Microsoft Research
FacultySummit

FUTURE/WORLD

2011 — 2031

Microsoft Research FacultySummit

LATAM Workshop 2011

Information Technology Applied to Bioenergy Genomics: Probabilistic Annotation using Artificial Intelligence

Advisor: Prof. Dr Ricardo Z. N. Vêncio

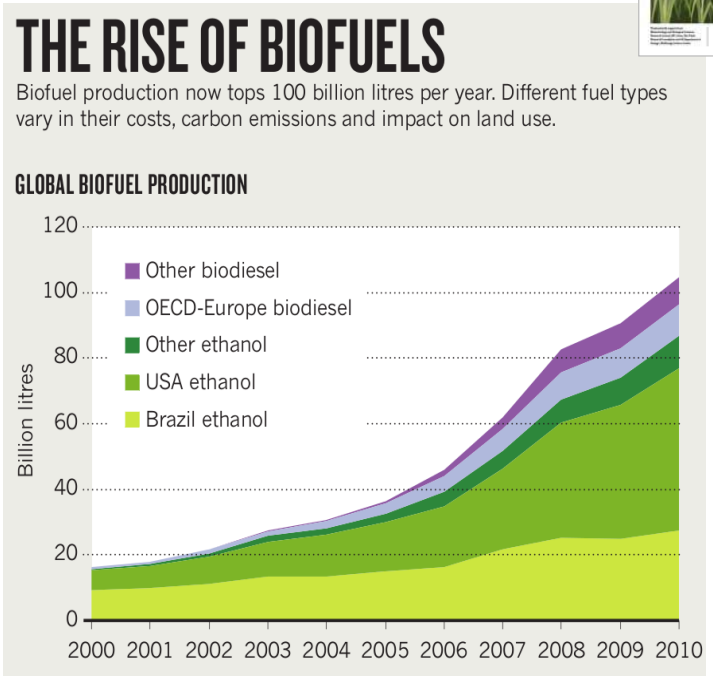
Danillo C. Almeida-e-Silva, grad student
Department of Computing and Mathematics - FFCLRP
University of Sao Paulo – Brazil

FAPESP USP FUTURE WORLD 2011 2031

adapted from:

Introduction: Next generation biofuels

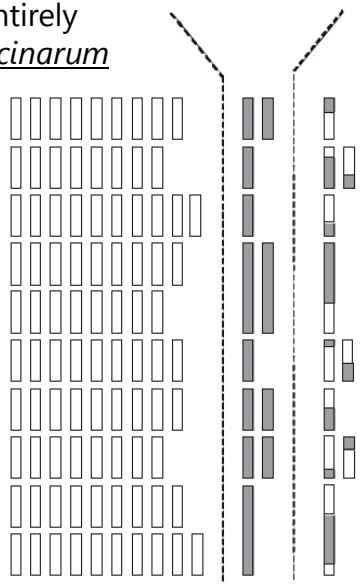
Fairley P., Nature (2011)



why?

70% to 80% are inherited entirely from *S. officinarum*

10% to 20% of their chromosomes are inherited in their entirety from *S. spontaneum*



Around 10% are the result of recombination between chromosomes from the two ancestral species.

adapted from:

Genomics of tropical crop plants

Moore & Ming (2008)

Motivation

Function: **unknown**
limited technological use

Function: **known**
possible technological use

Function: **unknown**
limited technological use

Function: **known**
possible technological use

Information Technology Applied to Bioenergy Genomics: Probabilistic Annotation using Artificial Intelligence

Function: **unknown**
limited technological use

Function: **known**
possible technological use

The aim is to develop user-friendly software to rationally guess the biological functions of genes from sugarcane

Information Technology Applied to Bioenergy

Genomics: Probabilistic Annotation using Artificial Intelligence

Function: **unknown**
limited technological use

Function: **known**
possible technological use

We propose to estimate the probability:

p = **P**(gene X has function A | evidence E)

instead of just saying gene X is A.

Information Technology Applied to Bioenergy

Genomics: Probabilistic Annotation using Artificial Intelligence

Function: **unknown**
limited technological use

Function: **known**
possible technological use

We propose to estimate the probability:

p = **P**(gene X has function A | evidence E)

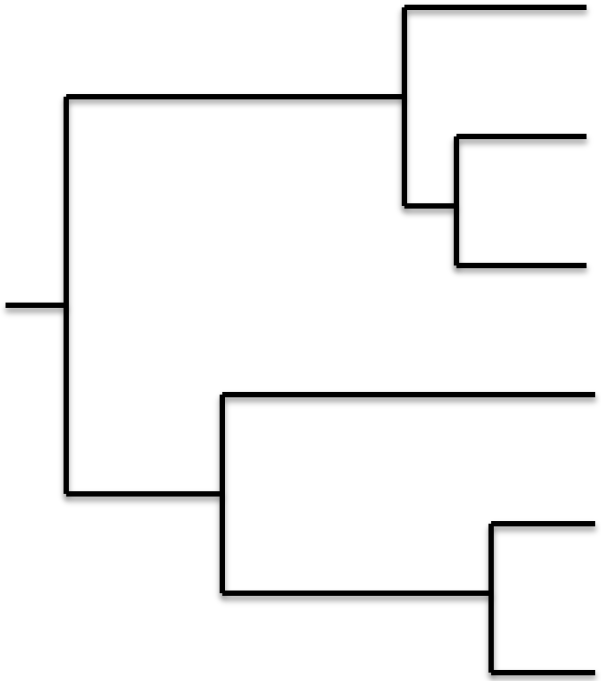
instead of just saying gene X is A.

modelling quantitative and qualitative uncertainty

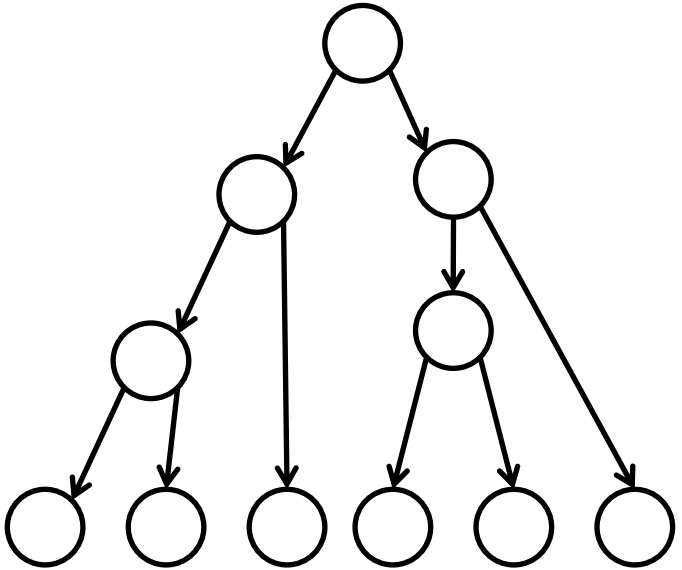
Information Technology Applied to Bioenergy

