PARConnector : sans détour du Langage R au Cloud Big Data

Application à l'analyse Metagénomique

Ndeye Aram GAYE - Laurent PELLEGRINO

Plan

- Présentation des sociétés
- ProActive Workflows and Scheduling
- Package “PARConnector”
- Cas d’utilisation: analyse métagénomique quantitative

Présentation des sociétés

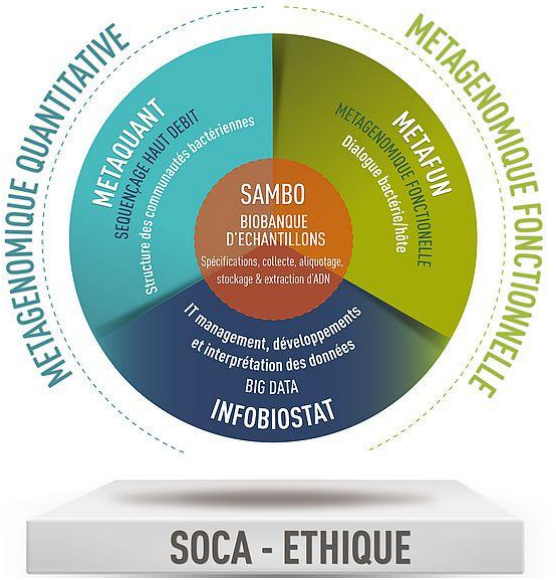
MetaGenoPoliS

Financement des Investissements d'Avenir 19M€ pour 2012-2019

Centre d'excellence en métagénomique humaine : www.mgps.eu

4 plateformes

1 centre de recherche en éthique



Objectifs :

- Explorer le rôle du microbiome intestinal dans la santé et les maladies complexes humaines
- Identification de biomarqueurs et d'outils de diagnostique et de prédiction dans le cadre d'une médecine préventive.
- Ouvrir de nouvelles pistes thérapeutiques centrées sur le microbiome
- Transfert de ces technologies dans le domaine de la nutrition

INRA SCIENCE & IMPACT

ICAN TOGETHER

Université Catholique de Lyon

ActiveEon

- Essaimage de l'INRIA, créée en octobre 2007
- 15 collaborateurs (ingénieurs, docteurs, commerciaux)
- Éditeur de logiciel libre, membre de la communauté OW2
- Des solutions applicables à de nombreux domaines : Ingénierie, Bio Technologies, Finance, Technologies de l'Information, etc.
- Une croissance en France et à l'international

Nos clients

Technologies de l'Information

orange
hp
amadeus
Your technology partner
HUAWEI
AMD
Mastek
Home Office
BUILDING A SAFE, JUST AND TOLERANT SOCIETY

Ingénierie Energie / Aéronautique Espace

ifp Energies nouvelles
cea
gtd
AIRBUS
AN EADS COMPANY
cnes
Snecma
SAFRAN Group
SCP
Sea Eng
Open Ocean
Marine Energy Consul
Artelys
optimization solutions

Bio Technologies

INRA
SCIENCE & IMPACT
L'ORÉAL
enterome
bioscience
EMBL-EBI

Medias Distribution

Mediametrie
TAGSYS

Finance

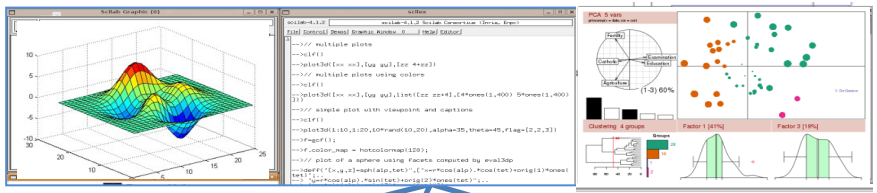
GMO
DEXIA
LA FRANÇAISE AM
Investing together

Notre mission

- L’innovation au service de nos clients
- Anticiper les défis technologiques
- Renforcer notre expertise sur des domaines clés
  - Systèmes Distribués et le Cloud Computing
- Être le partenaire privilégié de nos clients
- Être toujours à la pointe de l’innovation grâce à un investissement en R&D important

ProActive
Workflows & Scheduling

Workflows & Scheduling (1)



