Microbial Bioinformatics

France Génomique – Institut Pasteur, Paris – 25 Sept 2014

Accessibility and inter-operability of bioinformatics tools and resources for microbiology

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université (AMU), France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

A relevant question

- Last week, I started teaching my course
  "*Bioinformatics tools for genomics and proteomics*"
- First lesson: comparative bacterial genomics
  - Sequencing pace: number of available genomes increases exponentially.
  - Annotation pace: ratio between human-reviewed (~546,238) and un-reviewed (82,126,897) entries in Uniprot.
  - Difference between model organisms (*E.coli* 98% reviewed entries) and the vast majority of other organisms (almost no reviewed entry).
  - Benefits and limitations of similarity-based annotation.
- A student's question
  - ***"Since we will never be able to run the experiments to check the function of each protein of each of these organisms, does it make any sense to sequence more and more genomes ?"***
- This is actually an excellent question. Today we dispose of ~10,000 completely sequenced bacterial genomes.
  - What is the added value ?

What can we learn from >10,000 bacterial genomes ?

- Context-based annotation
  - Gene grouping : operons, directons, syntons.
  - Gene co-occurrence: phylogenetic profiles; co-occurrence networks.
    - Note: an advantage of multi-genome analyses is that we can gain knowledge not only from the presence, but also from the ***absence*** of genes/functions/processes.
- Regulation: phylogenetic footprints, prediction of co-regulation networks.
- Metabolism
  - Metabolic pathway inference.
  - Discovery of alternative pathways (by interconnecting enzymes).
  - Evolution of metabolic pathways.
- Genome evolution
  - Gene phylogeny (phylogenetic inference).
  - Species phylogeny at a genome scale (phylogenomics).
  - Gene mobility – horizontal transfers (cross-branches in the “tree” of life).
  - Genomic islands.
  - Genome variation (hundreds of strains for some species of interest).
- …
- In order to gain benefit of all these possibilities, we need usable tools.

Bioinformatics tools

The "Next Generation Gap"

- High-Throughput Sequencing technologies enable to produce huge amounts of data
- This technology is now used to obtain genome-scale profiles of
  - TF binding (ChIP-seq)
  - Chromatin accessibilty and histone modifications (ChIP-seq)
  - Expression profiles, alternative splicing (RNA-seq)
  - Inter-individual variations (SNPs, Whole-genome sequencing)
  - ...

- Quote from John D McPherson in his Nature Method commentary "*The Next Generation Gap*"
  - *There is a growing gap between the generation of massively parallel sequencing output and the ability to process and analyze the resulting data. New users are left to navigate a bewildering maze of base calling, alignment, assembly and analysis tools with often incomplete documentation and no idea how to compare and validate their outputs. Bridging this gap is essential, or the coveted $1,000 genome will come with a $20,000 analysis price tag.*

McPherson, J. D. (2009). Next-generation gap. Nat Methods 6, S2-5.

Semantic consideration: what is a bioinformatics tools ?

- Method != tool
- Every year, hundreds of bioinformatics methods are published in specialized journals.
- How many of these methods are associated to tools, i.e. ***software usable by other people than developer*** ?
- How many of these tools are interfaced for biologists ?

Issues for development and maintenance of bioinformatics tools

- Tool **accessibility** to human beings
  - Command-line
  - Web interface (remote usage by a human being)
- **Inter-operability**: remote access
  - Via custom libraries (e.g. Ensembl Perl API; R BioConductor)
  - Web services (SOAP/WSDL, REST)
- **Portability**: can the tool be installed on other machines ?
  - Mirroring
  - Virtualization
- **Documentation**
  - Quick help
  - Detailed manuals
  - Tutorials, protocols
- **Tractability**
  - Indicate the parameters used for the analysis in output files.
- **Reproducibility**
  - Benchmarking: run tool with test cases and check result consistency.
- **Maintainability**
  - Software design
  - Code documentation

How to maintain bioinformatics tools?

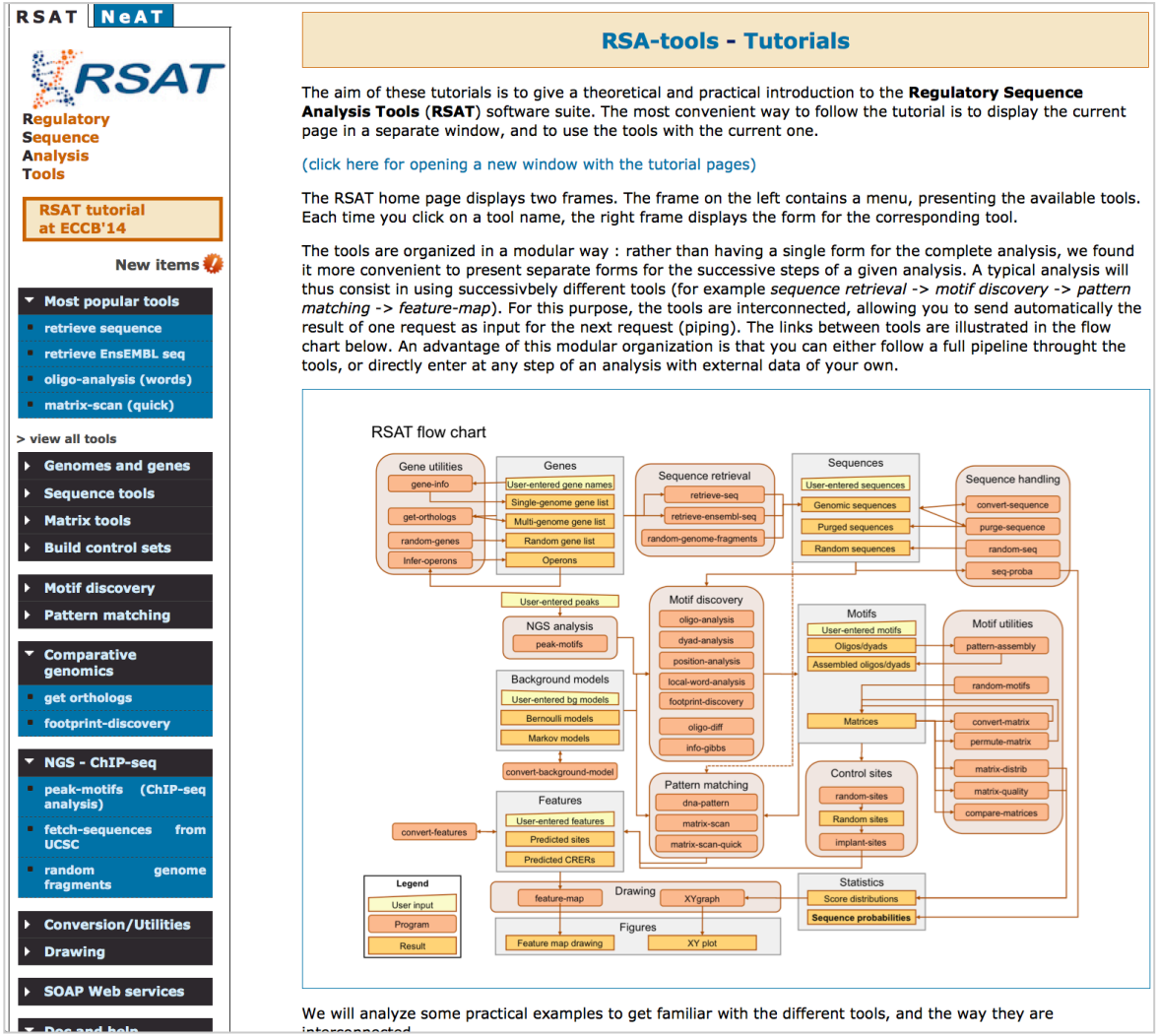
- What are bioinformaticians doing ?
  - Bioinfo-analysts
  - Bioinfo-developers
  - Bioinfo-curators

