THE BIOVEL PROJECT:
ROBUST PHYLOGENETIC WORKFLOWS RUNNING ON THE GRID

www.biovel.eu
BioVeL is a virtual e-laboratory that supports research on Biodiversity issues using large amounts of data from cross-disciplinary sources. It is a consortium of 15 partners from 9 countries, as well as an outer circle of ‘Friends of BioVeL’

- Access a worldwide network of expert scientists
- Sharing knowledge on Biodiversity research

Biodiversity Issues

- Species identification, discovery and distributions
- The changing nature of ecosystems altering organismal composition
- The increased risks of species extinction

Decision making in biodiversity management at multiple scales (genomic, organismal, habitat, ecosystem, landscape, etc...
Biodiversity Solutions

✓ **Services:** data processing techniques. Each technique is available as a single executable application which can be used either alone or within a workflow builder environment (e.g. Taverna)

✓ **Workflows:** examples of services use that can be modified

---

**Services and Workflows for Biodiversity Analysis:**

- Taxonomy
- **Phylogenetics**
- Metagenomics
- Ecological Niche Modeling
- Ecosystem Functioning and Valuation
- Geospatial Visualization
Example of Phylogenetic Services

Workflow View
- MultiFasta
  - Unaligned DNA protein coding
  - Nexus/NEXML quality Trimmed Aligned DNA
- Alignment
- Translate, Align to Selected PFAM Domain and back Align DNA
- WebServices View
  - Get PFAM Domain, frame and genetic codes
  - Translate DNA2Prot
  - HMMsearch
  - PFAM-A Mirror
  - HmmAlign
  - BackAlign DNA on Prot
  - Hmmer Domain Parser
  - Translator DNA/Protein
- Computation View
  - Job Submission Tool
  - Application DB
  - INFN WebService
  - Local farm
  - EGI farms (VO Biomed)
  - WebDav Server
  - User PC

Phylogeny Inference
- User
- GUI to select Evolutionary model
- Tree Posterior Distribution
- Phylogenetic inference
- Verify MCMC Convergence
- Goodness of Fit model
- Visualize Tree
- MrBayes
- GeoKS
- HyPhy
- PhyloWidget WebPage
Job Submission Tool: JST

Frontend:
- Username
- Task status
- Dependencies of each task
- Priority
- Job provenance
- Task description
- Number of failures
- Date and time of execution
- Infrastructure information (grid, local farm, interactive server)

Backend:
- Task submission at a given rate
- Stops jobs submission when no more unassigned tasks are found in the TaskList
Multiple Sequence Alignment Workflow

Workflow View

- MultiFasta
  - Unaligned DNA protein coding

- Nexus/NEXML
  - quality
  - Trimmed
  - Aligned DNA

- Aligned DNA protein coding
  - site specific quality scores
  - Stockholm

WebServices View

- Get PFAM Domain, frame and genetic codes
- Translate, Align to Selected PFAM Domain and back Align DNA

- Translate DNA2Prot
- HMMsearch
- PFAM-A Mirror
- Hmmer Domain Parser
- Translator DNA/Protein
- HmmAlign
- BackAlign DNA on Prot
Multiple Sequence Alignment Workflow

Multiple Alignment of DNA coding

Higher alignment precision given by:

- HMM search assigning a per site quality score (posterior probability)
- Back-align (amino acid -> DNA)

In progress:
Multiple Domain Coding sequences Alignment
Example Phylogenetic Inference Workflow

User

GUI to select Evolutionary model

Workflow View

Tree Posterior Distribution

Phylogenetic inference

Verify MCMC Convergence

Goodness of Fit model

Visualize Tree

WebServices View

MrBayes

GeoKS

HyPhy

PhyloWidget WebPage
Example Phylogenetic Inference Workflow

MrBayes partitioned models

**Partition 1**
- **Partition name:** Partition_1
- **Start:** 1, **End:** 400, **Lag:** 1
- **Evolutionary model:**
  - **Model for partition 1**
  - **Transition matrix:** GTR
  - **Rates variability:** gamma distribution
- **Markovian Integration Parameters:**
  - **Number of generation:** 1000000
  - **Frequency of sampling:** 100
  - **Number of Chain:** 4
  - **Temperature:** 0.2
  - **Output files prefix:**
  - **Number of Replicates:** 2
- **Matrix:** Not Grouped

**Partition 2**
- **Evolutionary model:**
  - **Model for partition 2**
  - **Transition matrix:** HKY
  - **Rates variability:** equal
- **Lset:**

**Link parameters across partitions?**

**Matrix**
- **Partition 1 (GTR):**
- **Partition 2 (HKY):**
Example Phylogenetic Inference Workflow in Taverna

MrBayes Web Interface

Bayesian Phylogeny Computation & Output Retrieval

GeoKS Execution

Consensus Tree Calculation

Tree Visualization

Peculiarity:
- Partitioned models
- Convergence calculation
- Short Computation time on the Grid (even for long jobs)

Other available Phylogenetic Services:
- Maximum Likelihood (RaxML)
- Phylogenetic Diversity (Phylocom)
Bioinformatic Scientists
Prof. Graziano Pesole
Dr. Saverio Vicario

ICT specialists
Dr. Giacinto DONVITO
Dr. Pasquale NOTARANGELO

Funding:
European Commission 7th Framework Programme (FP7), through the grant agreement: 283359