Genomics: Probabilistic Annotation using Artificial Intelligence

Phylogenomic



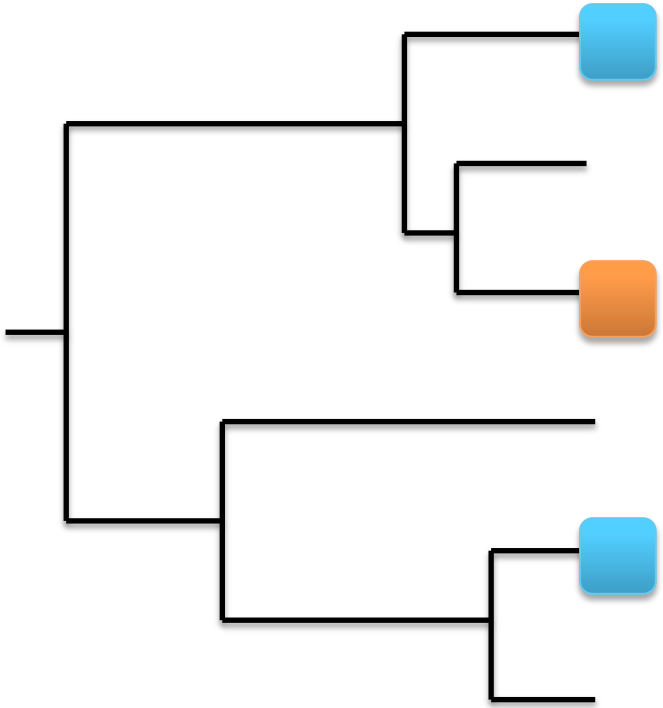
Probabilistic



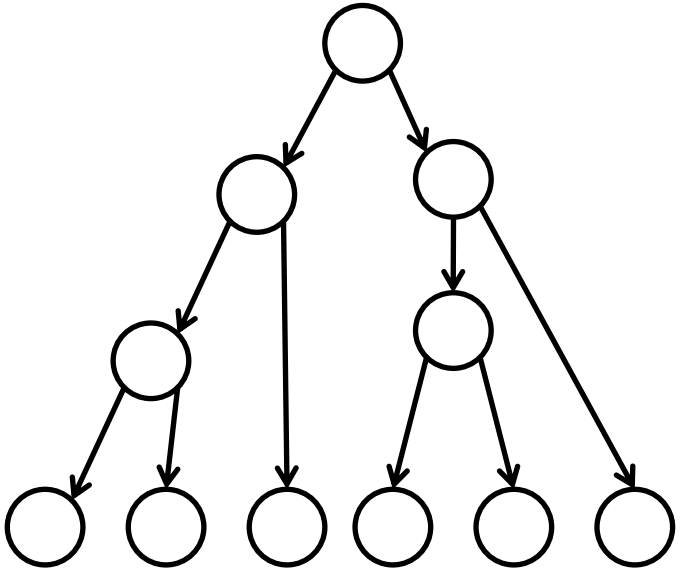
Bayesian Networks

Adapted from: Engelhardt *et al*., 2005

Phylogenomic



Probabilistic

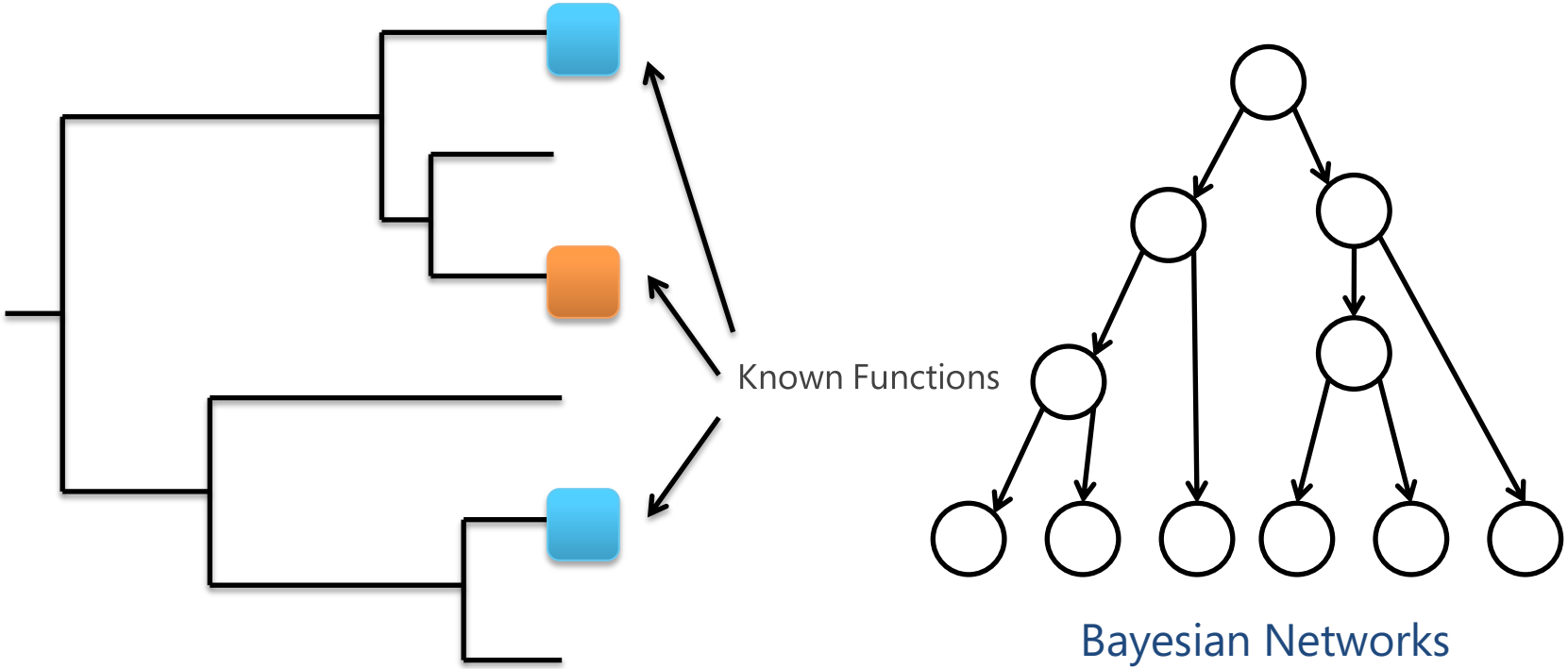


Bayesian Networks

Adapted from: Engelhardt *et al*., 2005

Phylogenomic

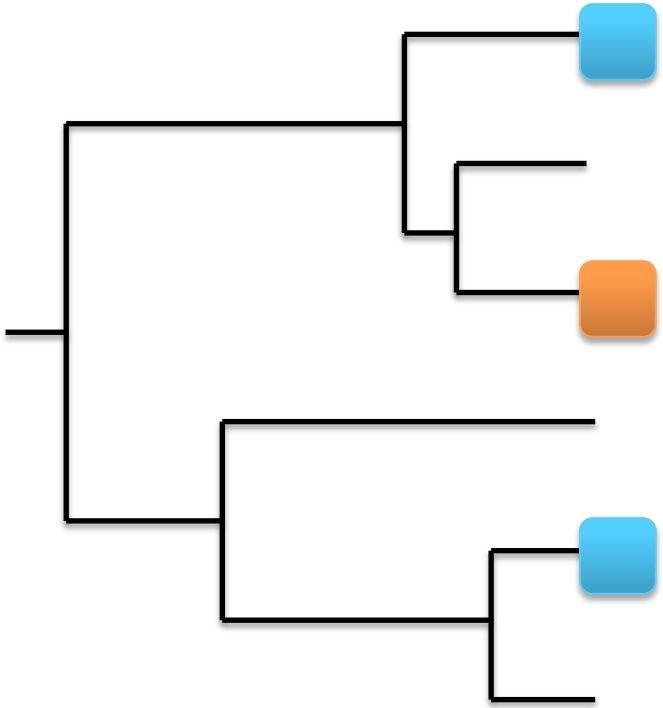
Probabilistic



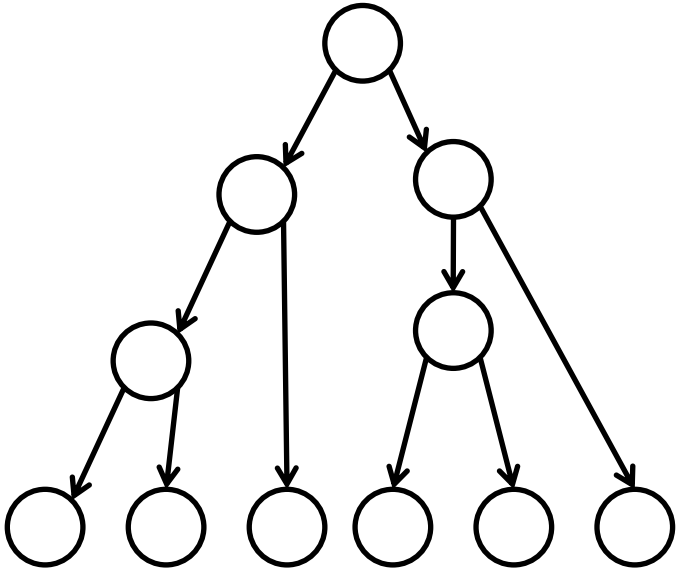
Bayesian Networks

Adapted from: Engelhardt *et al*., 2005

Phylogenomic



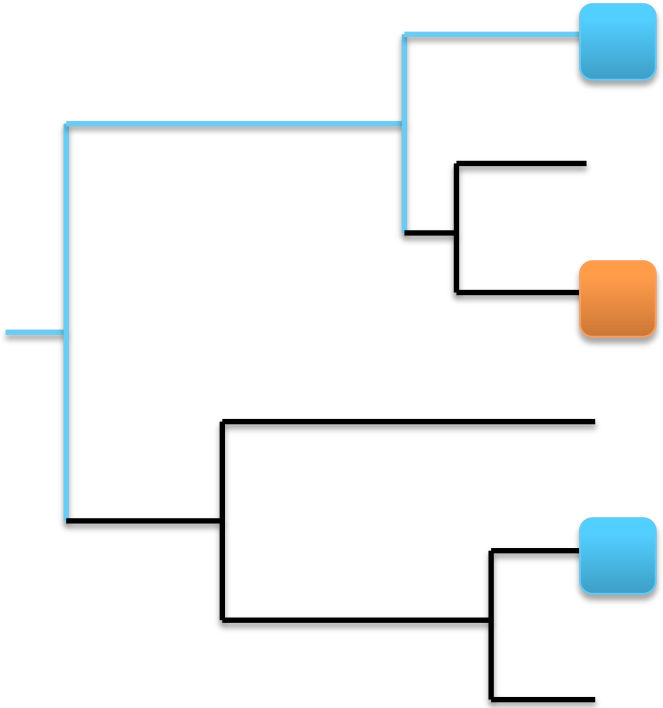
Probabilistic



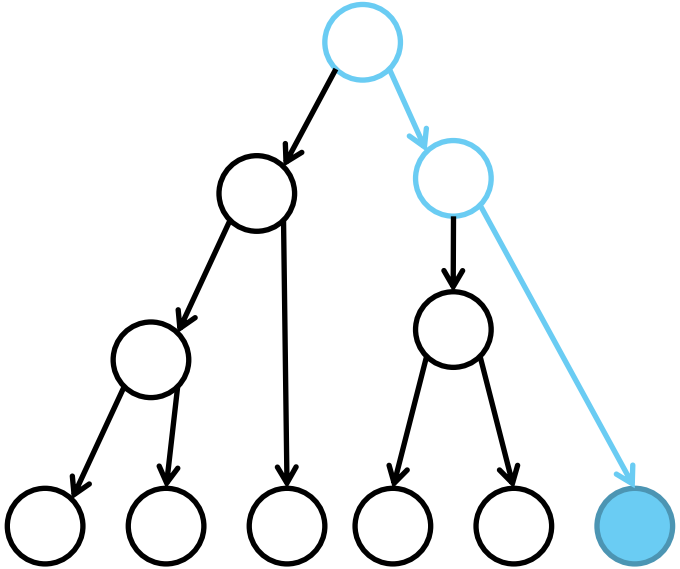
Bayesian Networks

Adapted from: Engelhardt *et al*., 2005

Phylogenomic



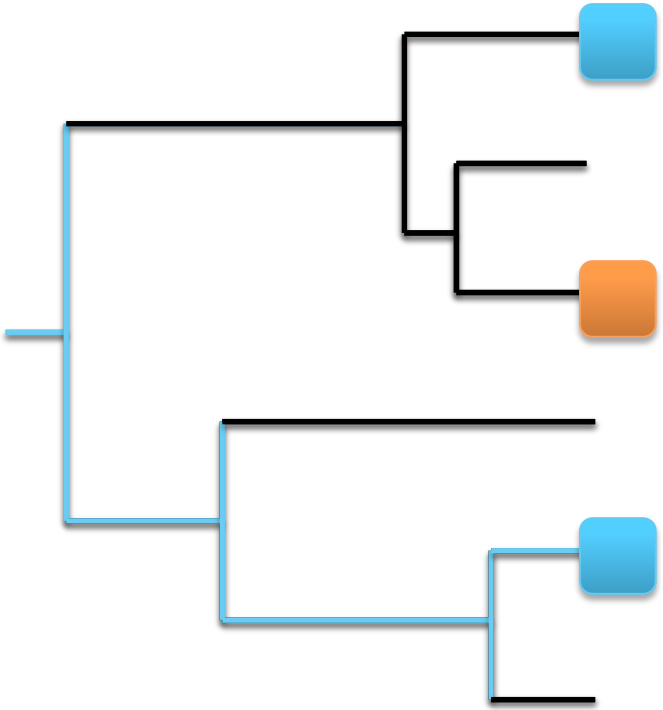
Probabilistic



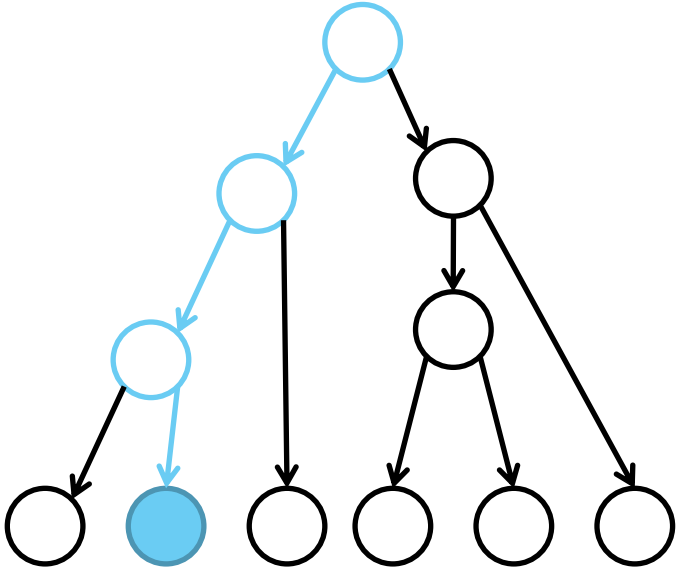