- Who develops tools
  - PhD students
  - Postdocs
  - Permanent researchers
  - Engineers

- Who can maintain tools >3 years after publication ?
  - Obsolescence of bioinformatics tools

Regulatory Sequence Analysis Tools (RSAT)
Network Analysis Tools (NeAT)

Regulatory Sequence Analysis Tools (RSAT)

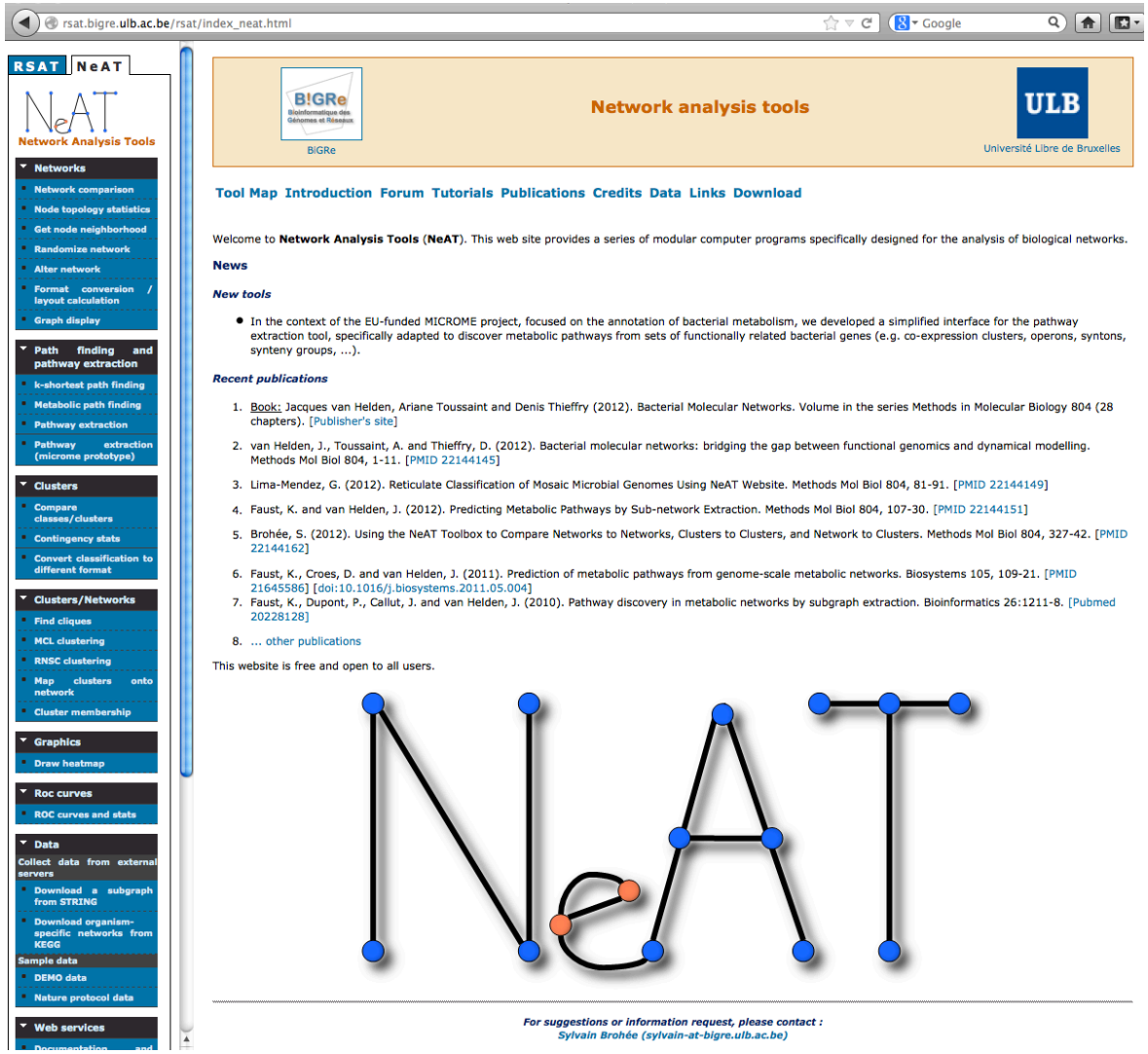


- User-friendly interface
  - http://www.rsat.eu/
  - Manuals
  - Demos
- Automated utilization
  - Stand-alone tools
  - Web services
- 2854 supported organisms (Oct 2013)
  - 1998 Bacteria
  - 150 Archaea
  - 104 Fungi
  - All EnsEMBL genomes (via Ensembl API)
  - ...
- Tasks
  - Sequence retrieval
  - Pattern discovery
  - Pattern matching
  - Feature-map drawing
- Applications
  - ChIP-seq peaks
  - Co-expression clusters
  - Phylogenetic footprints
  - ...

- Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S. and van Helden, J. (2008). RSAT: regulatory sequence analysis tools. Nucleic Acids Res 36, W119-27.
- Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D. and van Helden, J. (2011). RSAT 2011: Regulatory Sequence Analysis Tools. Nucleic Acids Res Web software issue 2011, in press.

Network Analysis Tools

- Biological networks
  - Protein interactions
  - Gene regulation
  - Metabolic pathways
- Approaches
  - Path finding.
  - Clustering.
  - Subgraph extraction.
  - Random or altered networks (controls).
  - Network comparisons
  - …



- Brohee S, Faust K, Lima-Mendez G, Sand O, Janky R, Vanderstocken G, Deville Y, van Helden J. 2008. NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. Nucleic Acids Res 36(Web Server issue): W444-451.
- Brohee S, Faust K, Lima-Mendez G, Vanderstocken G, van Helden J. 2008. Network Analysis Tools: from biological networks to clusters and pathways. Nat Protoc 3(10): 1616-1629.

Bioinformatique des Génomes et des Réseaux (BiGRe)
Université Libre de Bruxelles
http://www.bigre.ulb.ac.be/


Sylvain Brohée
Postdoc


Nicolas Simonis
Postdoc


Didier Croes
Postddoc


Didier Gonze
Premier assistant

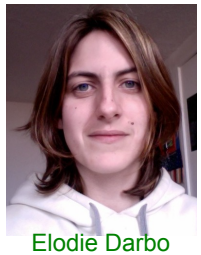

Myriam Loubriat
Secretary


Ariane Toussaint
Professor Emeritus


Jacques van Helden
Professor


Leon Juvenal Hajingabo
PhD Student


Maud Vidick
PhD Student (co-direction)


Elodie Darbo
PhD Student
co-direction Marseille


Alejandra Medina
PhD Student
co-direction Mexico


Morgane Thomas-Chollier
PhD student+postdoc


Matthieu Defrance
Postdoc


Olivier Sand
Postdoc


Jean Valéry Turatsinze
PhD student


Raphaël Leplae
Postdoc


Gipsi Lima
PhD + Postdoc


Karoline Faust
PhD student


Rekin' s Janky
PhD student


Eric Vervisch
Research fellow

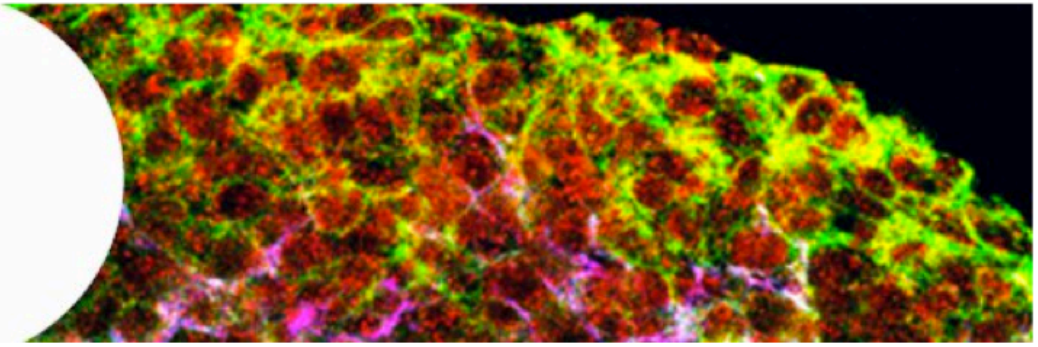

Guy Bottu
Postddoc

- **Conception, implementation, evaluation and application of bioinformatics methods for the analysis of genomes and biomolecular networks.**
- **Regulatory sequences**
  - Motif analysis algorithms *(Olivier Sand , Matthieu Defrance , Maud Vidick, Alejandra Medina-Rivera)*
  - Evolution of cis-acting elements in Bacteria (*Rekin's Janky, Alejandra Medina-Rivera*)
  - Regulation of development in Drosophila (*Jean Valéry Turatsinze, Elodie Darbo*)
  - Hox regulation in Vertebrates (*Morgane Thomas-Chollier)*
  - Work flows on transcriptional regulation (*Olivier Sand, Eric Vervisch)*
- **Biomolecular networks**
  - Network analysis tools (*Sylvain Brohée)*
  - Inference of metabolic pathways (*Karoline Faust, Didier Croes*)
  - Host-virus interaction networks (*Nicolas Simonis, Leon Juvénal Hagingambo)*
  - Analysis of regulatory networks (*Sylvain Brohée, Rekin's Janky*)
- **Mobile genetic elements in prokaryotes** *(Raphaël Leplae, Gipsi Lima, Ariane Toussaint)*
- **Modelling of dynamical systems** *(Didier Gonze)*
- **e-Learning for bioinformatics** *(Guy Bottu)*

New address (since Nov 2011)

TAGC Inserm U1090

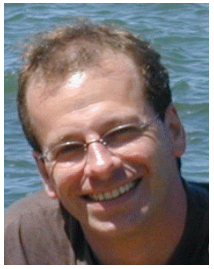
technological advances for genomics and clinics



- Home
- Research
- People
- Software
- TGML platform
- Publications
- Teaching
- About us
  - Contact us
  - Organization
  - Fundings
  - Seminars
  - Intranet

This month's TAGC

Collaborators involved in the development of the Regulatory Sequence Analysis Tools



Bruno André
(ULB, Bruxelles, Belgium)
Initiation of the RSAT project.
Conception of oligo-analysis.
Analysis of yeast regulation.

ULB

Denis Thieffry
(ENS, Paris, France)
ChIP-seq tools +
regulatory networks.

Carl Herrmann
(TAGC, Marseille, France)
ChIP-seq analysis (peak-motifs, compare-matrices).

