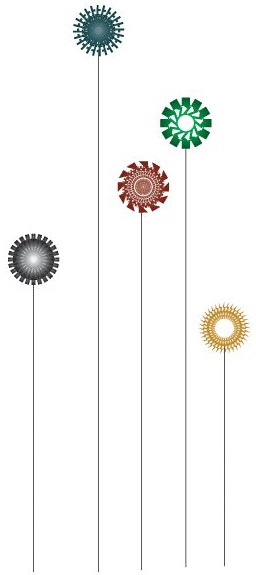
Unipro

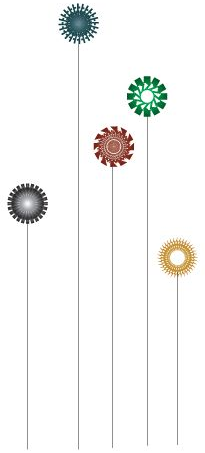
Bioinformatics workflows in the cloud with Unipro UGENE

Mikhail Fursov
Konstantin Okonechnikov
mfursov@unipro.ru
kokonech@unipro.ru



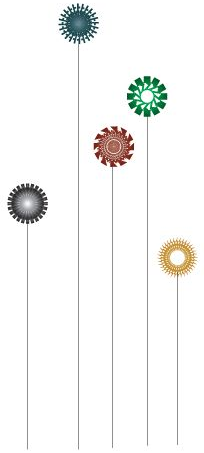
About this tutorial

- Intended audience
  - You are
    - Molecular biologist who uses bioinformatics tools in research
    - System administrator who looks for new software options for an organization
    - Martian secret service agent



Agenda

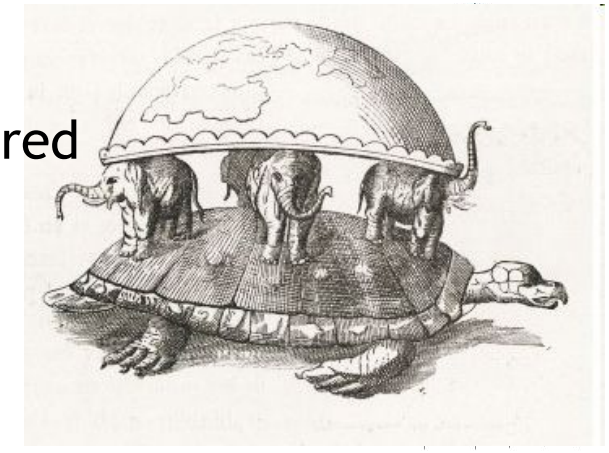
- Bioinformatics today
- Task complexity factors
- Computational workflows - overview
- Unipro UGENE project: features and capabilities
- UGENE Workflow Designer
- Cloud computing basics
- Running workflows in the cloud
- Free discussion



Unipro

Bioinformatics today

- **Bioinformatics 15 years ago**
  - Pharmaceutical companies were not interested;
  - Life scientists believed that it was an outlet for failed biologists that want to play around with computers;
  - Computer scientists did not even consider it important, they confused it with bio-inspired “computer sciences”.



Unipro

Bioinformatics today

- Bioinformatics in 2002

  - Pharmaceutical companies believe that it is the most efficient way to streamline the process of drug discovery;

  - Some life scientists believe it is the solution to all problems in life sciences and that it will allow them to avoid doing some experiments;

  - Computer scientists are very interested. The scope and complexity of the domain makes it the ideal field of application of new software techniques and specialized hardware developments.

Bioinformatics today

- Bioinformatics in 2010
  - Pharmaceutical companies use it routinely, but have realized that it complements rather than replaces experimental work;
  - Computer scientists may have jumped on another fancy subject.
  - The United Nations has designated 2010 the International Year of Biodiversity
  - ….
  - Life scientists use it every day and try to learn on how to deal with >9000 tools and methods available

Unipro

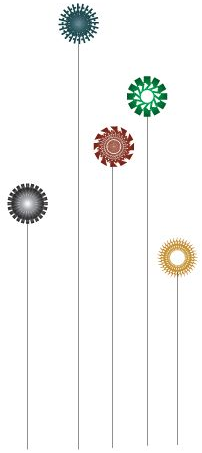
Bioinformatics today

Biological problems that computers can help with:

- I cloned a gene –is it a known gene?
- Does the sequence match? Is the sequence any good?
- Is the sequence similar to other known sequences?
- Which gene family does it belong to?
- The gene I´m interested in was found in another organism, but not in mine. How can I look for it?
- How is the gene expressed in different types of tissues?
- What is the biological function of the protein encoded by the gene?
- Is the gene associated with any disease?

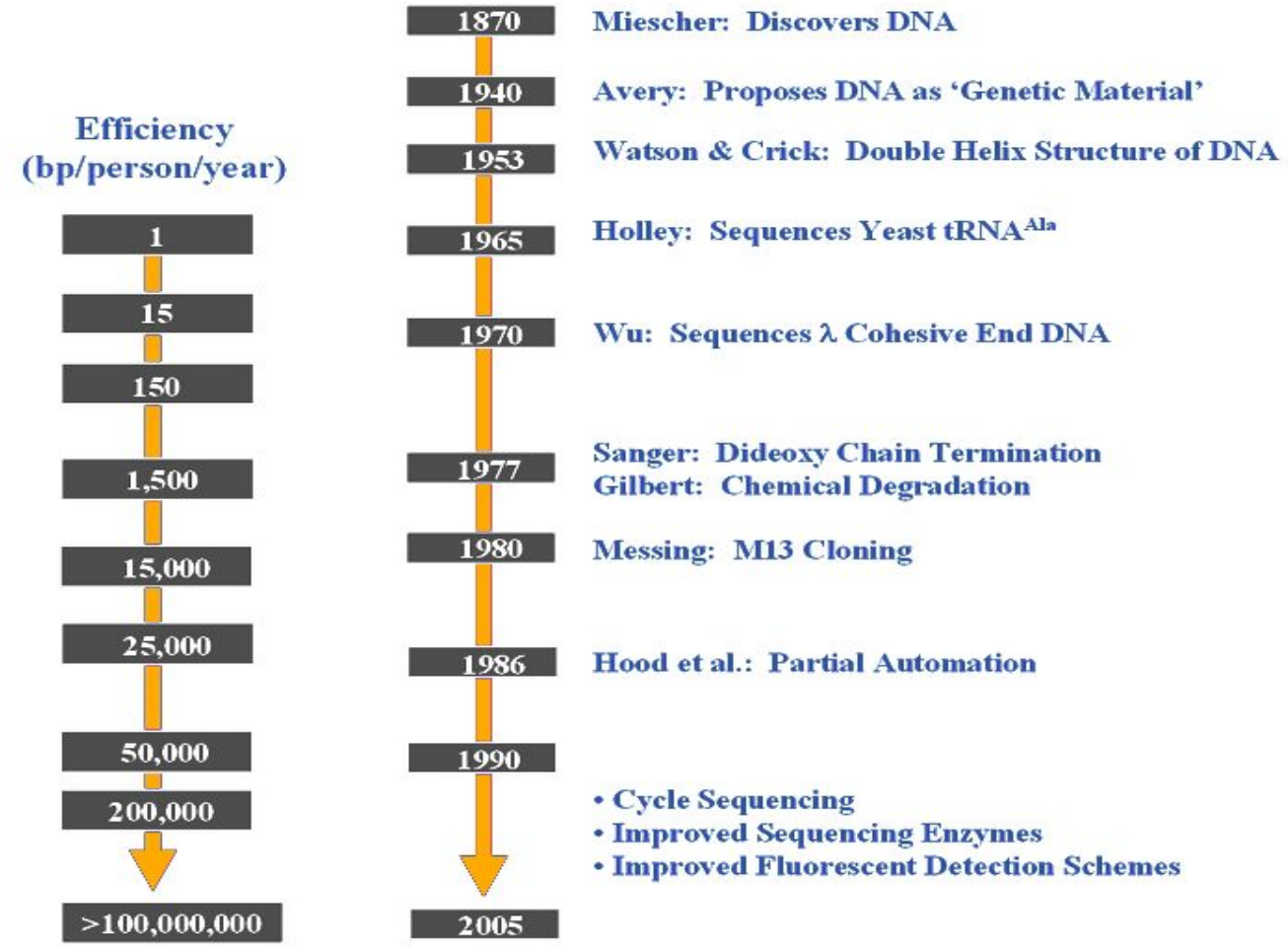
What computer science is to molecular biology is like what mathematics has been to physics