Bayesian Networks

Adapted from: Engelhardt *et al*., 2005

Phylogenomic



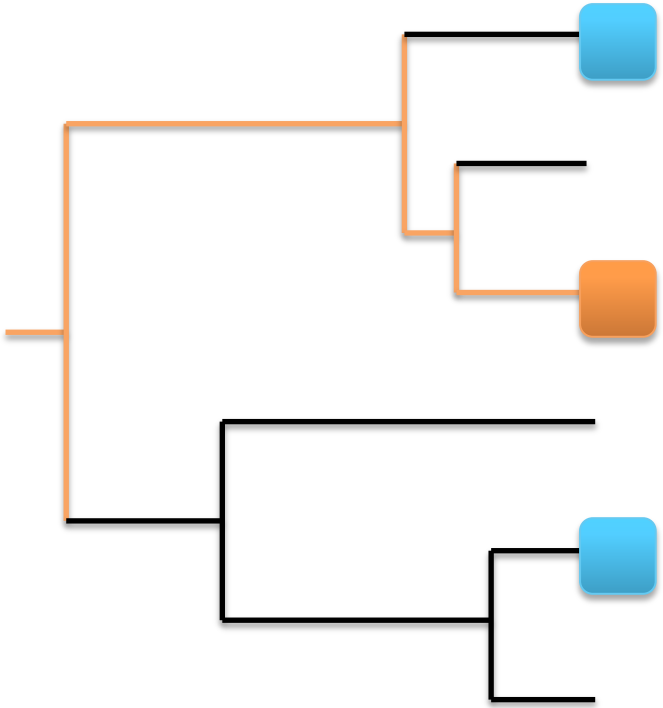
Probabilistic



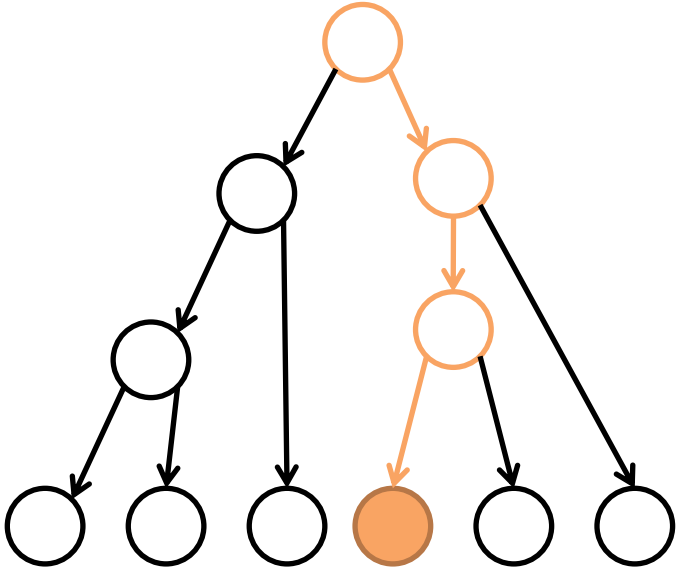
Bayesian Networks

Adapted from: Engelhardt *et al*., 2005

Phylogenomic



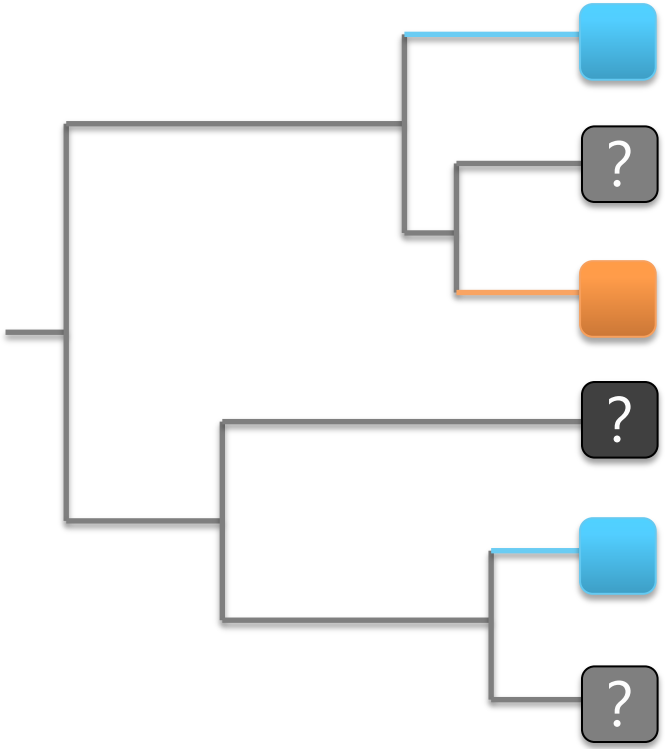
Probabilistic



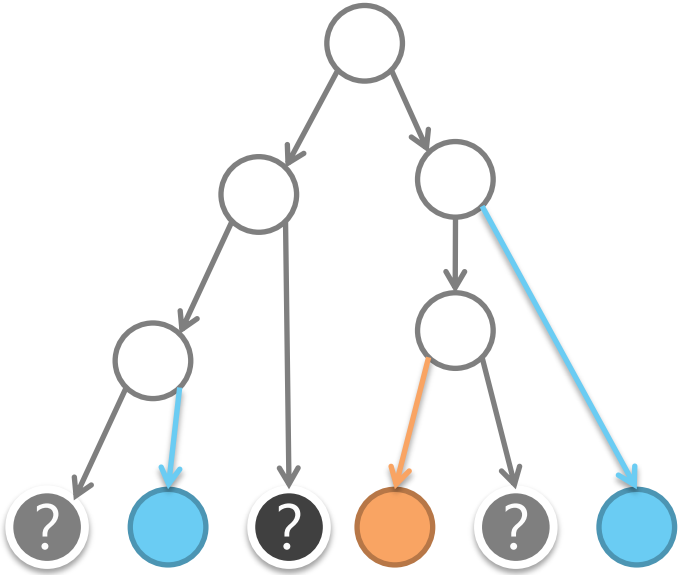
Bayesian Networks

Adapted from: Engelhardt *et al*., 2005

Phylogenomic



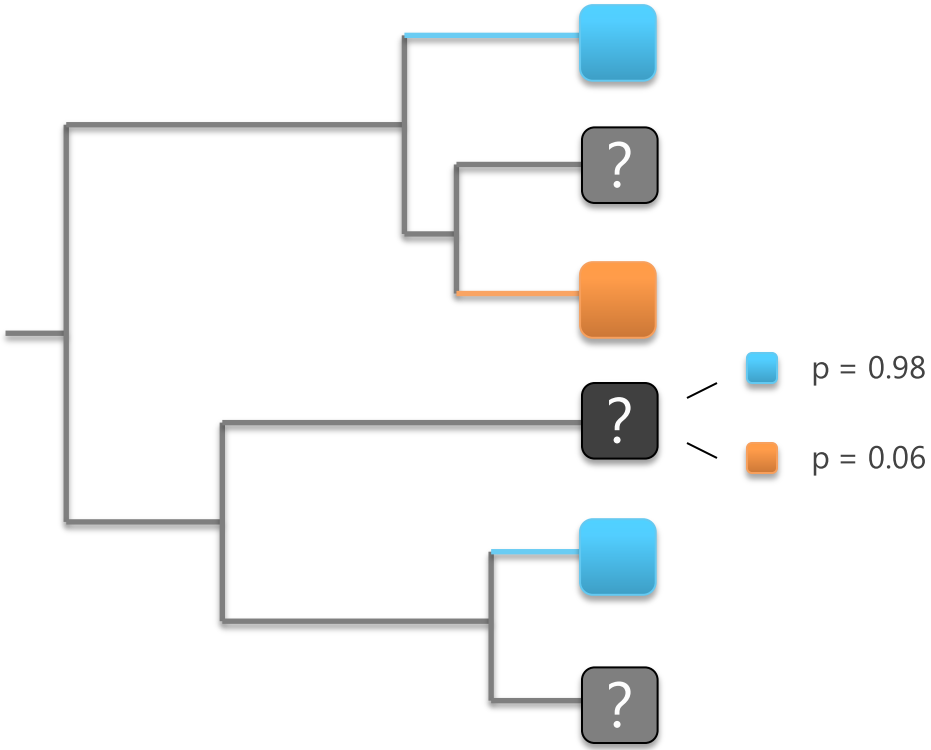
Probabilistic



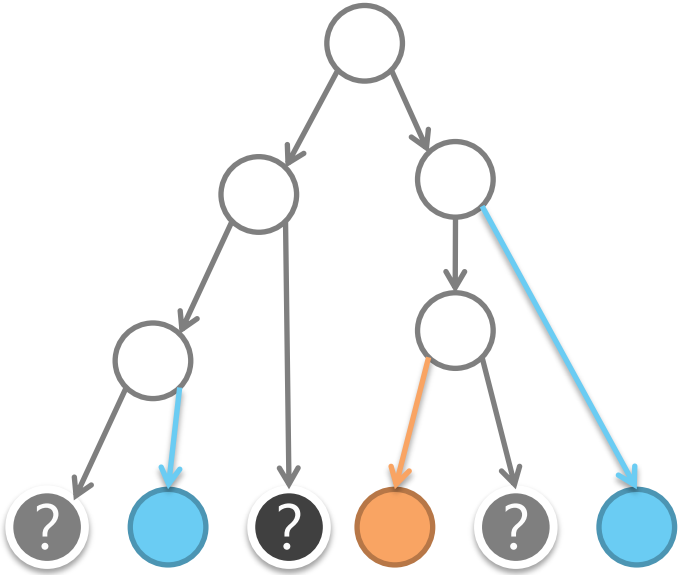
Bayesian Networks

Adapted from: Engelhardt *et al*., 2005

Phylogenomic

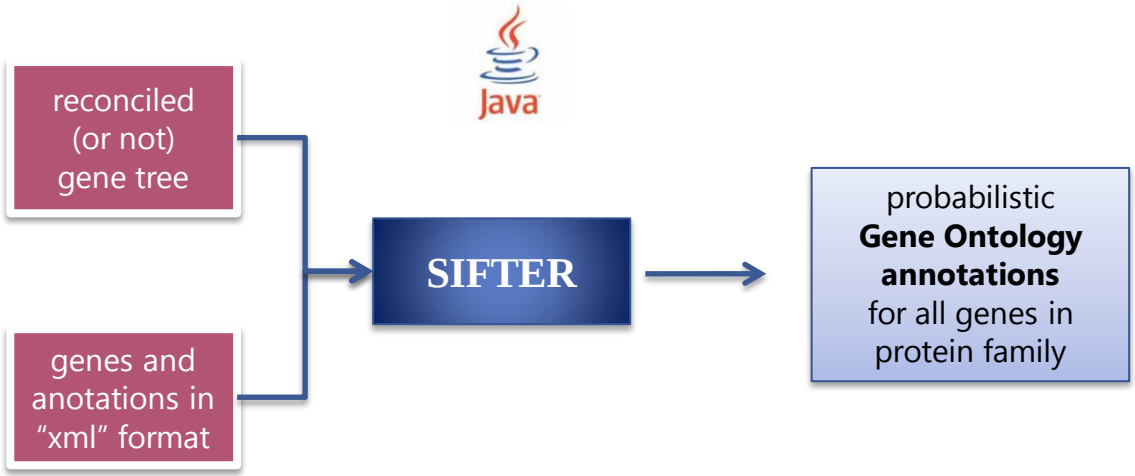


Probabilistic



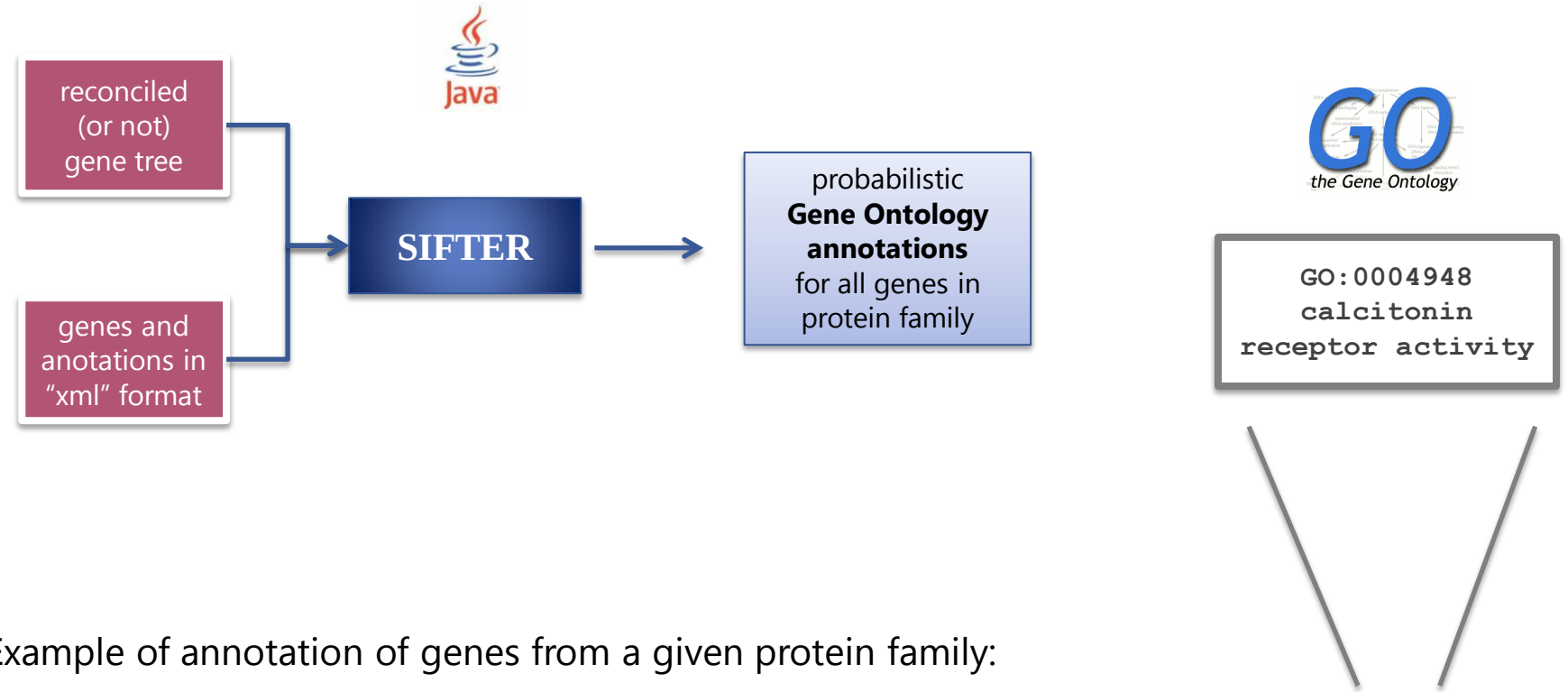
Bayesian Networks

Adapted from: Engelhardt *et al.*, 2005



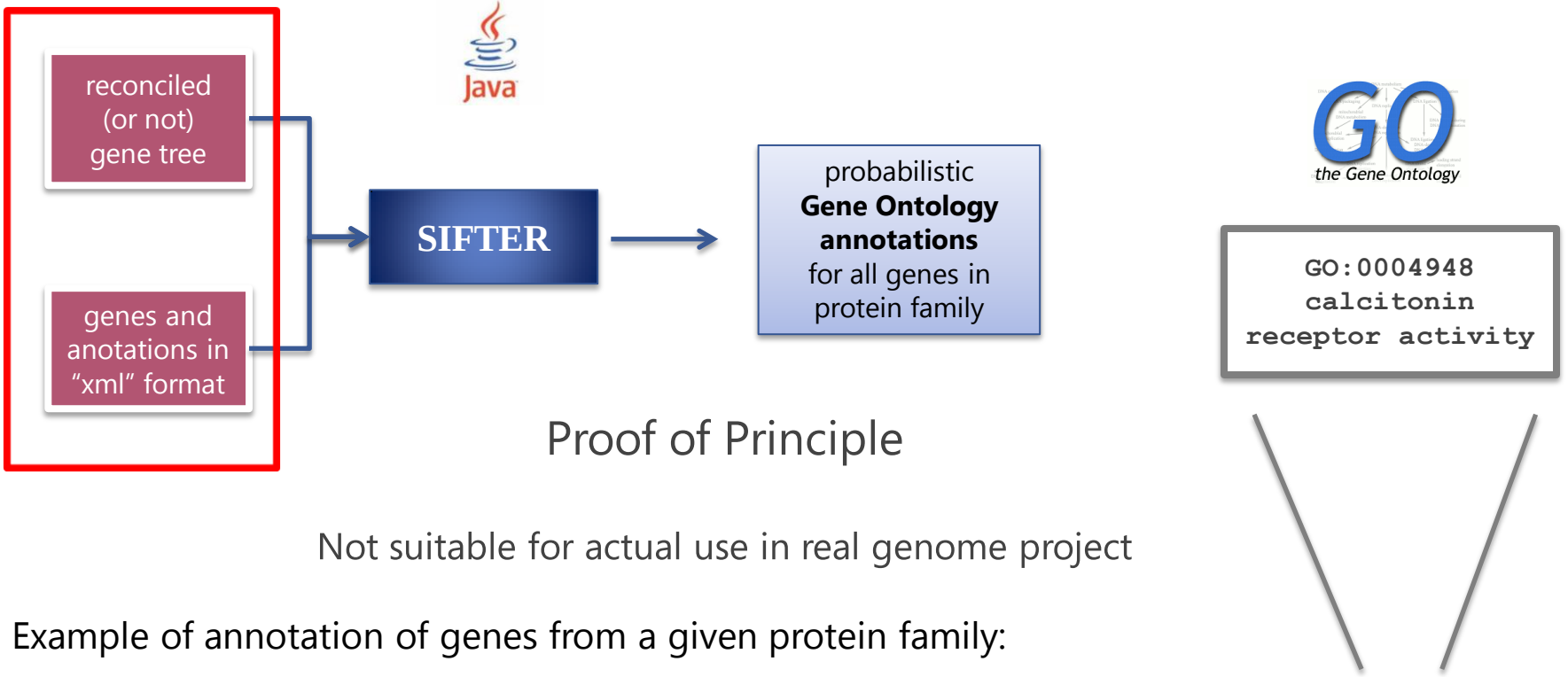
