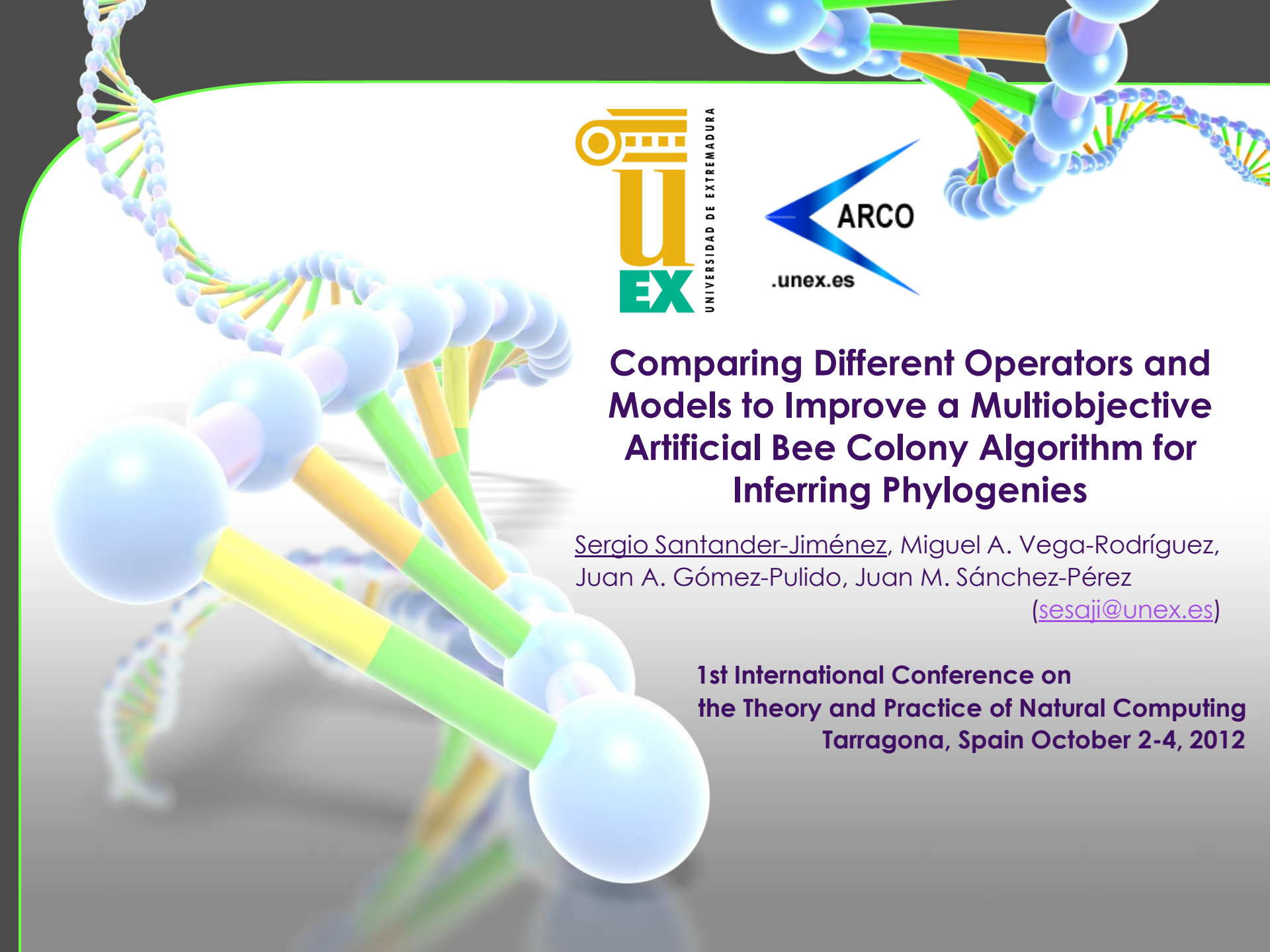
CONCLUSIONS AND FUTURE RESEARCH LINES

Conclusions

- We have reported a study on different neighbourhood-based proposals and evolutionary models to improve a multiobjective swarm intelligence approach to phylogenetic inference.
- Experiments have been performed on four real nucleotide data sets and Pareto solutions have been evaluated by using the hypervolume metrics and the AIC and BIC tests.
- Results suggest that the use of PPN neighbourhood and evolutionary models like *GTR obtain high-quality* phylogenetic solutions in reasonable times.

Future Lines

- Study of high performance computing techniques to address phylogenetic analyses on data sets with thousands of species.
  - Applying fine and coarse-grained parallelism on hybrid architectures.
- Development of other bioinspired algorithms for inferring phylogenies.
  - Multiobjective Firefly Algorithm.
  - Multiobjective Bat Algorithm.
  - Differential Evolution.
- Analysis of complex data sets, and comparisons using other multiobjective metrics.
  - Attainment surface.
  - Set coverage.



Comparing Different Operators and Models to Improve a Multiobjective Artificial Bee Colony Algorithm for Inferring Phylogenies

Sergio Santander-Jiménez, Miguel A. Vega-Rodríguez, Juan A. Gómez-Pulido, Juan M. Sánchez-Pérez
(sesaji@unex.es)

1st International Conference on the Theory and Practice of Natural Computing
Tarragona, Spain October 2-4, 2012