Elodie Darbo
(TAGC, Marseille, France)
Analysis of co-expression clusters + ChIP-seq data (transcription factors, chromatin marks).

Julio Collado-Vides
(CCG, Cuernavaca - Mexico)
Initiation of the RSAT project
Analysis of regulation
in bacterial genomes

CCG
Centro de Ciencias Genómicas

Alejandra Medina-Rivera
(CCG, Cuernavaca - Mexico)
Evaluation of matrix quality.
Phylogenetic footprints in Baceria.

CCG
Centro de Ciencias Genómicas

Lionel Spinelli
(TAGC, Marseille, France)
Development of peak-footprints.

Cei Abreu-Goodger
(Sanger Institute, Hinxton, UK)
Evaluation of matrix quality
on bacterial regulons.

wellcome trust sanger institute

Regulatory Sequence Analysis (http://rsat.eu/)

RSAT in the cloud

Institut Français de Bioinformatique (IFB)

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

Demo

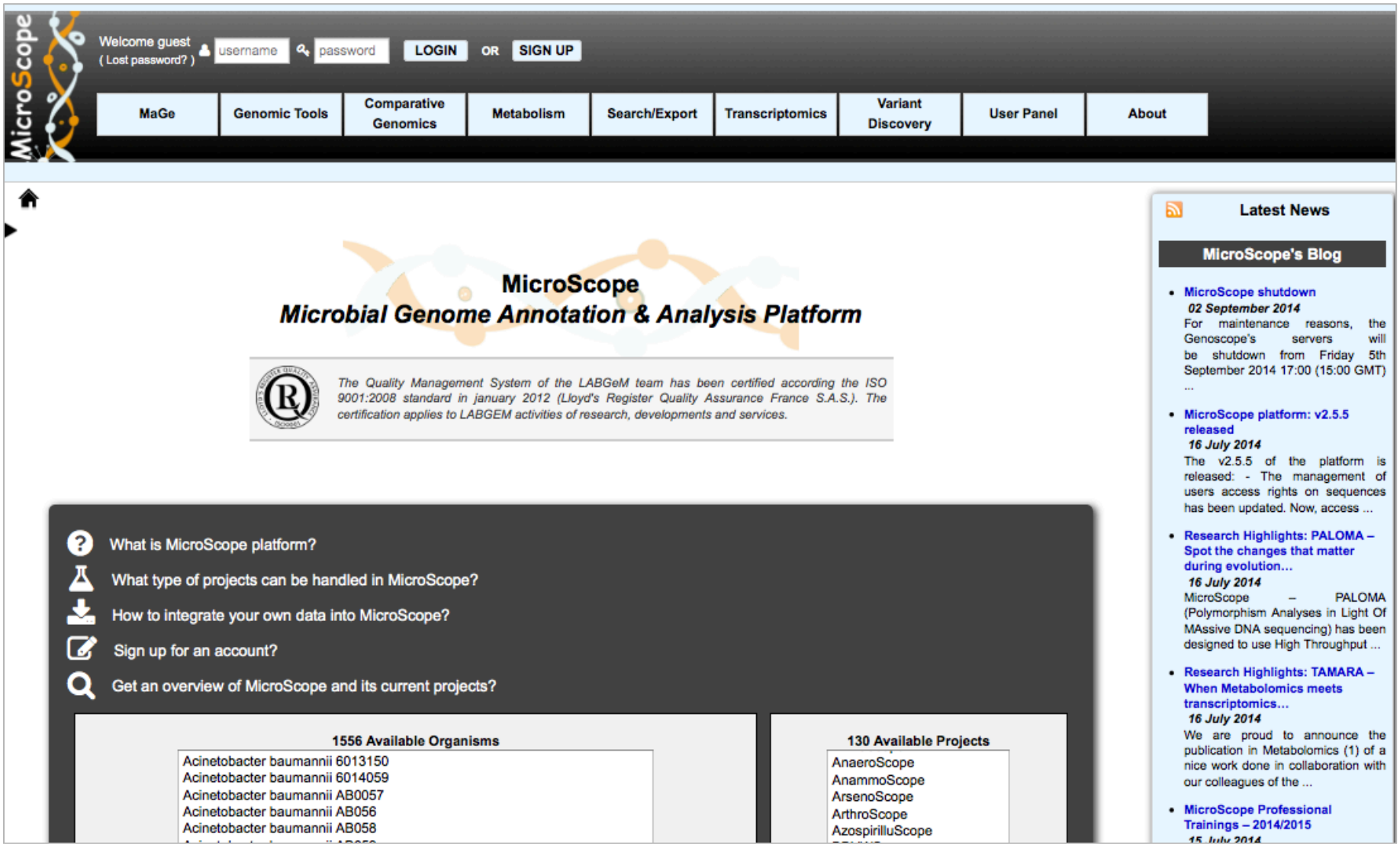
- https://cloud.france-bioinformatique.fr/

Interoperability

Accessing remote resources by Web services

The MICROSCOPE platform

- MICROSCOPE availability:
  - Web interface
  - Web services (SOAP/WSDL in 2013, REST since 2014)



REST clients for MICROSCOPE

- Aurélie Bergon (IR France Génomique) developed REST clients for MICROSCOPE (python).
- Functionalities
  - supported organisms
  - GPR tables (gene – protein – EC – reaction)
  - List of all reactions
  - Organism-specific reactions
  - Detailed information about a given reaction