Example of annotation of genes from a given protein family:

```
a7ryr2_nemve        0.9960239633105679      0.07628977950793243     0.07950043523590085   4948
b3rjg4_triad        0.9819255688728563      0.15671589849437728     0.15943461283254143   4948
```



Example of annotation of genes from a given protein family:

a7ryr2_nemve	0.9960239633105679	0.07628977950793243	0.07950043523590085	**4948**
b3rjg4_triad	0.9819255688728563	0.15671589849437728	0.15943461283254143	**4948**



Proof of Principle

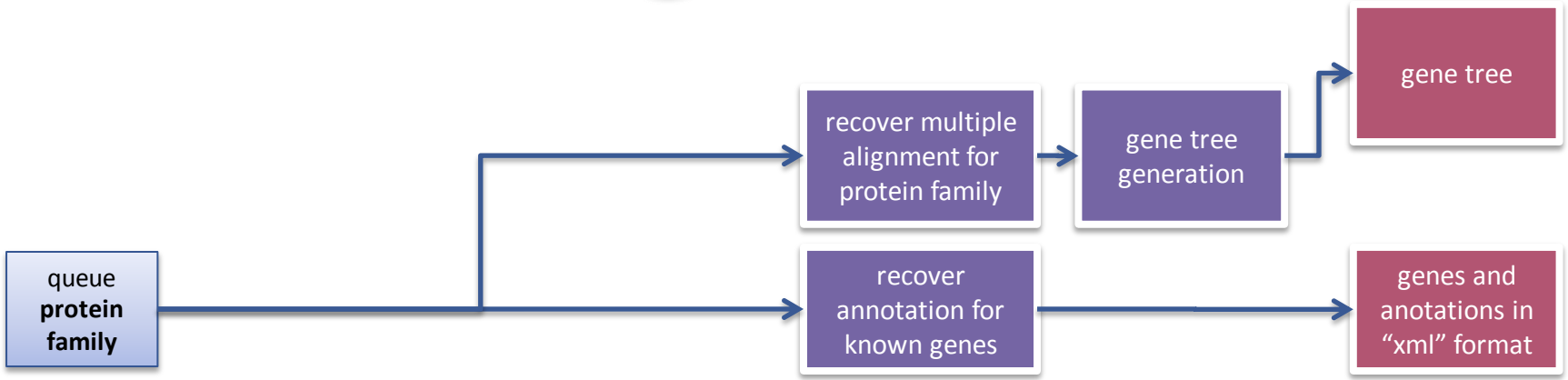
Not suitable for actual use in real genome project

Example of annotation of genes from a given protein family:

```
a7ryr2_nemve    0.9960239633105679    0.07628977950793243    0.07950043523590085    4948
b3rjg4_triad    0.9819255688728563    0.15671589849437728    0.15943461283254143    4948
```