Larry Hunter, ISMB'94



Unipro

Task complexity factors
Data Volumes



Adapted from Messing & Llaca, PNAS (1998)

Unipro

Task complexity factors
Number of databases and protocols

Some words about DNA databases:

unipro

Task complexity factors
Number of available methods, tools, data formats

BLAST,FASTA,SSEARCH,CLUSTAL,MUSCLE,MAFFT,KALIGN,UCLUST,HMMER2,
HMMER3,GARLIC,CONSED,CGVIEW,ERGO,EBBIE,MAUVE,MATTREE,COVE,
PSIBLAST,GOR,PSIPRED,EXPASY,EMBOSS,PHYLIPP,SAM,CASP,BLOCKS,PRIMER3,CSBLAST,HHPRED,BIOCONDUCTOR,M
UMMER,FEAST,BOWTIE,MAQ,SOAP,BIOPERL,POA, PRANK, FOLDALIGN,RMAP,SITECON,SHRIMP,BATWING,ASAP,
BEAST,MEGA,MESQUITE,SEMPHY,TNT,BIOEDIT,BIOPYTHON,GALAXY, TAVERNA, GENEMARK,
AMAP,MEME,PPSEARCH,ELPH,GENESCAN,ARTEMIS,CLANN,GENLUX,CRNPRED,BRAGI,DIP4FISH
ANGIS,AFFYMETRIX,GENECHIP,ARLEQUIN,BIOPHP,BIORUBY,BIOEXTRACT,BIOSLAX,BISKIT,CYTOSCAPE,DAVID,DIALIGN
-T,DIALIGN-TX,DNASTAR,ETBLAST FOLDX,FORMATDB,GENSCAN,GENTLE,GESS,GENMAPP,GENE,ACE, UGENE,ARGO,
DESIGNER,GENEDATA,ENEPATTERN,GENEVESTIGATOR,JALIGNER,MEGAN,ARKA
MODELLER,OLIGO,JPRED,STRIDE,TESS,GLIMMER,BIOECLIPSE,ENSEMBL,ASTERIAS,DPVIEW,
PAUP,PSORT,PHYLOSCAN,PUPASUITE,PYMOL,RAPTOR,RASMOL,STING,SIMBIOSYS,SNAGGER,SOAPLAB,SPLITSTREE,ST
EMLOC,T-COFFEE,PILER,USEARCH,DELTASTAT,DCSE,ASID,ARB,ANGLER,
TREEFINDER,UCSF CHIMERA,UTOPIA,VECTOR NTI,YASS,MUSCA,JASPAR

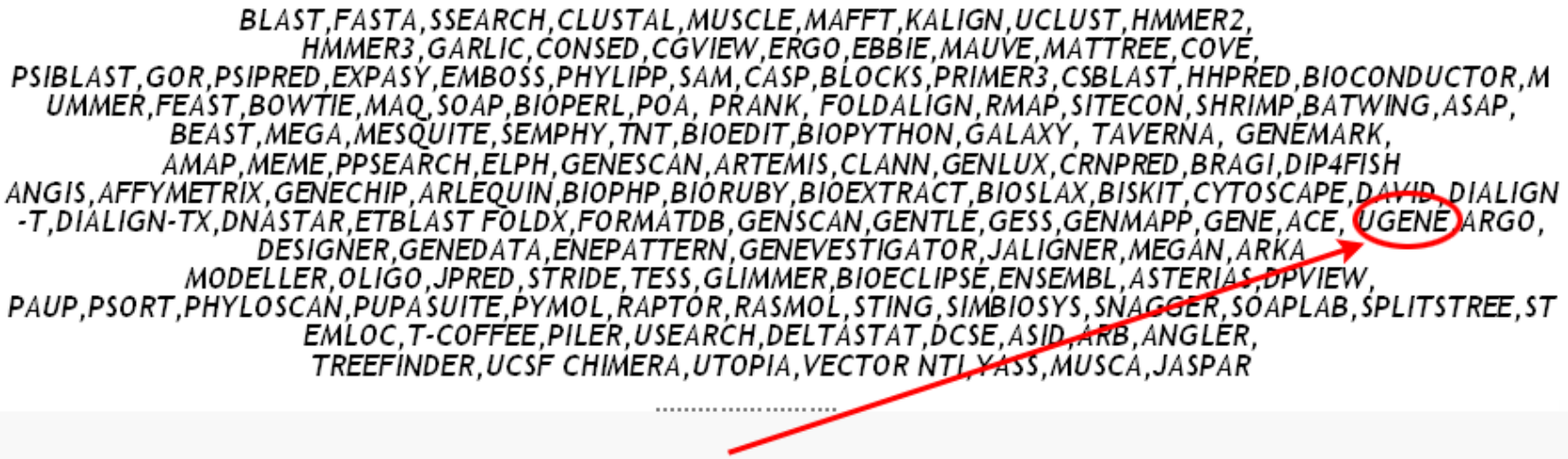
……

+ 9000 more (from Wikipedia)