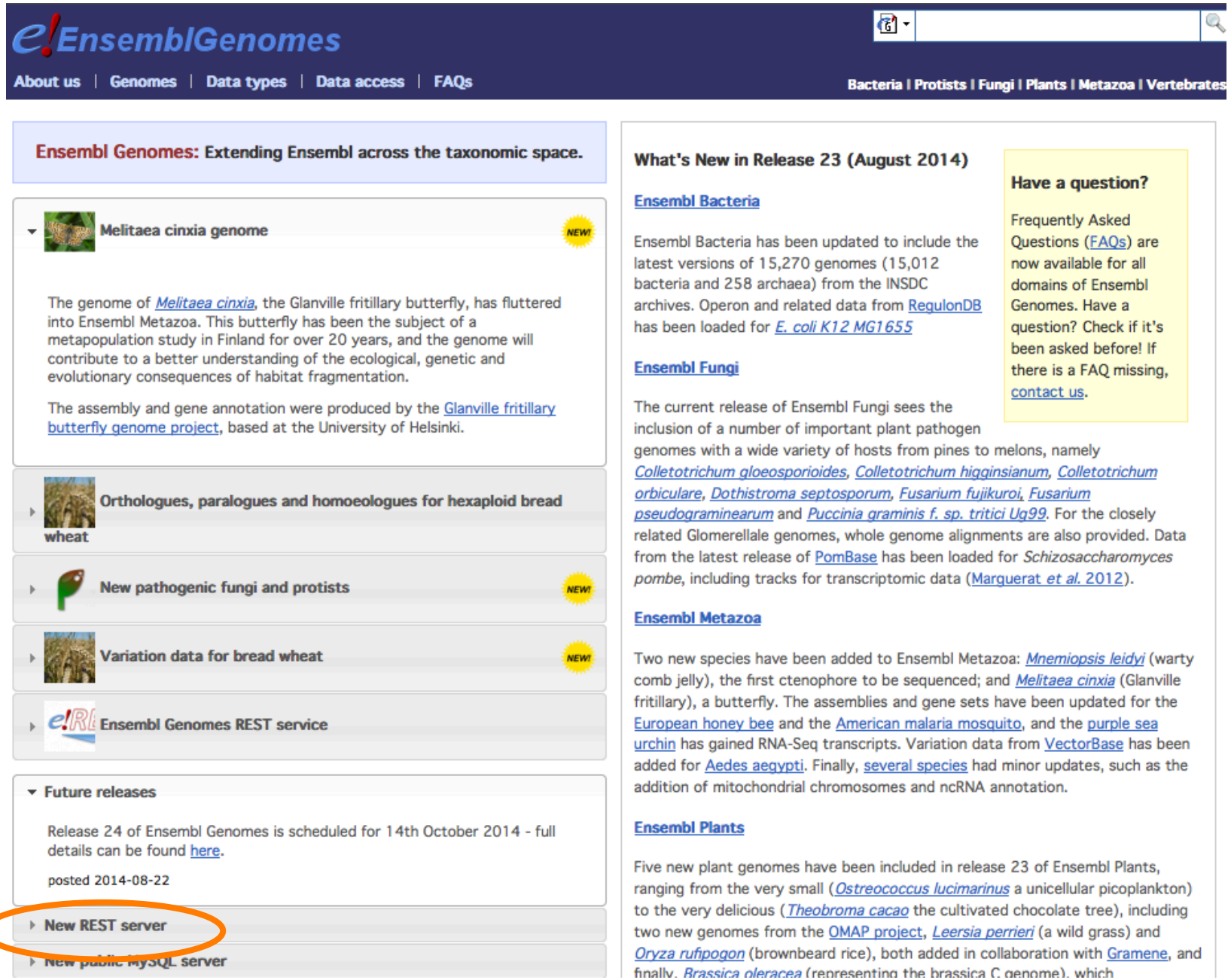
REST clients for MICROSCOPE – online help

- Two-level help
  - task list (python “positional arguments”
  - detailed help for each task

```
microscope_get --help
```

```
usage: microscope_get [-h] [--version]
                      {reactions,reactionInfo,gpr,organisms,reactionList} ...

This program fetches data from MicroScope database
(http://www.genoscope.cns.fr/agc/microscope/), such as the list of supported
organisms, and their annotations (genes, transcripts, proteins, reactions),
metabolic networks.

positional arguments:
  {reactions,reactionInfo,gpr,organisms,reactionList}
    organisms           Get information about organisms supported at
                        MicroScope
    gpr                 Collect gene-EC-protein-reaction table for a selected
                        species.
    reactionInfo        Get detailed information about a given reaction
    reactionList        Get reaction list for a given species
    reactions           Get the list of reactions supported at Microscope (all
                        species).

optional arguments:
  -h, --help            show this help message and exit
  --version             show program's version number and exit
```

http://ensemblgenomes.org/



Ensembl / EnsemblGenomes

- Genome availabilty
- EnsembGenomes version 76-23
  - 11,010 “Bacteria” (actually procaryotes, incl. Bacteria + Archaea)
  - 52 Metazoa
  - 45 Fungi
  - 33 Plants
  - 29 Protists
- Programmatic interfaces
  - BioMart.
  - Perl API (object-oriented model).
  - Direct SQL access (requires to know their relational schema).
  - Since June 2014: REST Web services.
  - Good documentation, with examples of code in Perl, Python, java, …

Demo: RSAT client to EnsemblGenomes, via Perl API

```
install-ensembl-genome  -help
```

```
NAME
    install-ensembl-genome

VERSION
    $program_version

DESCRIPTION
    Install on RSAT genome sequence, genomic features and (optionally)
    variation features for a genome form Ensembl (http://www.ensembl.org/).

    The connection to ensembl is ensured by a combination of their Perl API
    and their ftp site (some information can not be obtained directly from
    the API, e.g. the taxonomy).

AUTHORS
    Jeremy.Delerce@etu.univ-amu.fr
    Jacques.van-Helden@univ-amu.fr

CATEGORY
    util
```

Demo: RSAT client to EnsemblGenomes, via Perl API

- download-ensembl-genome -available_species -ensembl_genomes

```
install-ensembl-genome -available_species -ensembl_genomes
```

```
Acaryochloris_marina_mbic11017  genome, features        bacteria
Acetobacter_pasteurianus_ifo_3283_01    genome, features        bacteria
Acetobacter_pasteurianus_ifo_3283_01_42c        genome, features        bacteria
Acetobacter_pasteurianus_ifo_3283_03    genome, features        bacteria
Acetobacter_pasteurianus_ifo_3283_12    genome, features        bacteria
Acetobacter_pasteurianus_ifo_3283_22    genome, features        bacteria
bacteria
...
Acyrthosiphon_pisum     genome, features        metazoa
Aedes_aegypti   genome, features        metazoa
Aegilops_tauschii       genome, features        plants
Albugo_laibachii        genome, features        protists
Amborella_trichopoda    genome, features        plants
Amphimedon_queenslandica        genome, features        metazoa
Anopheles_darlingi      genome, features        metazoa
Apis_mellifera  genome, features        metazoa
Arabidopsis_lyrata      genome, features        plants
Ashbya_gossypii genome, features        fungi
Aspergillus_clavatus    genome, features        fungi
Aspergillus_flavus      genome, features        fungi
...
```