State of the Art

python



Pfam

GO
the Gene Ontology

UniProt

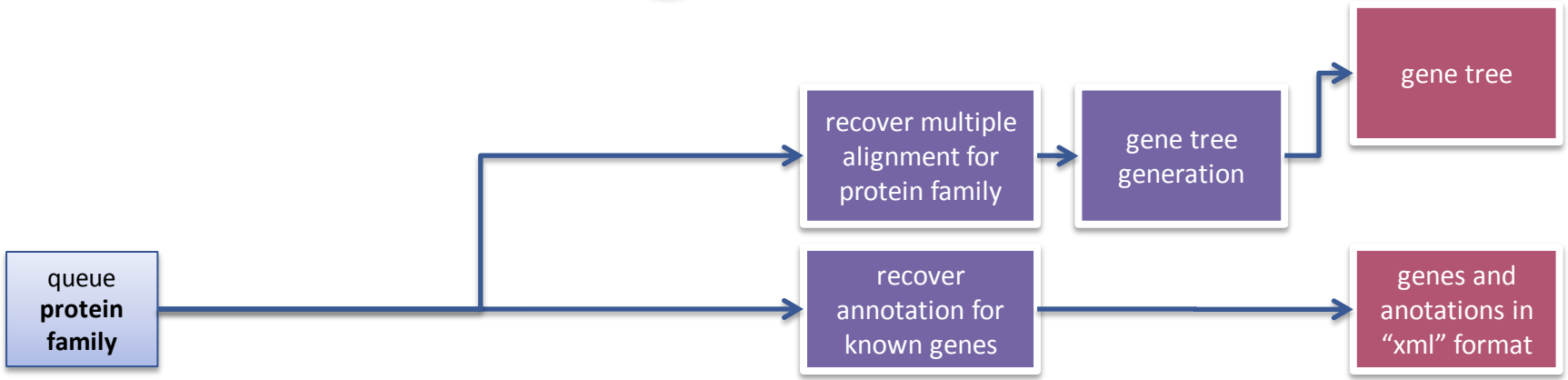
Fast Tree

Starting Point for Improvement

Microsoft Research FacultySummit

State of the Art

python™




```
PF00258
PF05681
(...)
```

Pfam

GO
the Gene Ontology

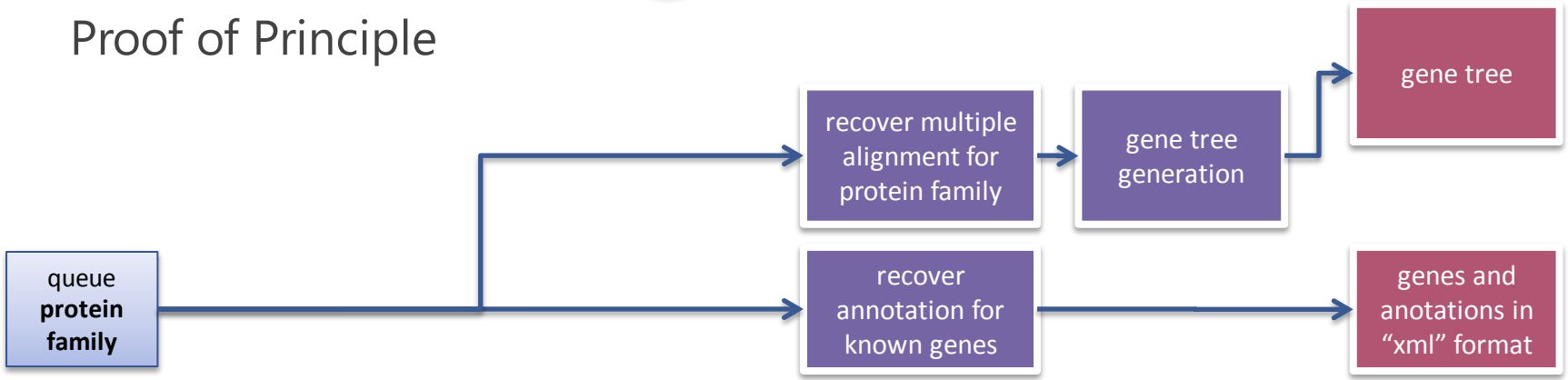
UniProt

Fast Tree

Starting Point for Improvement

State of the Art

Proof of Principle



Pfam

GO the Gene Ontology

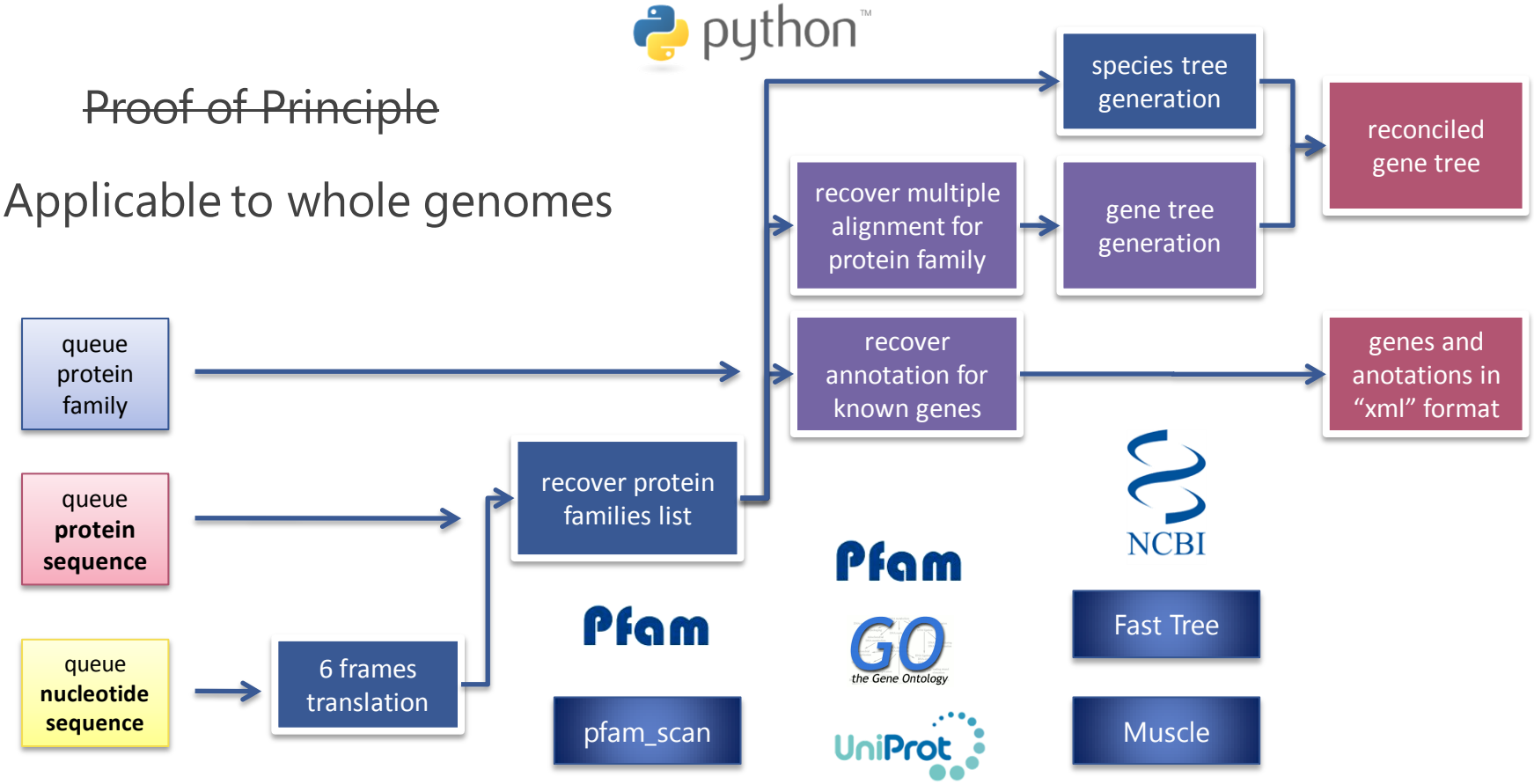
UniProt

Fast Tree

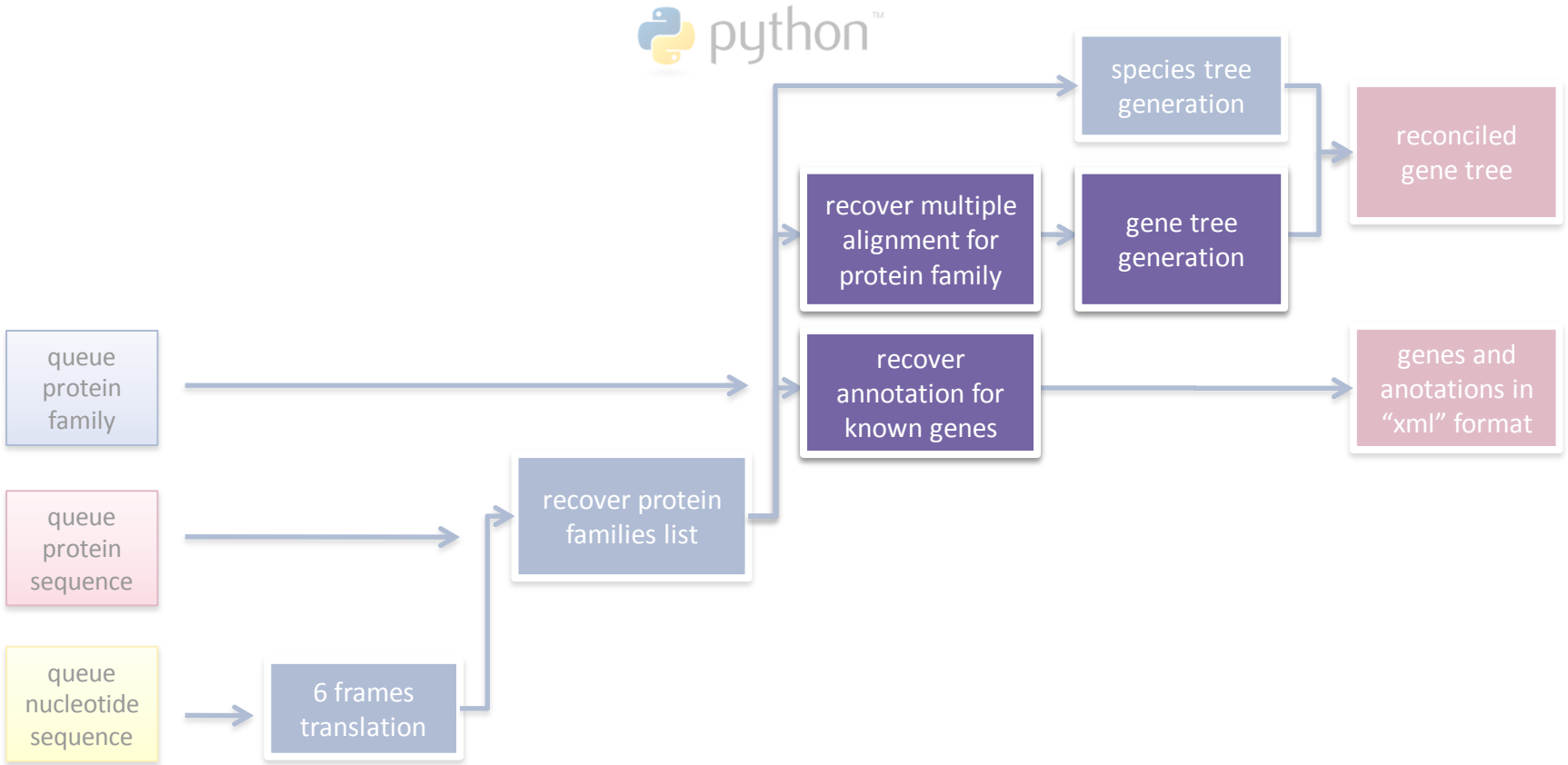
Starting Point for Improvement

~~Proof of Principle~~

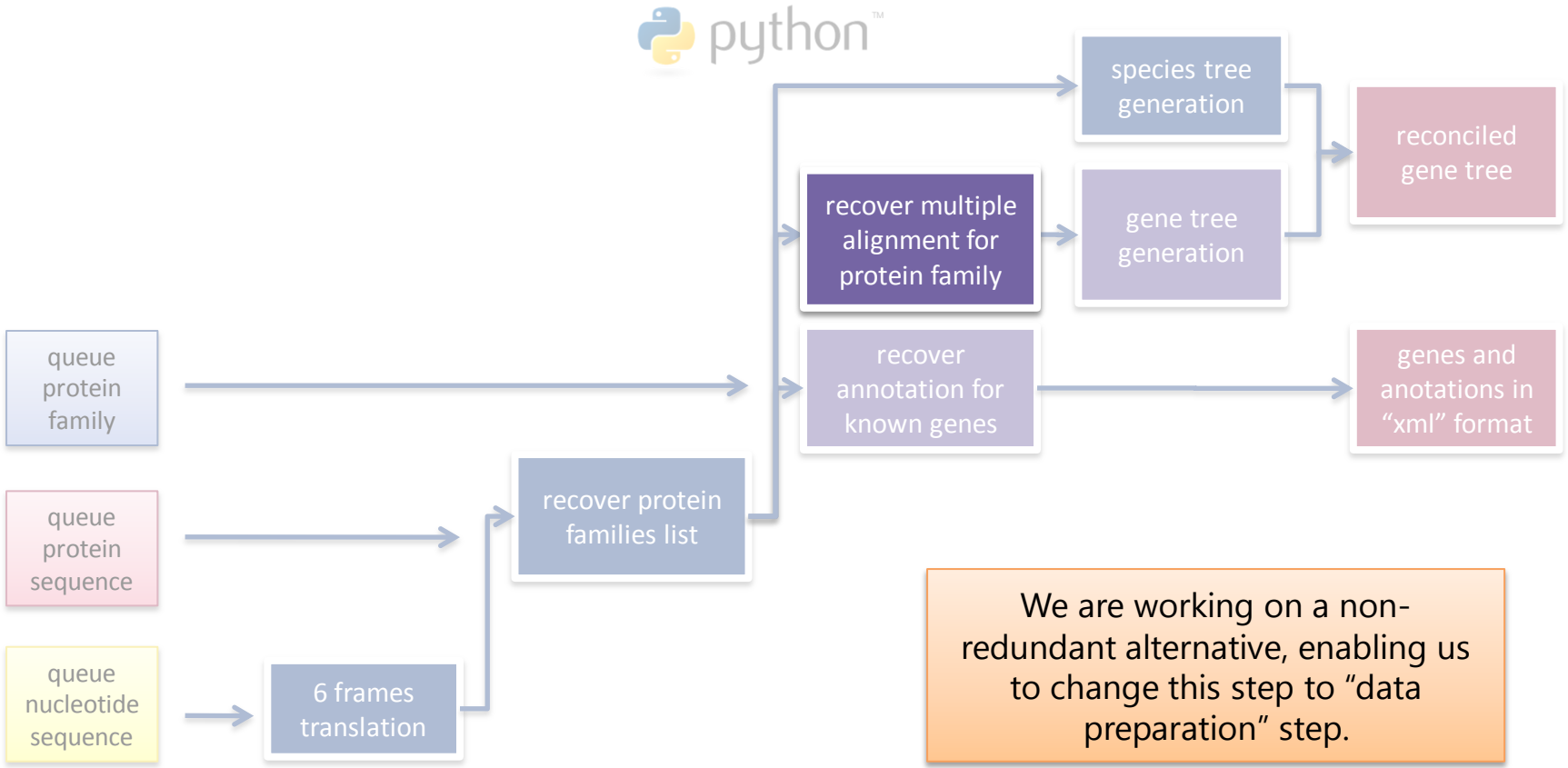
Applicable to whole genomes



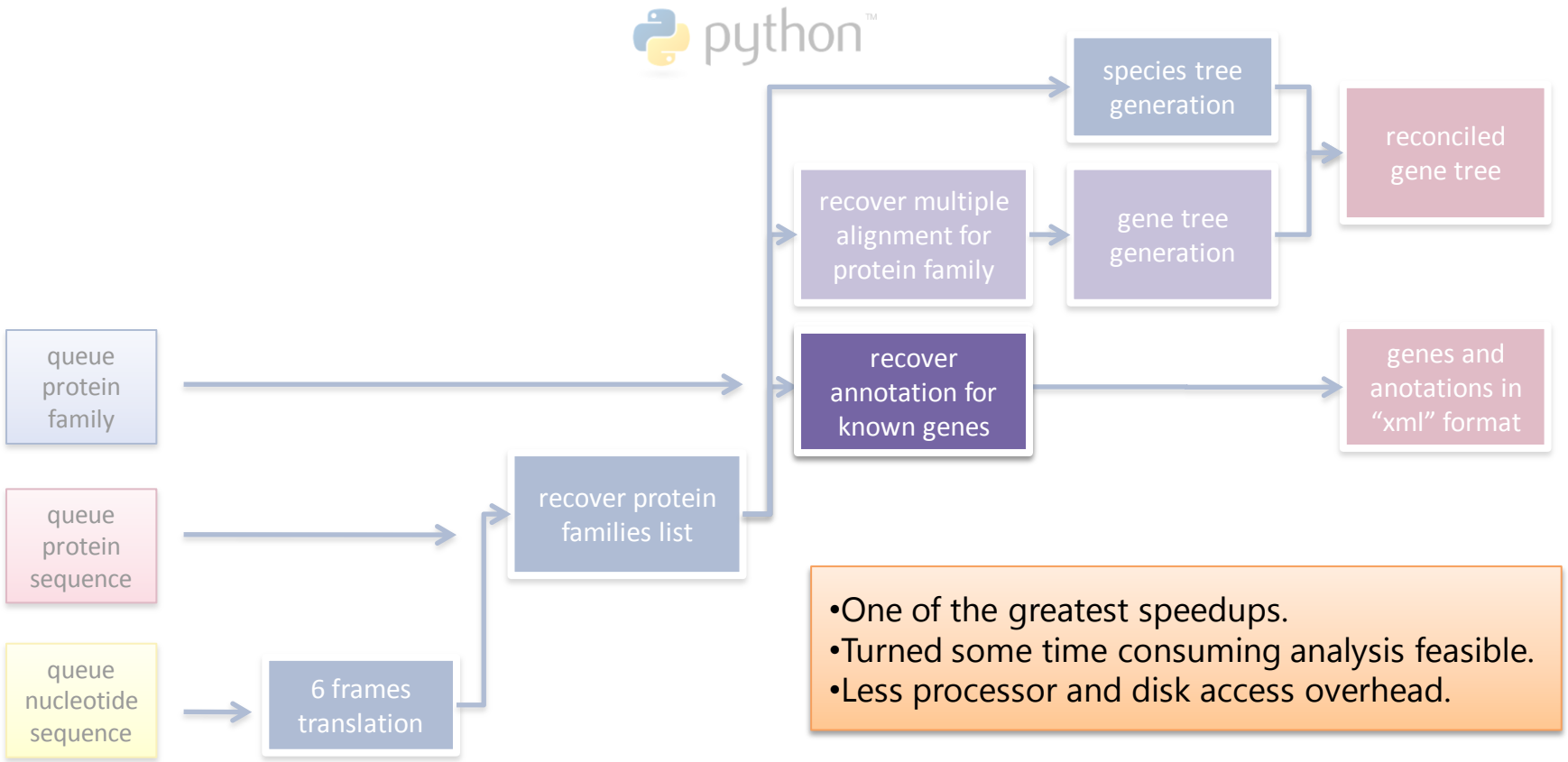
Current Results



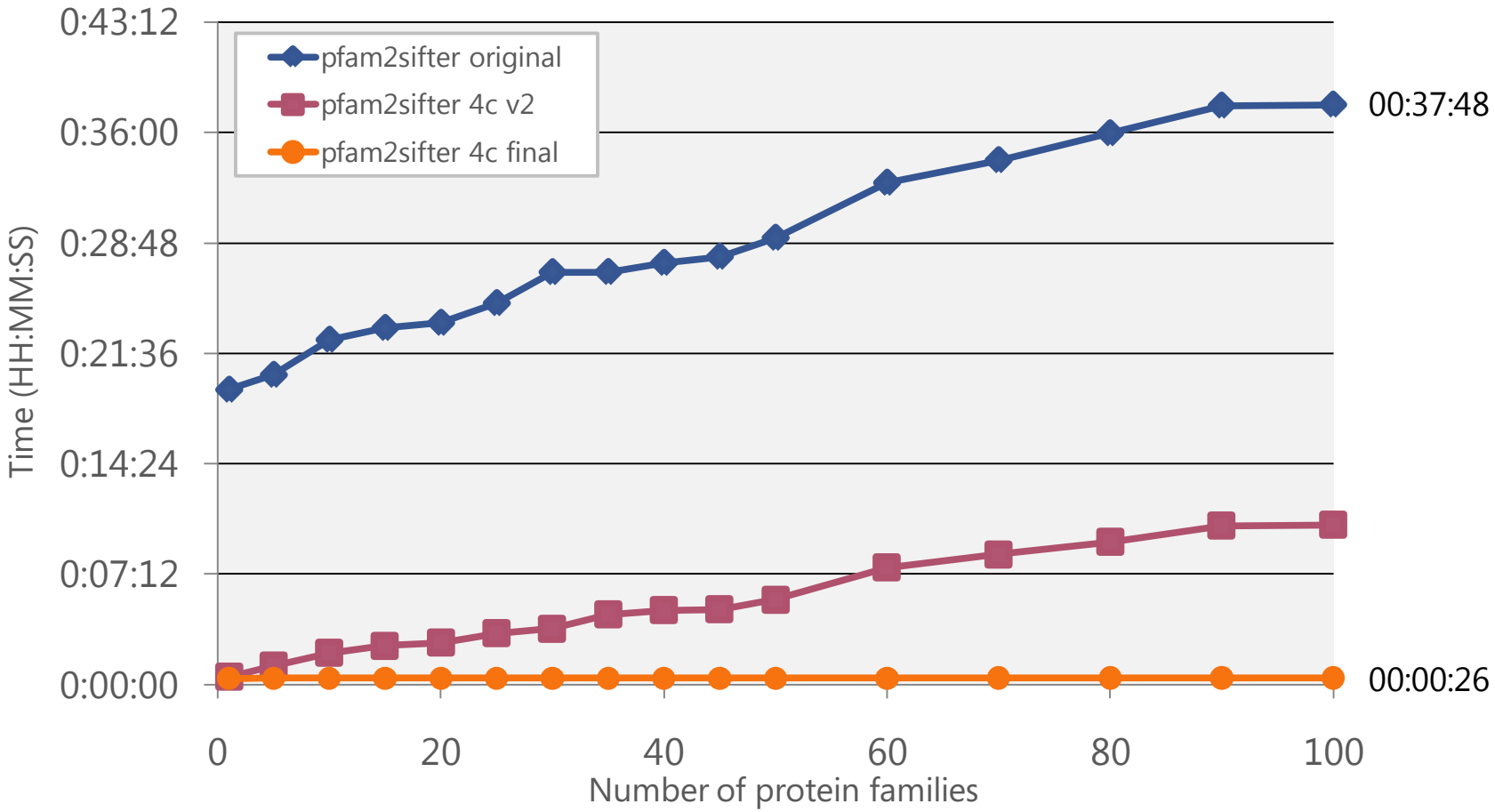
Current Results



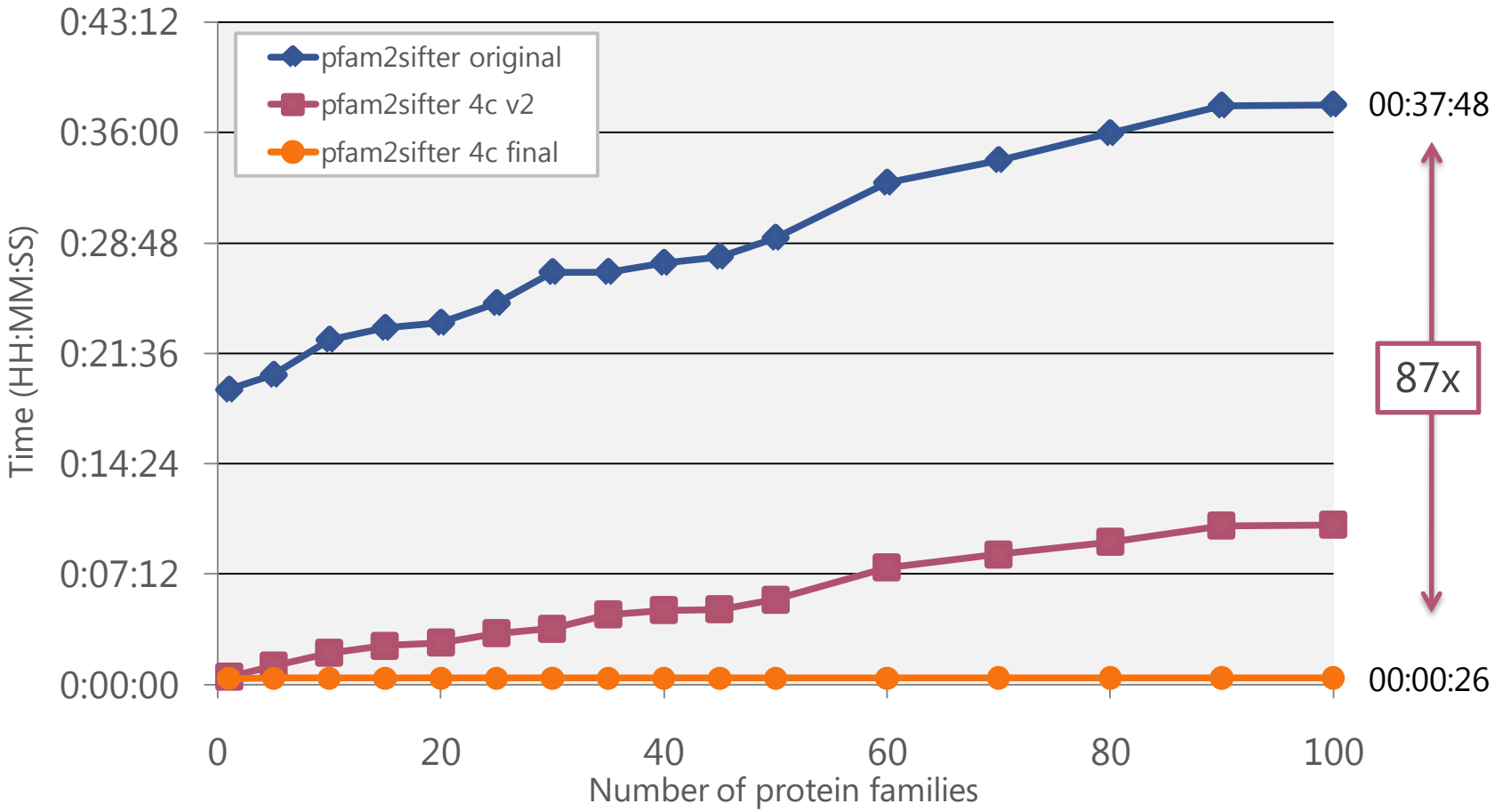
Current Results



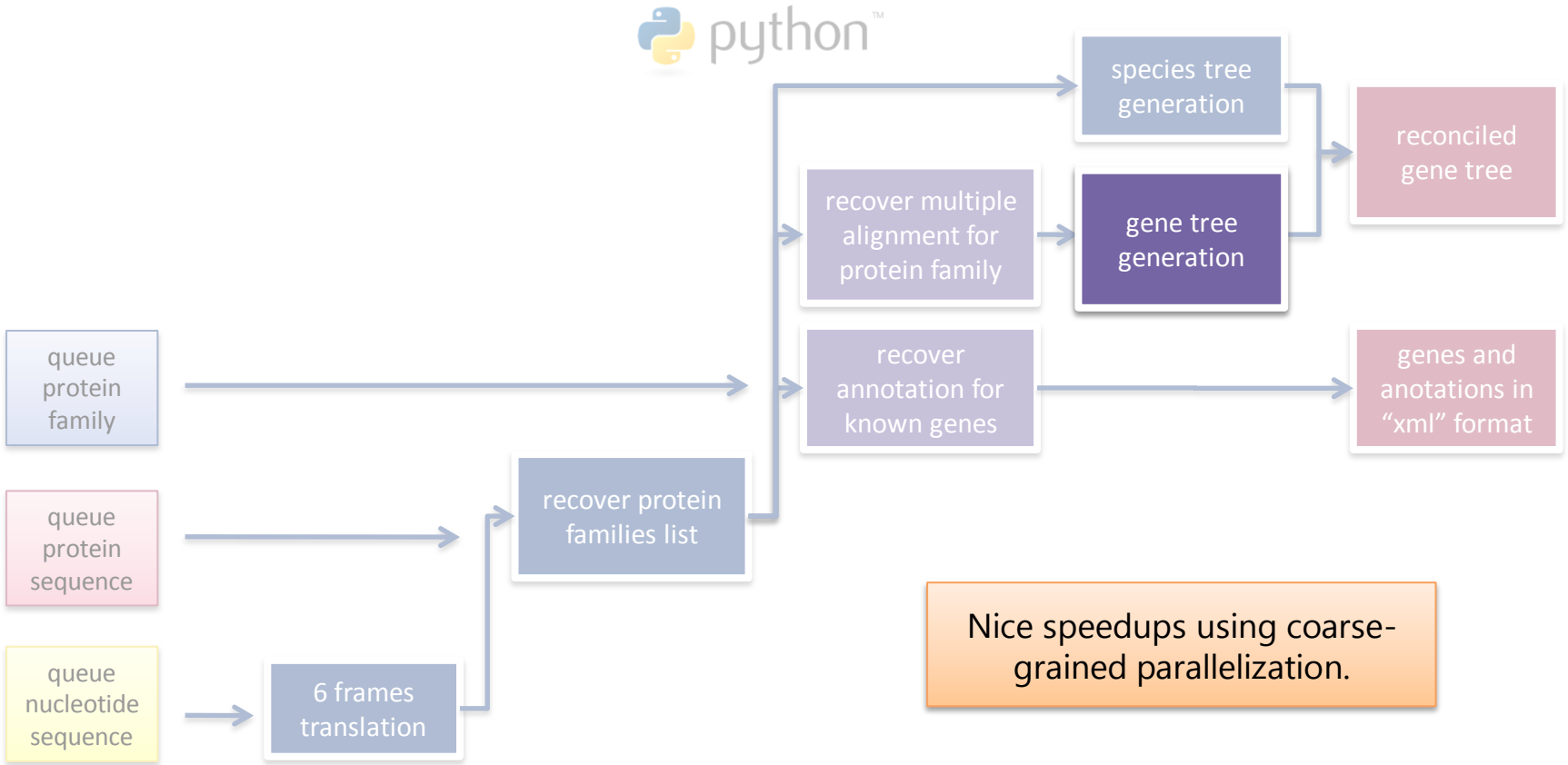
Current Results



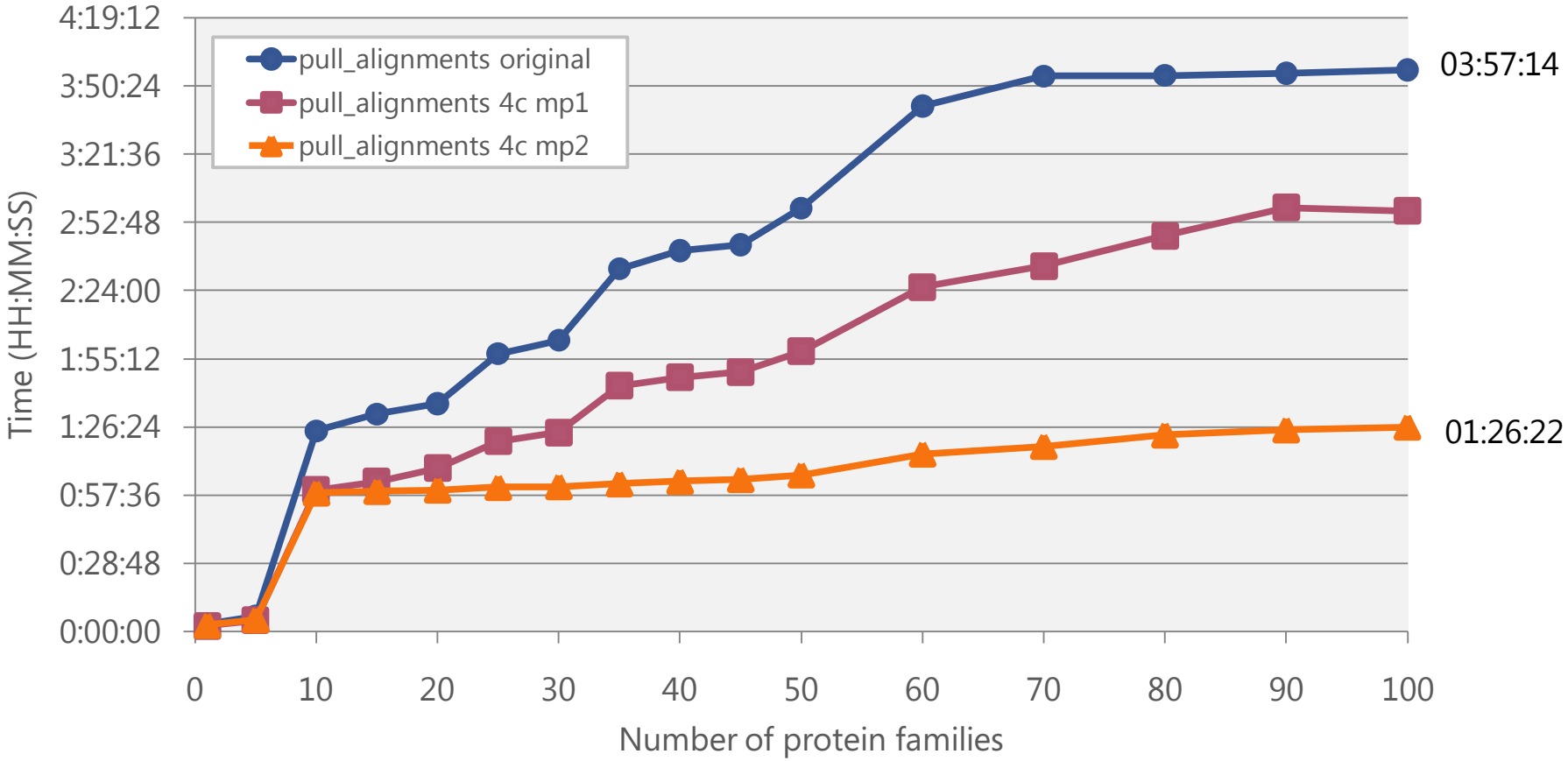
Current Results



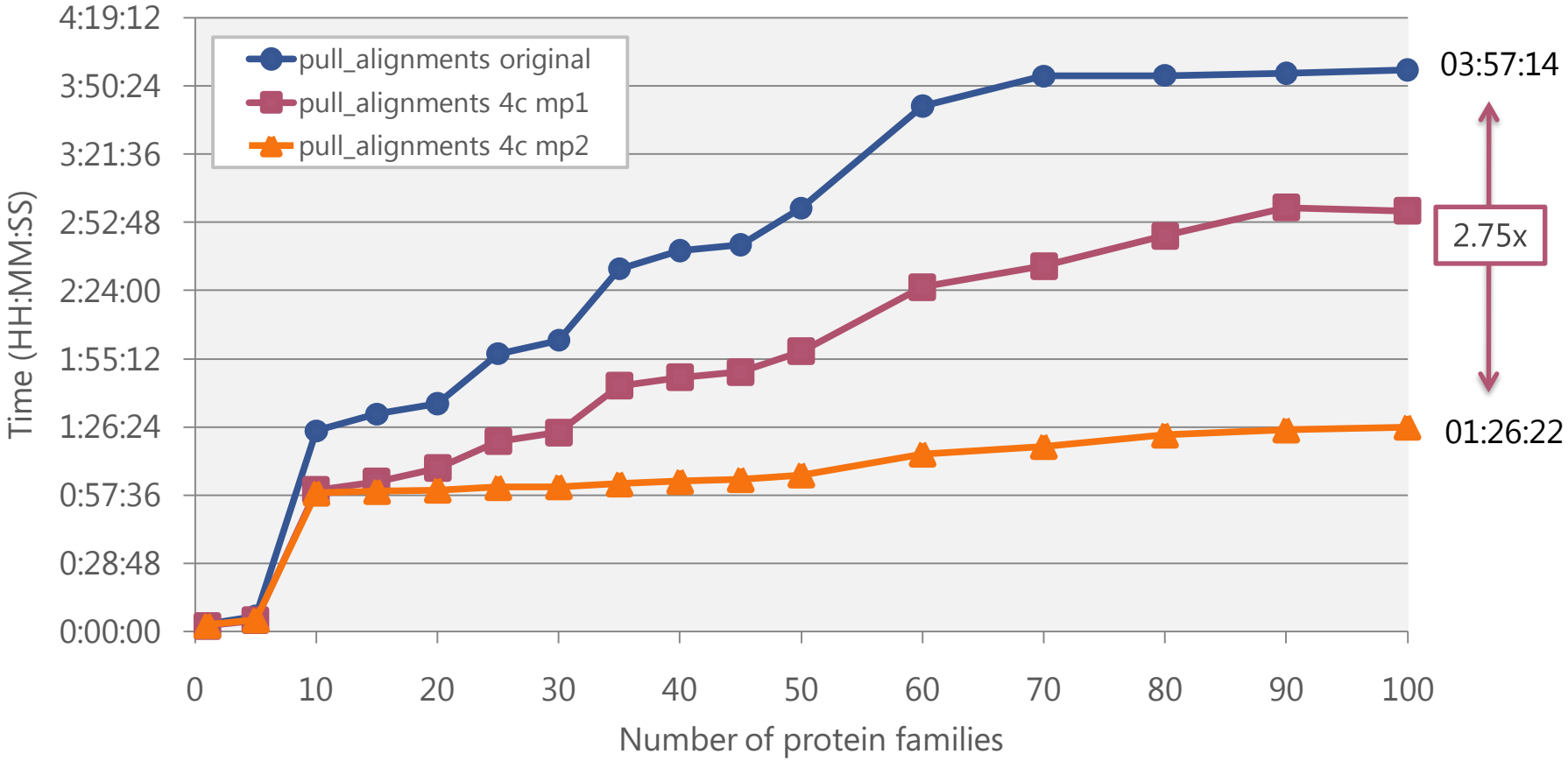
Current Results



Current Results



Current Results



Current Results

Microsoft Research

Search Microsoft Research

Videos | Projects | Publications | People | Downloads

Home | Our Research | Connections | Careers

Worldwide Labs | Research Areas | Research Groups

The Buxton Collection
Explore the history of interactive devices