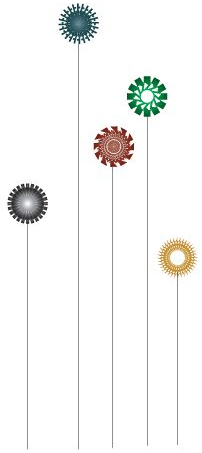
- **Facts:**
  - A part of these tools use incompatible data formats
  - A part of these tools are not available for your computer platform
  - A part of these tools are too slow to be used effectively
  - A part of these tools cost money and have not free license
  - A part of these tools has critical bugs that can affect your results
  - A part of these tools lacks documentation
  - A part of these tools are not supported nor developed anymore
  - ...
  - You have no idea what most of these tools about or how to use them …

Unipro

Unipro UGENE project about



- UGENE is one of a number of bioinformatics projects
- The goal of UGENE project is to integrate all other popular bioinformatics tools within a
  - Single application
  - Unified data model
  - Powerful user interface
    - » Do all above in open source



Unipro

UGENE capabilities (v1.7.2)

Algorithms:

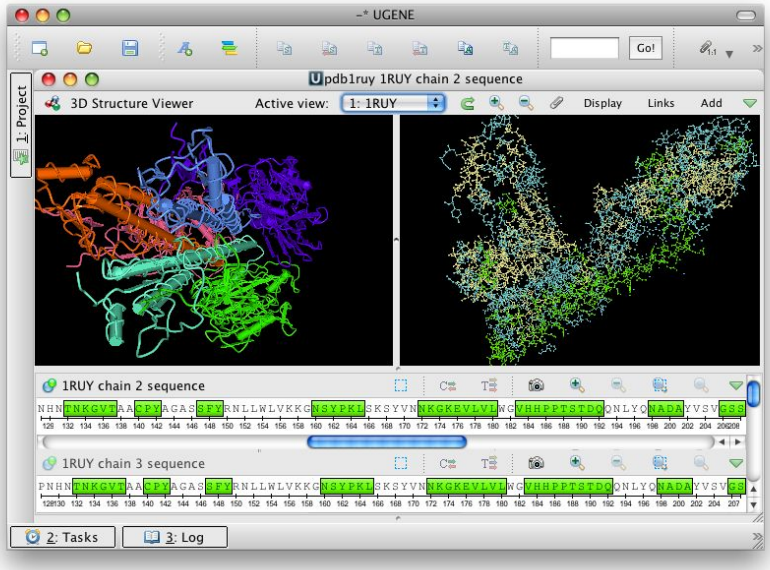
- Multiple sequence alignment using MUSCLE 3,4, KAlign, ClustalW, MAFFT;
- HMM profiles search, based on the source of HMMER2 and HMMER3 tools;
- PCR Primers design using Primer3;
- Protein secondary structure prediction using GORIV and PSIPRED;
- Phylogenetic analysis with Phylip;
- Search for restriction enzymes and integration with REBASE;
- Extremely fast repeat finder;
- DNA to reference assembly using Bowtie and own algorithms;
- Search for transcription factor binding sites using collection of weight and frequency matrix based algorithms (SITECON, JASPAR, UniProbe);
- Various statistical reports: GC content/deviation, I.Entropy, Karlin …
- Classic algorithms: ORFs, Smith-Waterman, protein back-translation;
- Comparing whole genomes using dotplot viewer;

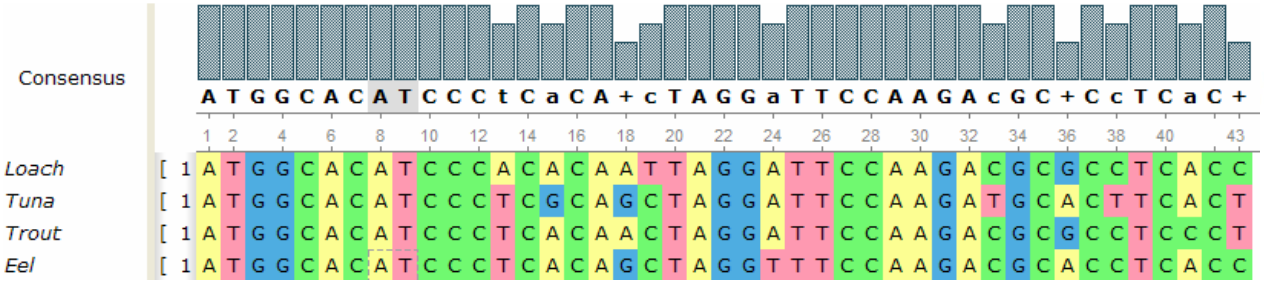
» *more*

UGENE capabilities (v1.7.2)

Visualization:

- DNA/Protein sequence & annotations
- Multiple sequence alignments
- 3D structures and surfaces
- Graphs and statistics
- Phylogenetic trees
- Chromatograms
- Plasmid maps
- Dot-plots





Unipro

UGENE capabilities (v1.7.2)

Data formats

- Genbank
- EMBL
- CLUSTAL
- MSF
- STOCKHOLM
- FASTA
- FASTQ
- NEWICK
- NEXUS
- ABI
- SCF
- HMMER2/3
- MMDB
- PDB
- GFF
- SAM
- CSV
- Plain text

- UGENE supports import and export data between all these formats
- All of these formats are auto-detected automatically
- User can work with files stored in gzipped (compressed) form
- Data files can be located on remote servers
- User is provided with powerful scripting language to extend the supported formats list

Unipro

UGENE capabilities (v1.7.2)

High performance computing

- UGENE is created to utilize effectively computational power as well as small workstations (P1 MMX, 64Mb RAM) as large data centers

- For the algorithms included UGENE has special optimizations for:
  - Multi-core CPUs
  - SIMD instruction sets (SSE,SSE2,AltiVEC)
  - CELL CPU architecture
  - NVIDIA GPUs with CUDA library
  - ATI GPUs with ATISTREAM library
  - Direct x86/x86_64 assembler snippets
  - Distributed computations