RSAT clients to collect GO annotations from EnsemblGenomes

- Developer: Lucas Françoise (M1 student; 2014)
- Generic tool: go_analysis
- Modularity: tasks (via python “positional arguments”)
  - GO term definitions: downloading, parsing
  - GO annotation (organism-specific): download, inheritance, enrichment analysis
- Two-level help
  - for the tool
  - for each task

```
python $RSAT/python-scripts/go_analysis.py
```

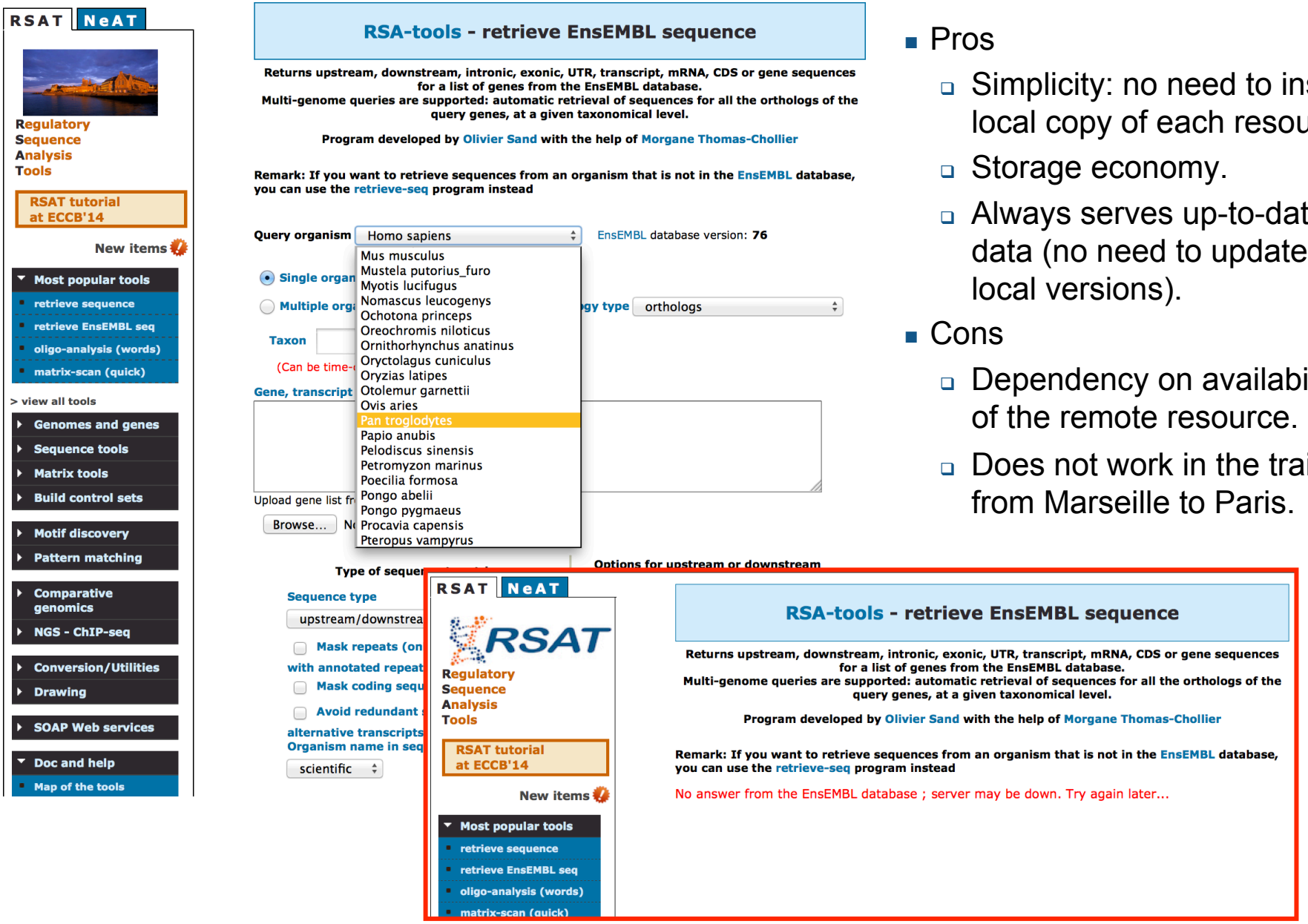
```
Usage : python go_analysis.py <task> [<options>]

Command 'task' expects one argument, received zero.
List of accepted arguments:
          download_go
          parse_go
          get_annotations
          ancestor
          enrichment
```

Integrating remote queries in Web tools

Using remote resources to flesh the tools

- Between RSAT servers
  - supported-organisms-server -url http://rsat.sb-roscoff.fr/
- From RSAT to external resources
  - supported-organisms-genoscope (REST Web services, python client)
  - supported-organisms-galileo (REST Web services, Perl client)
  - supported-organisms-ucsc (DAS server, via HTTP)
  - supported-organisms-ensembl (Perl API, requires SQL port)

Web tools relying on “life” queries to remote resources



- Pros
  - Simplicity: no need to install local copy of each resource.
  - Storage economy.
  - Always serves up-to-date data (no need to update local versions).
- Cons
  - Dependency on availability of the remote resource.
  - Does not work in the train from Marseille to Paris.

Documenting the tools

Multiple levels of documentation

- Method publication
  - Detailed description in bioinformatics journals
  - Assessment of tool performances, accuracy, relevance of the results.
  - In many case no proper publication of the method as such, but as part of result papers.
- Web interface
  - Manual pages
  - Demo buttons filling the form with relevant study cases
- Tutorials (documented study cases)
  - How does it work ? Principle of the method.
  - How to choose parameters ?
  - How to interpret the results ?
  - Selected study cases (should include successes, but also difficulties or failures)
- Protocols
  - A step-by-step detailed tutorial.
  - Structured according to some publisher's standard (Nature Protocols, Methods in Molecular Biology, …)
  - Comment critical choices.
- Command-line tools: on-line help
- Code documentation: essential for maintainability
- On-line manuals: general description + list of options.
- Demo buttons: quick access to illustrative conditions of utilization.