> Projects > Microsoft Biology Initiative > Microsoft Biology Foundation

Microsoft Biology Foundation

Microsoft Biology Initiative | Microsoft Biology Foundation | Microsoft Biology Tools | MBF Training | MBF Sample Applications

The Microsoft Biology Foundation (MBF) is a language-neutral bioinformatics toolkit built as an extension to the Microsoft .NET Framework to help researchers and scientists work together and explore new discoveries. This open-source platform serves as a library of commonly-used bioinformatics functions. MBF facilitates collaboration and accelerates scientific research by enabling different data sets to communicate. Several universities and corporations use MBF tools, which help reduce processing time and enable scientists to focus on research.

Microsoft Biology Foundation 2.0 Beta 1 Video

Simon Mercer, director in the Microsoft Research Connections group, provides an overview of the improvements and features of the beta release of MBF version 2.0.

Have you used this Research Accelerator?

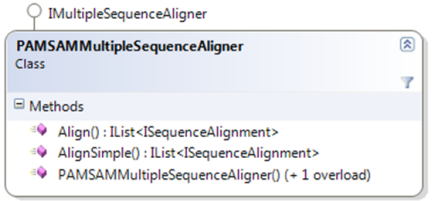
- Take a quick survey

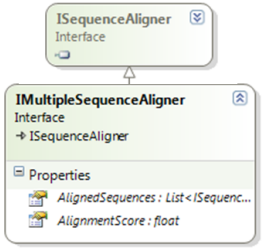
Downloads

- MBF 2.0 Beta 1
- Sequence Assembler 2.0 Beta 1
- Biology Extension for Excel 2.0 Beta 1
- MBF 2.0 Beta 1 Source Code
- MBF 1.0

Multiple alignment sequences