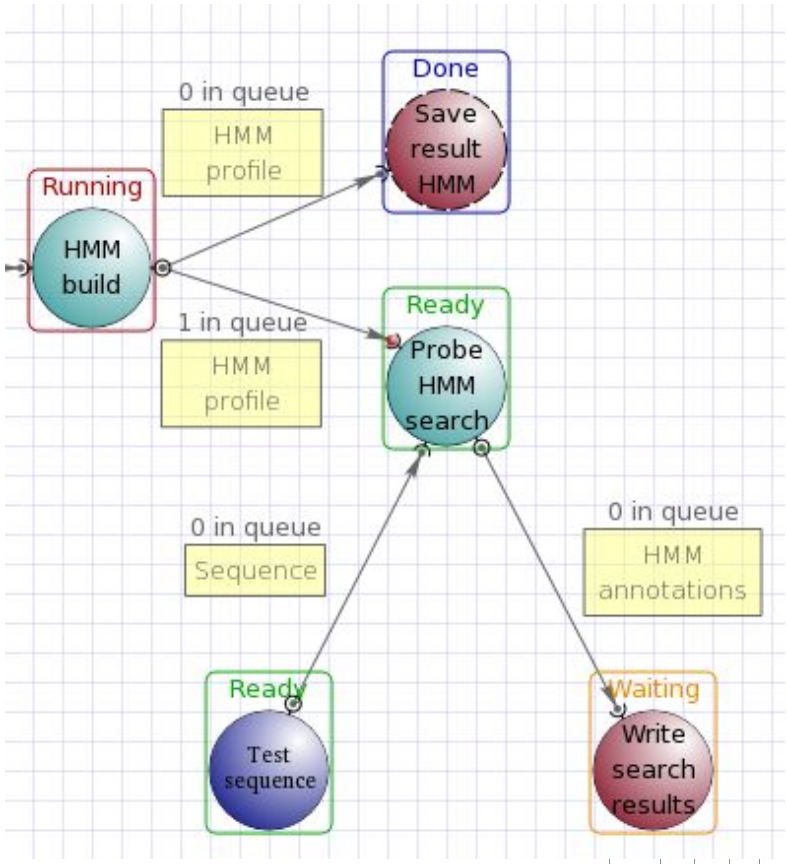
Unipro

UGENE capabilities (v1.7.2)

Joining all together

- Task complexity
- Rich algorithm libraries
- Unified data formats
- High performance
- Powerful user interface

Workflow Designer



Unipro

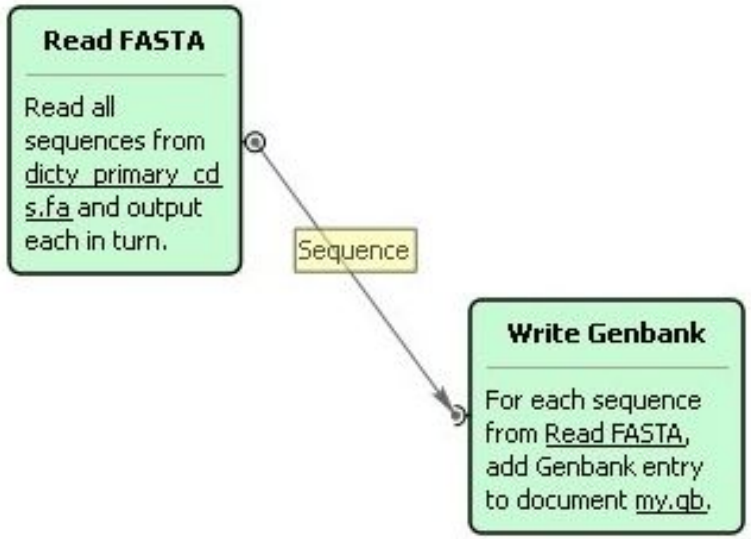
UGENE Workflow Designer

- Goal
  - to make complex things simple
- Key feature
  - Simplicity
- How?
  - See next slide

Unipro

UGENE Workflow Designer

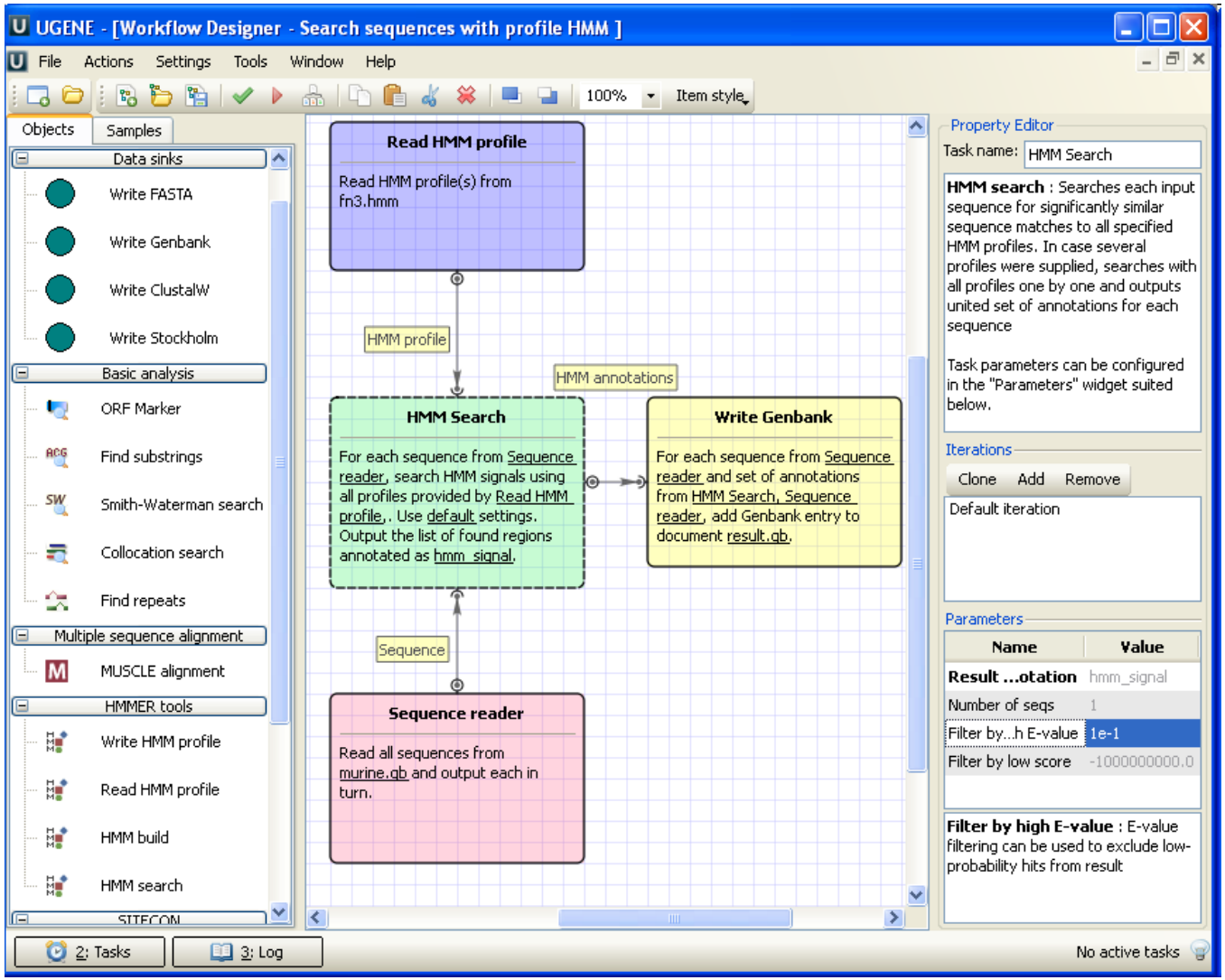
- A complete workflow sample
  (*a screenshot of User Interface)*



- Q: What will this workflow do?