Protocols

TABLE 1 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
16	The matrix-scan result does not appear after a long waiting time	For a large data set with computer-intensive options, the analysis can take several minutes, or even hours to run	Select the 'email' output in the matrix-scan form instead of 'display'. Alternatively, a more stringent threshold can be applied (e.g., discard negative scores with a lower threshold of 0 on the weight score)
16	There are too many matches to display and results are difficult to interpret	The threshold may be too loose and allows many false predictions. Another possibility is that your background model does not correspond to the composition of the sequence	Relaunch the analysis with a more stringent threshold. To obtain an idea about the number of matches expected by chance for a given weight threshold, you can use the program *matrix-distrib*

ANTICIPATED RESULTS

At the end of this protocol, the user should be able to visualize predicted sites and CRERs along the input sequence. In addition, several random controls should have been run to evaluate the quality of the predictions. **Figure 8** shows the feature-maps obtained with the *even-skipped* study case.

Figure 8a displays the binding sites (top) and CRMs (bottom) annotated in the 5,500-bp region located upstream the *even-skipped* TSS. TFBSs were extracted from ORegAnno[29] and CRMs from REDfly[31,37]. The right side of the figure corresponds to the region immediately upstream of the *even-skipped* gene whereas the left side of the figure corresponds to the most distal region. The coordinates of these elements are provided as supplementary material on the website (see *oreganno_eve_annotation.ft* for TFBSs and *redfly_eve_annotation.ft* for CRMs). **Figure 8a** can be reproduced by copy-pasting the content of these files in the *feature-map* form. The figure shows that the annotated TFBSs for the 12 transcription factors of interest fall into the four CRMs of the *even-skipped* promoter. The perfect correspondence between the annotated TFBSs and CRMs probably reflects some experimental or annotation bias and should by no means be taken as an evidence that these factors do not bind anywhere else in the region.

Figure 8b displays *matrix-scan* predictions in the *even-skipped* promoter. Individual site predictions are mostly found inside or in the neighborhood of the annotated CRMs. CRER predictions with the first set of parameters (Pval ≤ 0.001, crer_sig ≥ 0) show a landscape with many overlapping CRERs.

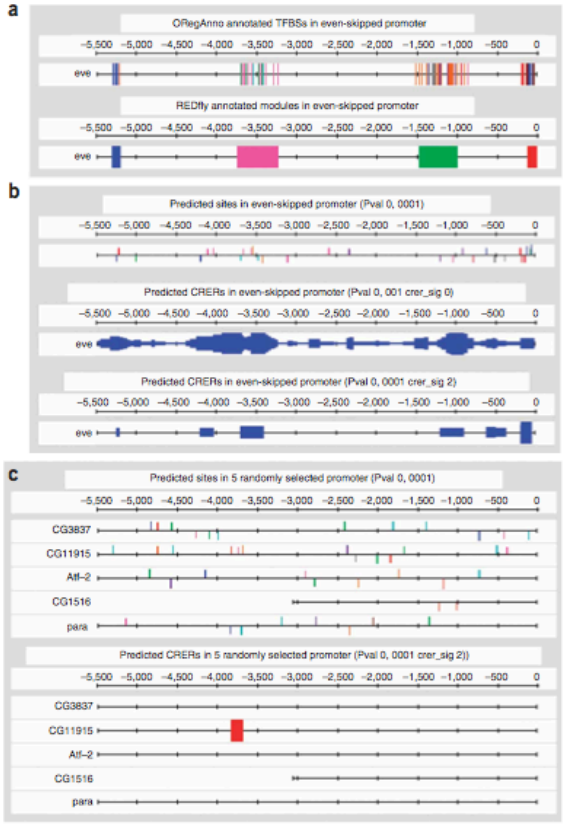
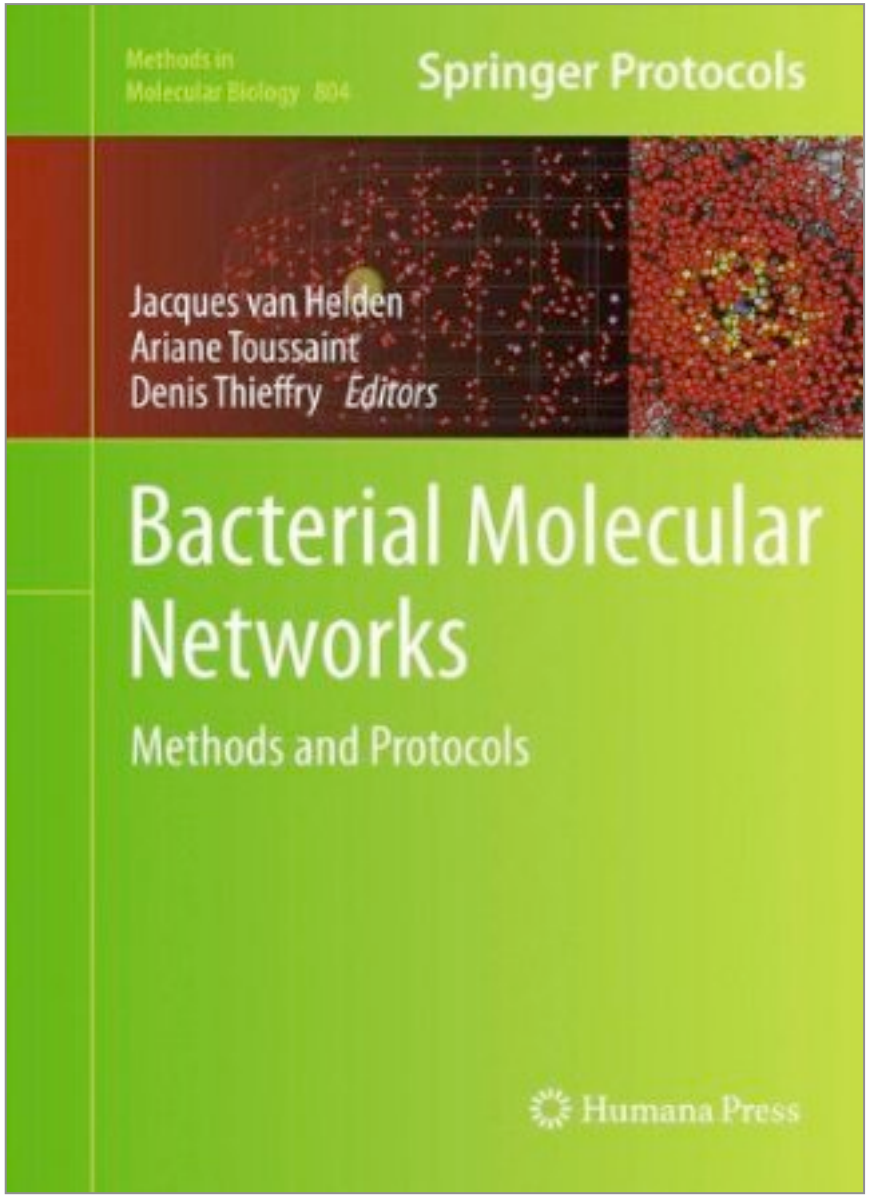