- **`PAMSAMMultipleSequenceAligner` implements optimized multiple sequence alignment algorithm**
  - produces a sequence alignment using three or more sequences
  - based on MUSCLE algorithm
  - uses .NET 4.0 Parallel Extensions to take advantage of multicores
  - included in **`Add-ins\Bio.PamSam.dll`**
- **Algorithm defined by `IMultipleSequenceAligner` interface**
  - defines score as floating point value





Python scripting environment

adapted from:

Microsoft Biology Foundation v2.0 Training Material (Module 06)

BioInformatics

Sequence (re)annotation:

- pathogenic bacteria genome (*Leifsonia xyli xyli*)
- cellulitic fungi genome (*Neurospora crassa*)

useful "sandboxes"

- sugarcane expressed genome (SUCEST 2001)
- sugarcane genome (what is available 2012)

Information Technology:

- automatic probabilistic functional annotation tool
- "technology transfer" from bioenergy to global health

BioInformatics

Sequence (re)annotation:

- pathogenic bacteria genome (*Leifsonia xyli xyli*)
- cellulitic fungi genome (*Neurospora crassa*)

useful "sandboxes"

- sugarcane expressed genome (SUCEST 2001)
- sugarcane genome (what is available 2012)

SUCEST
The Sugar Cane EST Project

Vettore *et al.*, 2001

Information Technology:

- automatic probabilistic functional annotation tool
- "technology transfer" from bioenergy to global health

BioInformatics