Unipro

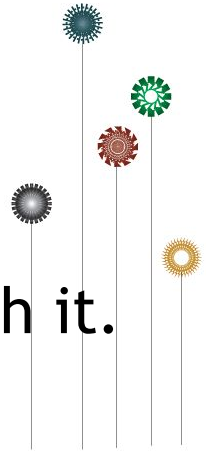
UGENE Workflow Designer layout



Unipro

UGENE Workflow Designer basics

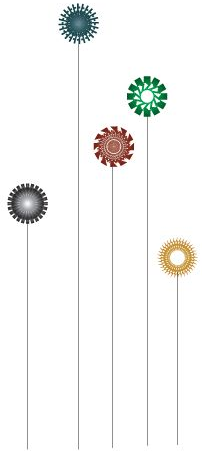
- Using drag&drop interface construct computational diagram from a set of algorithmic blocks or processes.
- Processes can be connected with each other using data-flow channels if they have input and output ports of the same data types.
- Each process in a diagram has a set of configurable parameters, specific to the logic of the algorithm it represents.
- Validate the workflow, correct any errors and launch it.



Unipro

UGENE Workflow Designer practice

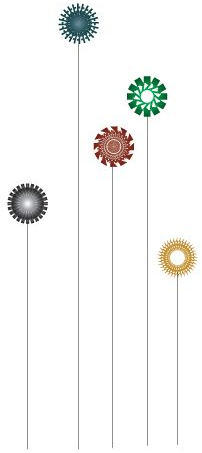
- Lesson 1 – Data format conversion
  - Goal: extract sequences from PDB files



Unipro

UGENE Workflow Designer practice

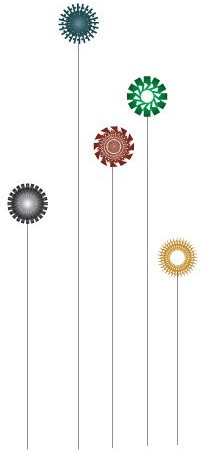
- Lesson 2 – Multiple sequence alignment
  - Align multiple files and save the result into desired file format



Unipro

UGENE Workflow Designer practice

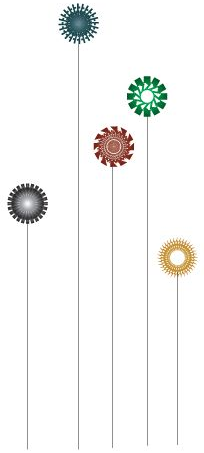
- Lesson 3 – Assemble a sequence
  - Run DNA assembly with Bowtie



Unipro

UGENE Workflow Designer reuse

- Run workflow multiple times
  - Run the same workflow multiple times for different data sets



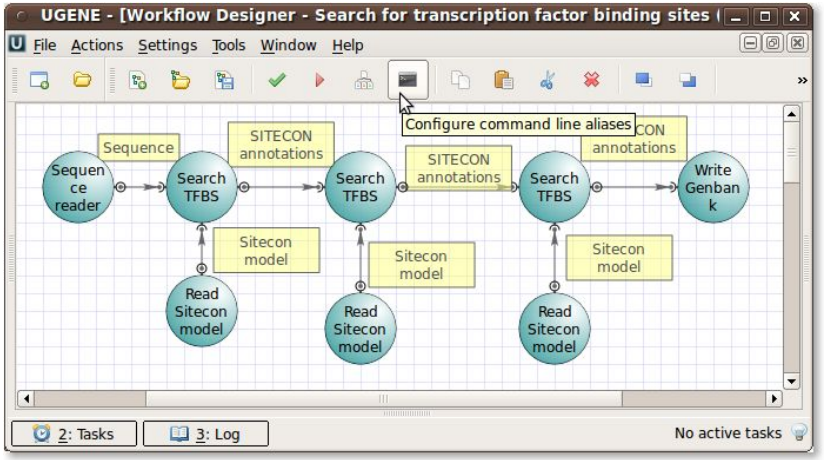
Unipro

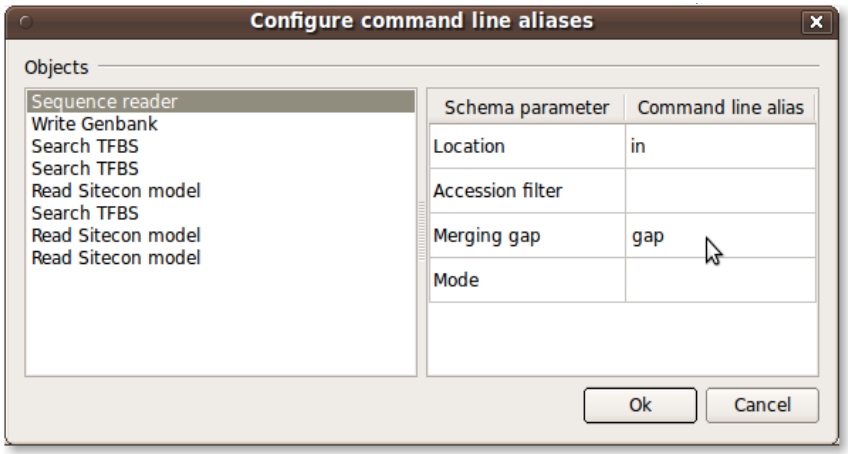
UGENE Workflow Designer reuse

- Create new shell command from workflow
  - Use your own workflow as a stand alone command line tool
  - Example:

```
ugene align --in=file1.aln --out=file2.ali
```

  - Where
    - ‘align’ is the name of the workflow
    - ‘--in’ and ‘--out’ are cmd-line aliases for workflow parameters