Static Policy

Clusters

Timing Policy 12/24

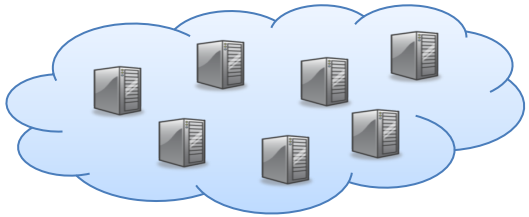
Desktops

Dynamic Workload Policy

EC2, Azure, HP Cloud, …

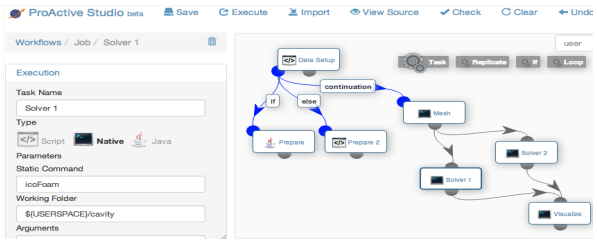
Dedicated resources

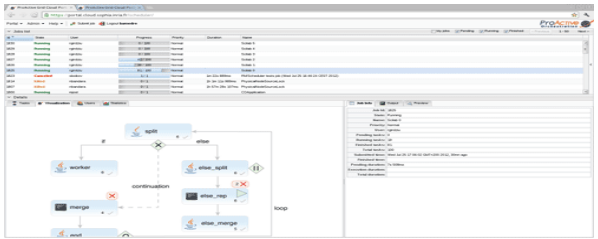
Desktops

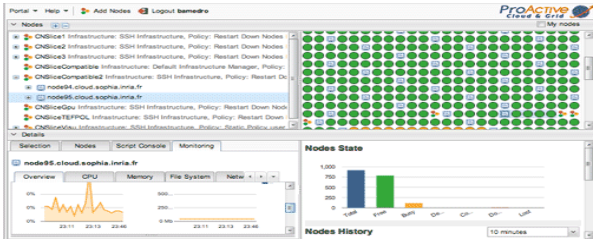


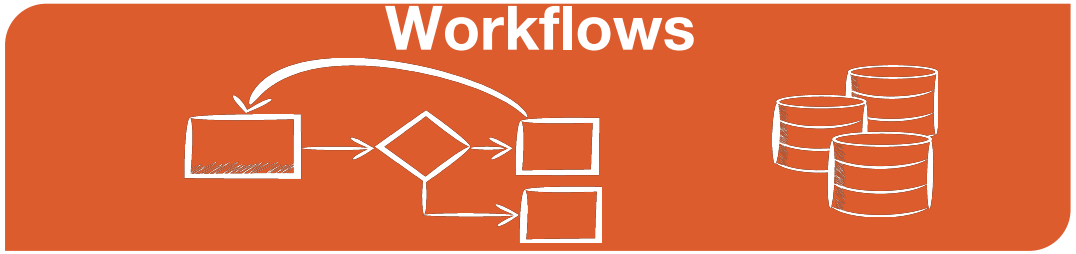
Amazon EC2

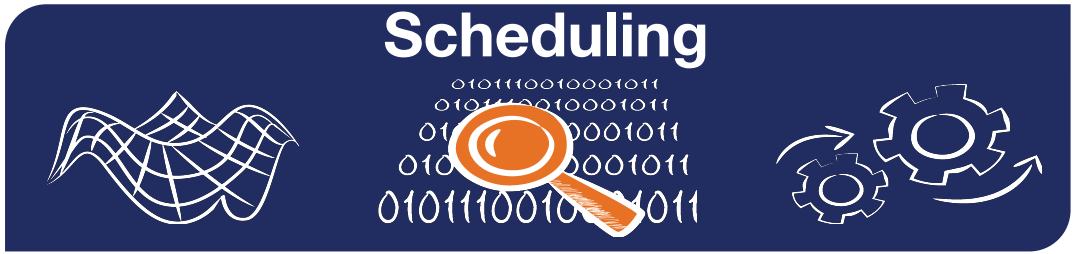
Workflows & Scheduling (2)