Sequence (re)annotation:

- pathogenic bacteria genome (*Leifsonia xyli xyli*)
- cellulitic fungi genome (*Neurospora crassa*)

useful "sandboxes"

- sugarcane expressed genome (SUCEST 2001)
- sugarcane genome (what is available 2012)

SUCEST
The Sugar Cane EST Project

Vettore *et al.*, 2001

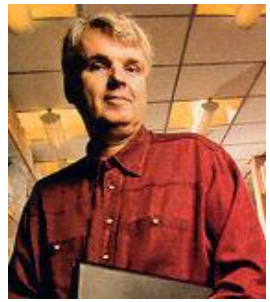
Information Technology:

- automatic probabilistic functional annotation tool
- "technology transfer" from bioenergy to global health

Easy to use
Lightweight
Broader public

Partnership

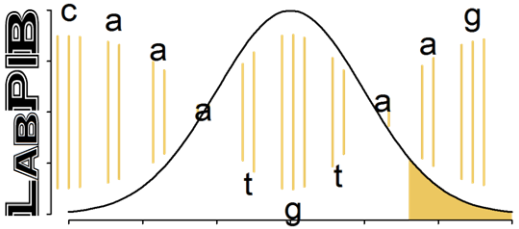
- Microsoft Research–FAPESP Institute for IT Research
  - Dr. David Heckerman
  - http://research.microsoft.com/



- FAPESP Bioenergy Program
  - Prof. Dr. Gláucia Mendes Souza
    - Program Coordinator
    - Project: *Sugarcane Signaling And Regulatory Networks*
    - http://bioenfapesp.org/

Special Thanks

- Laboratory for Biological Information Processing - **LABPIB**
  - Prof. Dr. Ricardo Z.N. Vêncio
  - Undergrad Ricardo Cacheta Waldemarin



http://labpib.fmrp.usp.br

Microsoft Research
FacultySummit

Thank you for your attention!

Danillo C. Almeida-e-Silva, grad student
Department of Computing and Mathematics - FFCLRP
University of Sao Paulo – Brazil
Contact: danillosilva@usp.br

FUTURE/WORLD 2011 2031

Microsoft Research
FacultySummit

FUTURE/WORLD 2011 2031