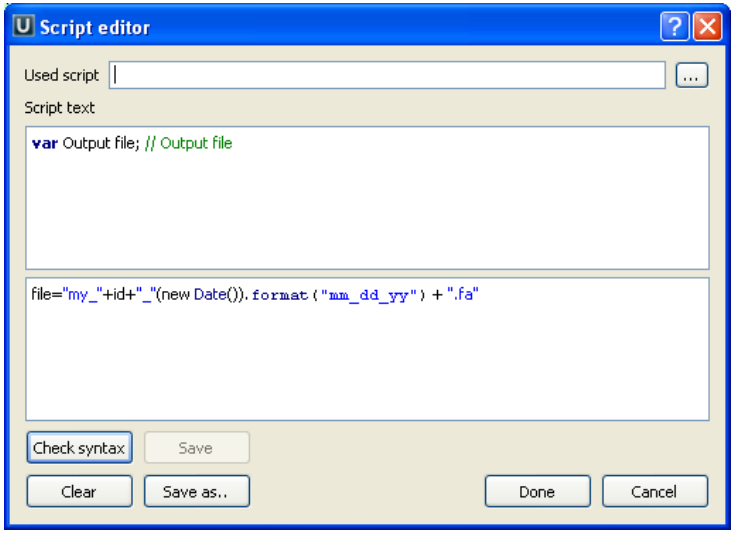


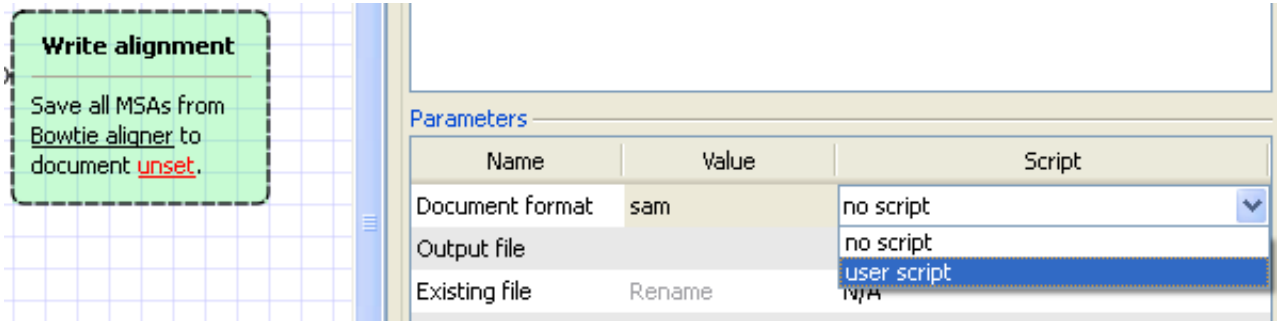


Unipro

UGENE Workflow Designer reuse

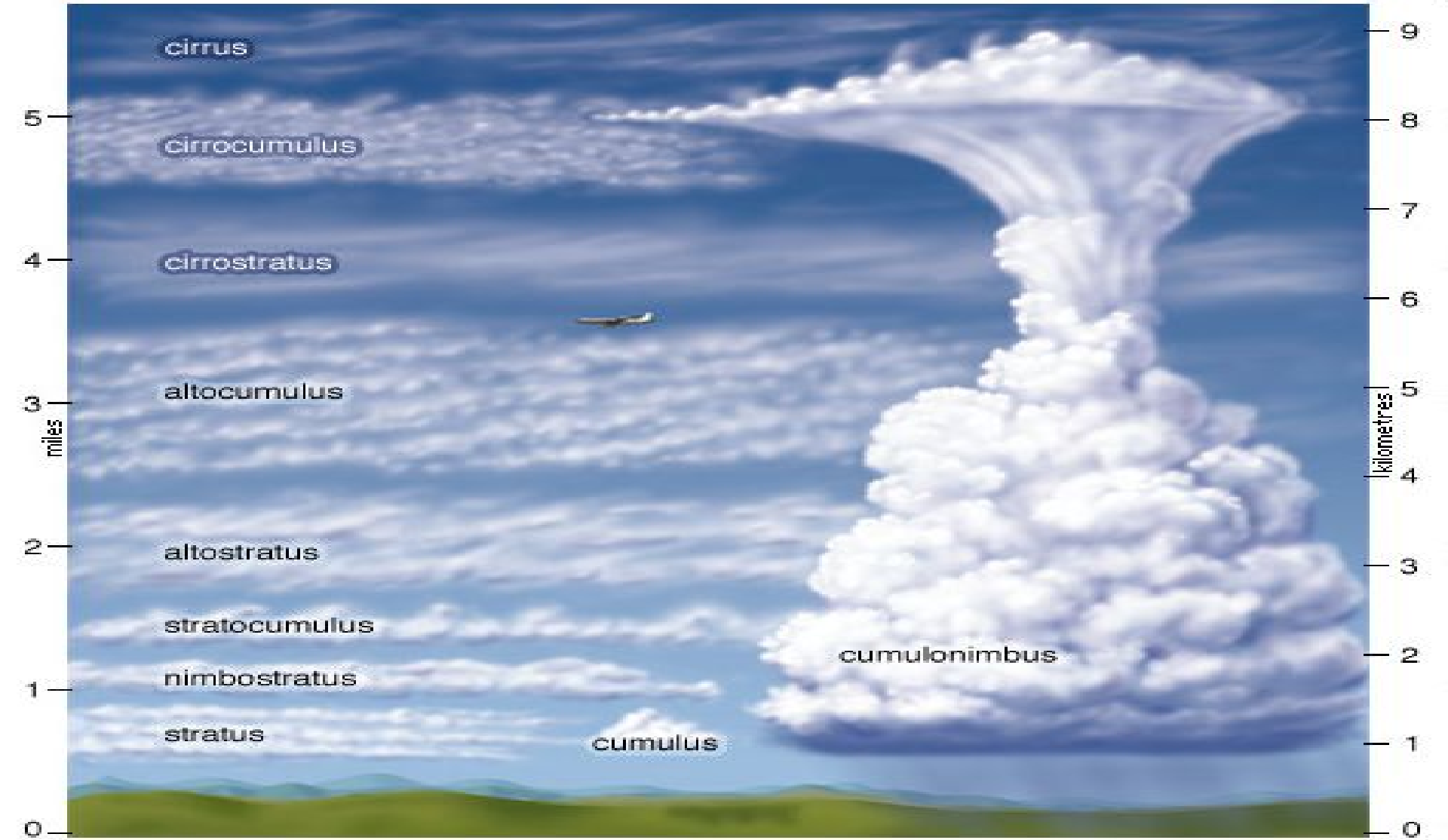
- Script new features
  - use embedded scripting language to design new workflow building blocks





Unipro

Part 2.
Cloud Computing

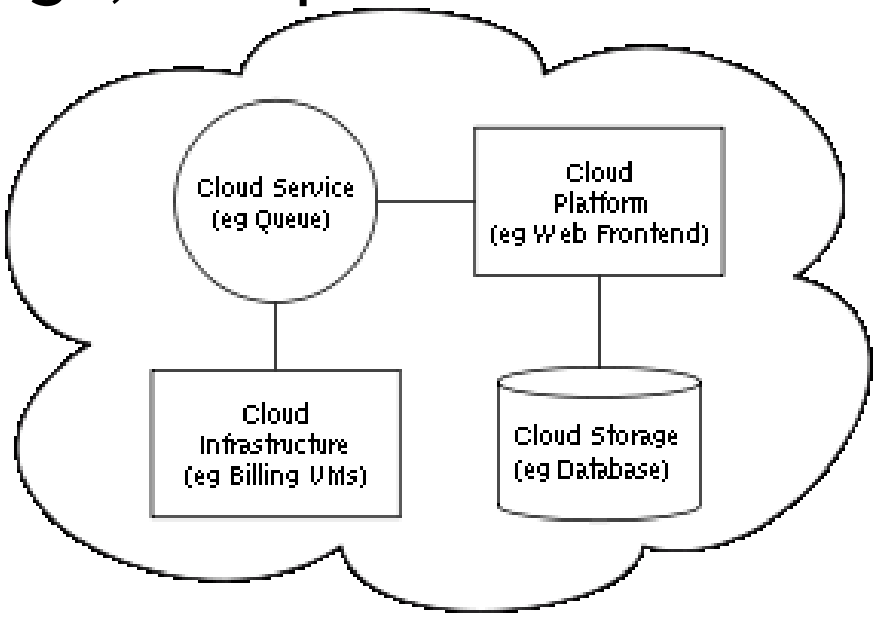


Cloud Computing basics

- Basically: the virtual servers available over the Internet.