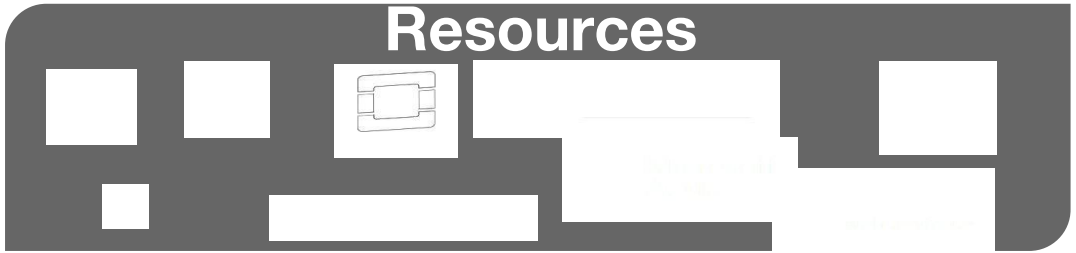






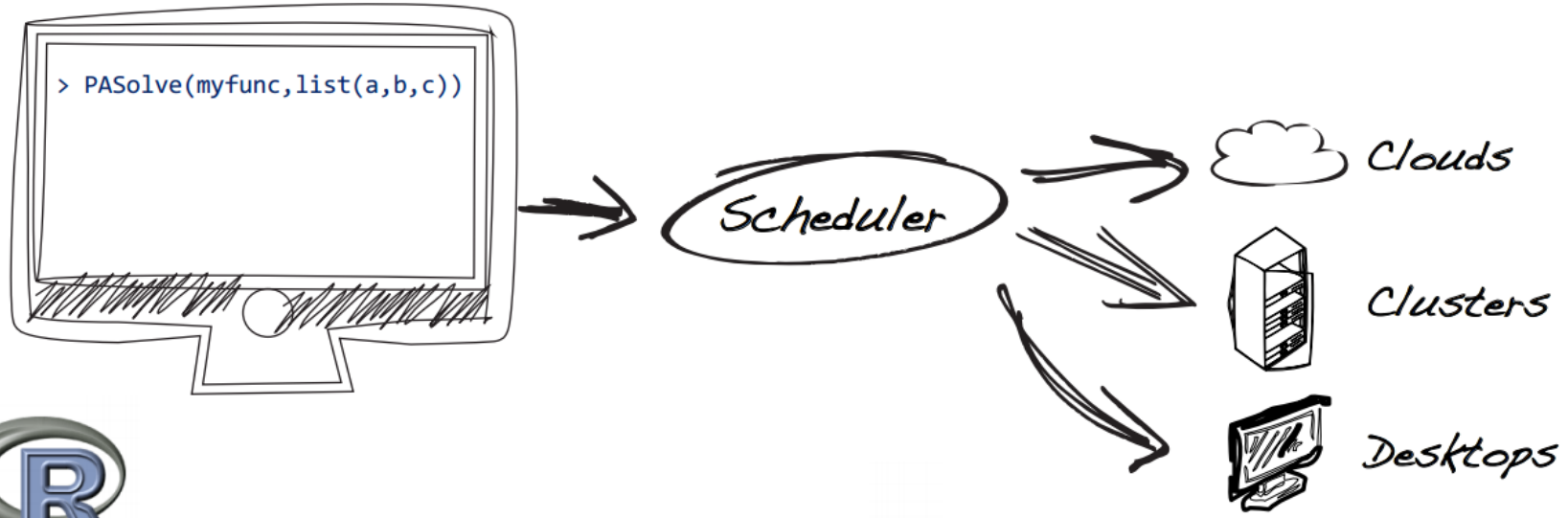






Package “PARConnector”

PAR Connector



Intégration transparente
dans votre environnement scientifique

Fonctions ProActive R

CONNEXION

PAConnect

TRANSFERTS

PAPushFile

PAPullFile

SOUMISSION

PASolve

PA

PAM(erge)

PAS(plit)

PAWaitAny

PAWaitFor

MONITORING

PAState

PAJobState

PADebug

Fonctions ProActive R

CONNEXION

PAConnect

TRANSFERTS

PAPushFile

PAPullFile

SOUMISSION

PASolve

PA

PAM(erge)

PAS(plit)

PAWaitAny

PAWaitFor

MONITORING

PAState

PAJobState

PADebug

Soumission d'un calcul

- La soumission d'un calcul s'effectue à l'aide de `PAsolve`
- PASolve permet des invocations paramétriques
  - executer une même fonction avec différentes valeurs en entrée
  - chaque exécution est une tâche indépendante
- PASolve a une syntaxe similaire à *mapply*

  `> res = PAsolve(func, arg`$_1$`, arg`$_2$`, ..., arg`$_n$`)`

  `func` peut être:
  - une fonction définie par function (x, y, ...)
  - le nom d'une fonction

L'évaluation de PASolve est asynchrone

Exemple PASolve

- Invocation paramétrique simple :

```
> job = PASolve('cos',1:5)
```

  - Crée 1 job et 5 tâches
  - Exécutes cos(1), cos(2), … , cos(5) sur 5 noeuds différents
  - Le nombre de tâches est défini par la taille max des paramètres
  - Les appels à PASolve ne bloquent pas la session R
  - `job` est un élément de substitution qui informe sur l'avancement des tâches soumises à l'exécution
    - `job` est mis à jour dynamiquement au fur et à mesure que les résults sont reçus

Récupérer des résultats

- `PAWaitFor`: permet d'attendre la fin de l'exécution des calculs distants et de récupérer les résultats

```
> res = PAWaitFor(job, timeout=1000)
```

  - Arguments optionnels :
    - `timeout` définit la période maximum d'attente
    - `callback` fonction appelée lorsque les résultats sont disponibles