Figure 8 | Feature-maps for the *even-skipped* example. (**a**) Annotated transcription factor-binding sites (TFBSs) and *cis*-regulatory modules (CRMs) in the *even-skipped* promoter. (**b**) Matrix-scan predictions of sites and *cis*-regulatory element-enriched regions (CRERs) in the *even-skipped* promoter. (**c**) Matrix-scan predictions of sites and CRERs in randomly selected *drosophila* promoters.

Published protocols

1. Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J. 2012. A **complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs**. Nat Protoc 7(8): 1551-1568.
2. Defrance M, Janky R, Sand O, van Helden J. 2008. **Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences**. Nat Protoc 3(10): 1589-1603.
3. Sand O, Thomas-Chollier M, Vervisch E, van Helden J. 2008. **Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services**: an example with ChIP-chip data. Nat Protoc 3(10): 1604-1615.
4. Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J. 2008. **Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules**. Nat Protoc 3(10): 1578-1588.
5. Brohee S, Faust K, Lima-Mendez G, Vanderstocken G, van Helden J. 2008. **Network Analysis Tools: from biological networks to clusters and pathways**. Nat Protoc 3(10): 1616-1629.
6. Faust K, van Helden J. 2012. **Predicting Metabolic Pathways by Sub-network Extraction**. Methods Mol Biol 804: 107-130.
7. Sand O, van Helden J. 2007. **Discovery of motifs in promoters of coregulated genes**. Methods Mol Biol 395: 329-348.
8. Janky R, van Helden J. 2007. **Discovery of conserved motifs in promoters of orthologous genes in prokaryotes**. Methods Mol Biol 395: 293-308.

MiMB Volume on Bacterial Molecular Networks



- A collaborative volume on biological and bioinformatics methods for analyzing bacterial molecular networks.
- From data acquisition (experiments) to dynamical modelling.
- Variety of network types: interactome, gene co-occurrences, metabolic networks, regulatory networks, …
- Two chapter types
  - Reviews on the state-of-the art of a domain.
  - Commented protocols explaining how to use software tools, and how to interpret their results.

- Jacques van Helden, Ariane Toussaint and Denis Thieffry (2012). Bacterial Molecular Networks. Methods in Molecular Biology, Volume 804 (28 chapters).

Conclusions & Perspectives

How can we cope with the increase of microbial genomic data ?

- Combining resources maintained at various locations
  - Genome centers (e.g. EnsemblGenomes, UCSC, NCBI)
  - Specialized databases (e.g. MICROSCOPE, RegulonDB)
  - Generic environments with add-on capabilities (e.g. R, CytoScape, Galaxy, …)
  - Specialized tools (e.g. RSAT, NeAT)
- Local integration
  - Cost in time, space, updates
- Programmatic interfaces
- Workflows

Table-ronde: quelles formations pour le traitement des données NGS ?

Participants

- Erwan Corre (corre@sb-roscoff.fr)
- Jean-François Gibrat (jean-francois.gibrat@france-bioinformatique.fr)
- Hélène Chiapello (helene.chiapello@toulouse.inra.fr)
- BLANCHET Christophe (Christophe.BLANCHET@france-bioinformatique.fr)
- Guy Perriere (Guy.Perriere@univ-lyon1.fr)
- Lionel Frangeul (lfrangeu@pasteur.fr)
- Jacques van Helden (Jacques.van-Helden@univ-amu.fr)

Questions

- Quel(s) environnement(s) informatique(s) pour les formations ?
  - Galaxy ?
  - Ressources spécialisées (MICROSCOPE, RSAT, …) ?
  - Outils en ligne de commande ?
  - Instances personnelles sur le cloud ?
  - Machines Virtuelles préinstallées + jeux de données (Virtualbox)
  - Environnement virtuel (docker) ?
- Comment adapter les formations aux besoins particuliers des participants ?
  - Analyse de cas, conception de workflows.
  - Traitement des données propres des participants.
- Equilibre théorie/pratique
- Peut-on générer du matériel didactique réutilisable?
  - Portail des ressources pédagogiques existantes (sous Galaxy, R, console Unix, …)
  - Ressources de support aux formation « directes »: tutoriels, vidéos, vidéos des cours théoriques,…
  - Effort post-formation: diapos et exercices commentés suite à la formation
  - E-learning
  - MOOC
- Adaptation des outils d'analyse et de formation à l'augmentation des données
  - Comment traiter les milliers de génomes bactériens los de formations ?
- Forum France Génomique / IFB: utilisateurs <-> spécialistes-consultants
- Partenariats avec universités
  - Matériel des formations en master, licence, ….

Ecole de Bioinformatique Aviesan (EBA)

- http://ecole-bioinfo-aviesan.sb-roscoff.fr/
- 3ème édition: du 5 au 10 octobre 2014