- One can use those servers to execute specific functions: storage, computation etc.

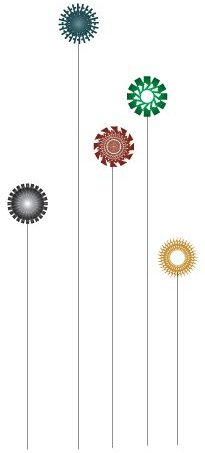
The functions provided as ***services***.



Unipro

Cloud Computing basics

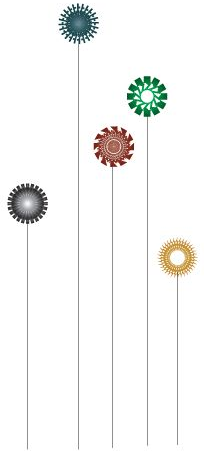
- Customers **do not own** the physical infrastructure.
  - They consume resources as a service and pay only for resources that they use.
- Cloud providers:
  - Google App server, Amazon EC2, Microsoft Azure, etc…



Unipro

Cloud Computing basics

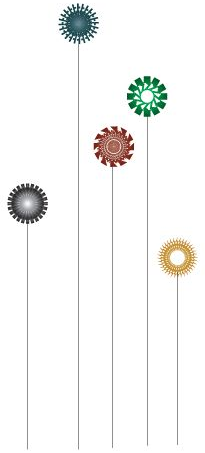
- The difference between cloud-providers is the type of resources/services their sell
- Example of cloud resources
  - Storage (Box.net, DropBox, Oosah …)
    - Easy and chip to rent a huge amount of storage for your documents
    - Available in different geographical locations 24/7
    - Backup services
    - Can be used for any purpose or limited (only photos, videos…)
    - Have special tools to upload data
      - *Pay per GB, plus web traffic*
  - Hardware and application infrastructure (Amazon, GoogleApp, Azure…)
    - Run your own programs (sometimes limited to certain architecture)
    - Rent storage/CPU/memory resources
    - Use various additional services: load balancing, billing, database accounts
      - Pay per storage, traffic, CPU and memory resources, additional services



Unipro

Cloud Computing
basics

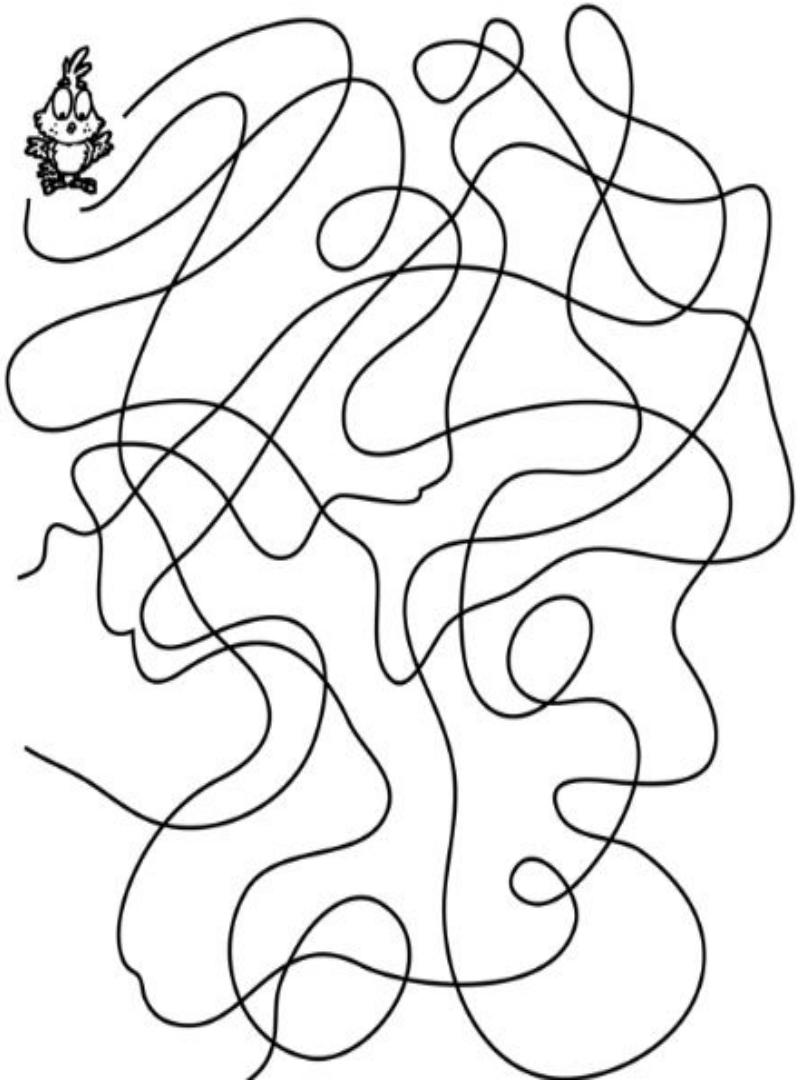
- Example of cloud resources (continued…)
  - Software applications
    - Access to preinstalled software application
    - Database
    - Office-like document processing suites
  - Web services
    - A protocols to run complex computations remotely



Unipro

Cloud Computing
UGENE

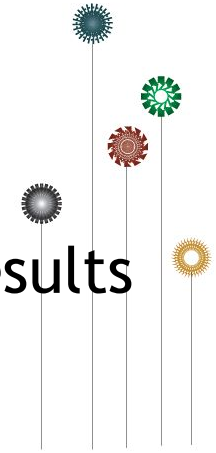
- UGENE cloud model
  - Server part of UGENE can be installed and utilize resources of infrastructure-level cloud providers (like Amazon EC2)
  - UGENE users will see the UGENE server as remote service and utilize it as a service-level cloud provider



Unipro

UGENE Workflows in the cloud

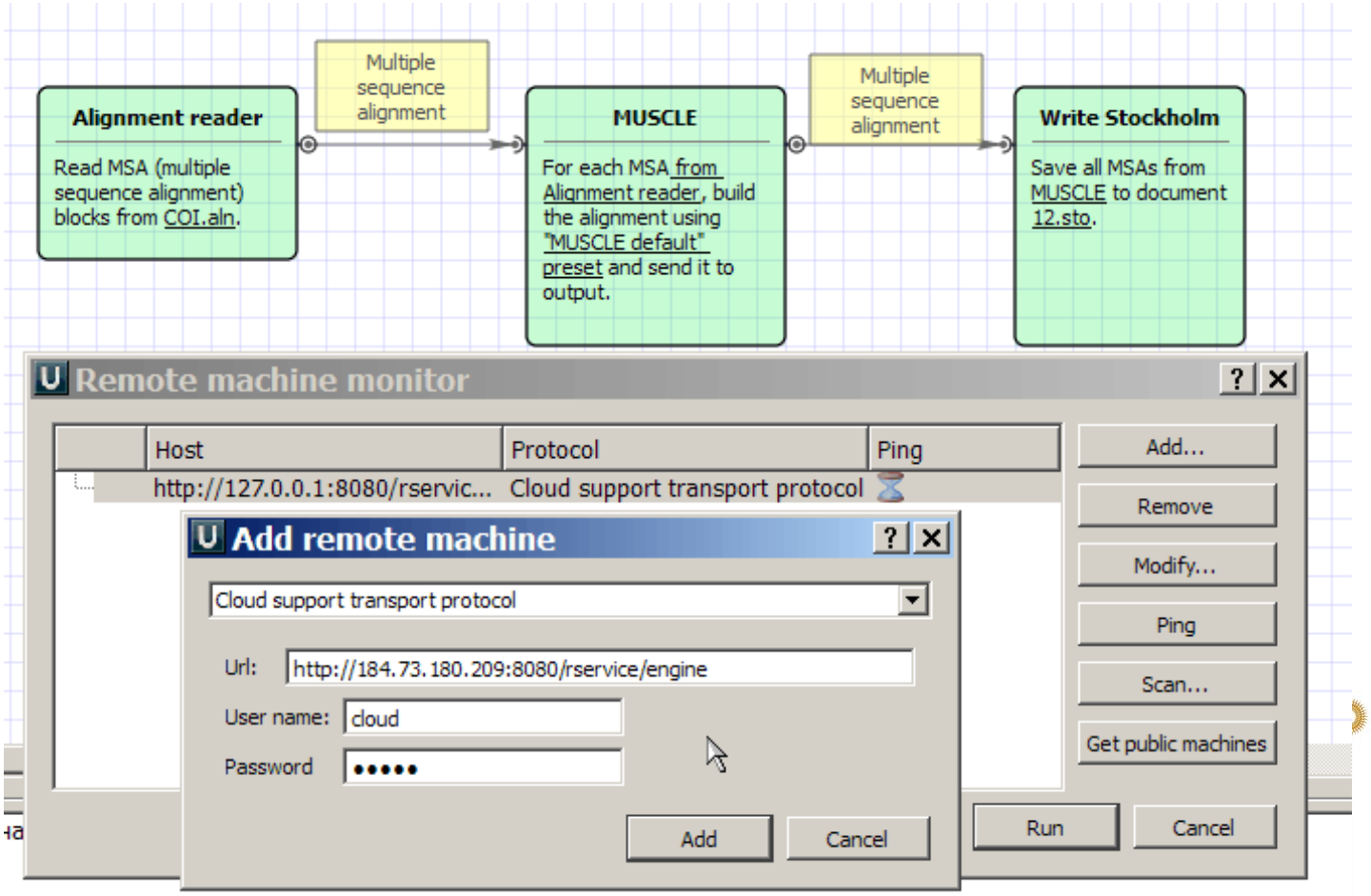
- As simple as local workflow:
  - Add remote machine on the cloud
  - Launch the workflow
    - All your local data will be sent to remote computer
    - Check the task progress in “Tasks” area
    - After task is finished, UGENE will fetch the results



Unipro

UGENE Workflows in the cloud

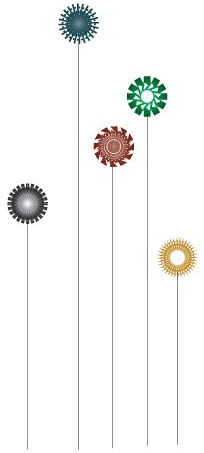
Using open Amazon EC2 installation of UGENE:



Unipro

UGENE Workflows in the cloud

- UGENE team provides you with a free account to access to UGENE server instance (see screenshot on the previous page)
  - This account is limited by hardware resources (Amazon is not free for us, unfortunately)
  - You can install the same UGENE service for you organization
    - Using popular cloud provider (example: Amazon)
    - Inside of the private network of your organization!

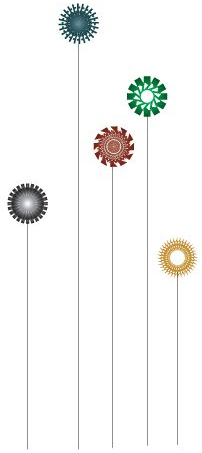


Unipro

UGENE as a Service for your organization

- Why UGENE as a Service in your private network/organization is a good solution

?

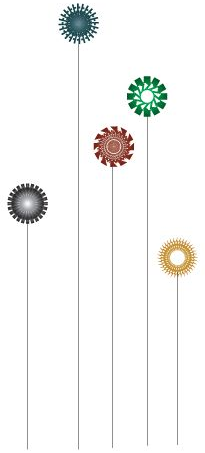


Unipro

UGENE as a Service for your organization

Show these facts to your network administrators and managers:

- No need to install a lot of software: UGENE package integrates dozens of tools and will get more in the future
- Users can create new algorithms in Workflow Designer that are
  - Safe for execution and secure
  - Effectively utilize distributed environment
- No need to pay license fee or ask additional funds for software
- *Ask us for support if needed!*



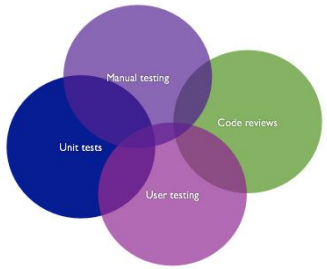
Unipro

Open Discussion

Unipro

Appendix
Some technical details

- UGENE is written in C++, using QT4 library
- Code license: GPLv2
- Distribution bundle size: 12Mb
- Platform supported:
  Windows, Linux, MacOS
- System requirements:
  Pentium 166MMX, 64Mb RAM
- Test base contains ~5000 tests
- Web-site with additional info
  http://ugene.unipro.ru



UGENE Test Report - precommit testing
generated by UGENE
ПН янв 25 2010 22:10:20

Total Runtime: 207 s	Number of tests	Pass	Fail	Excluded	Success Rate
Total:	1947	1947	0	26	100%
Annotator plugin tests	30	30	0	0	100%
Document format tests	266	266	0	0	100%
Enzymes plugin tests	16	16	0	0	100%
ORFMarker plugin tests	25	25	0	0	100%
Repeat finder tests	177	177	0	0	100%
SequenceWalker tests	10	10	0	0	100%
Smith Waterman Tests classic 2	35	35	0	0	100%
Task tests	10	10	0	0	100%

"ugene" package in Ubuntu
Ubuntu » "ugene" package

Unipro