Transférer des fichiers

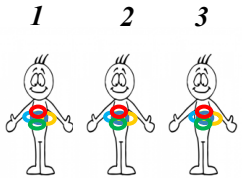
- Directement avec PASolve:
  - Arguments optionnels
    - `input.files`: liste de fichiers à transférer de la machine locale vers les noeuds distants
    - `output.files`: liste de fichiers à transférer des noeuds distants vers la machine locale
- En utilisant des fonctions dédiées:
  - `PAPushFile`
  - `PAPullFile`

Cas d’utilisation: analyse métagénomique quantitative

Pipeline de traitement

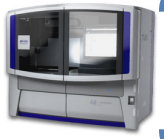
Prélèvement de l'échantillon → Séquençage de l'ADN → Alignement sur référence → Matrice de comptage → Analyses bio-{informatiques, statistiques}

Échantillons de fèces

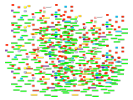


Extraction d' ADN

Sequençage de l'ADN



Courtes séquences 30-50 millions

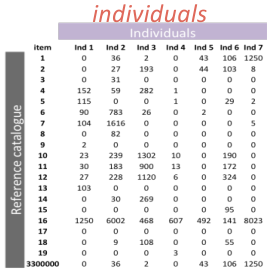


Alignement des séquences sur un catalogue de gènes

Catalogue de gènes de référence

Génomes existants

Comptage des gènes

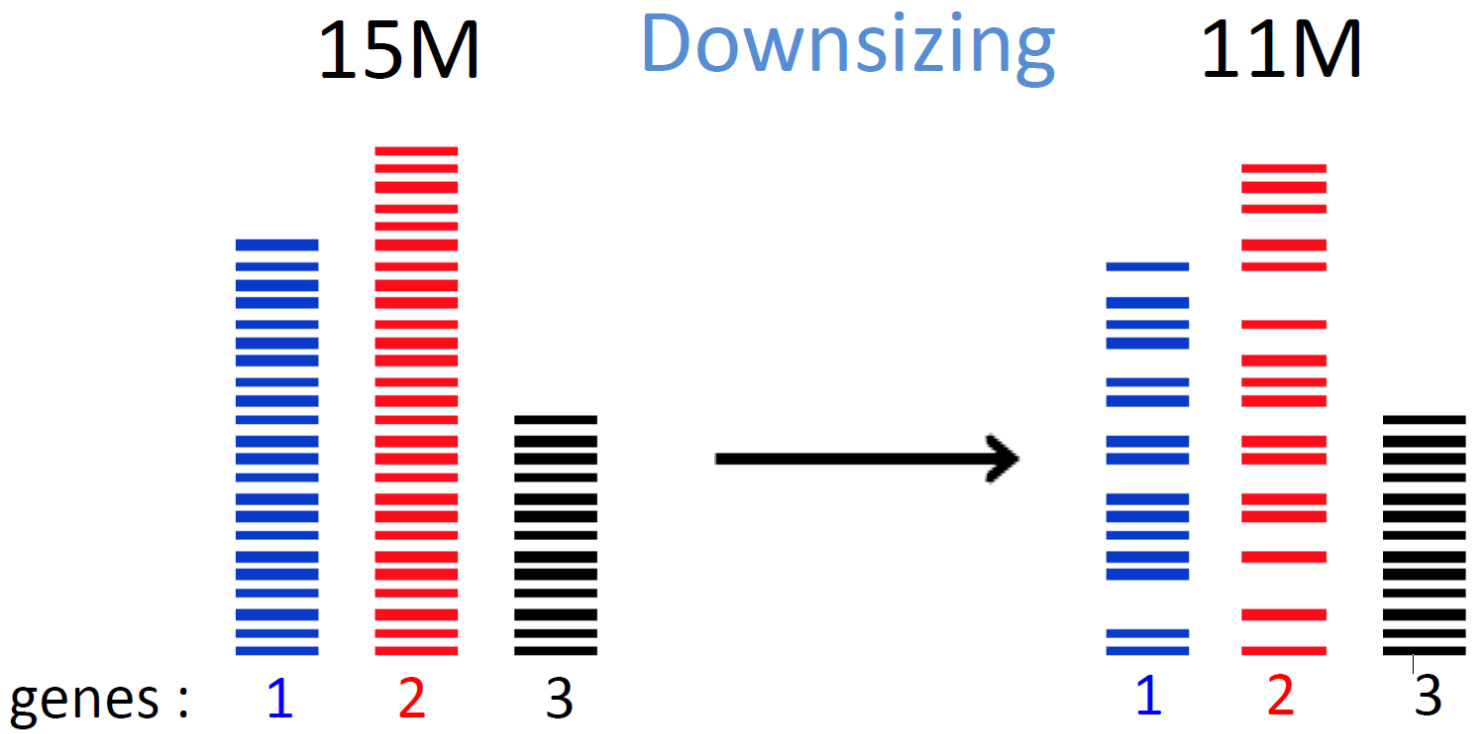


relation with clinical data

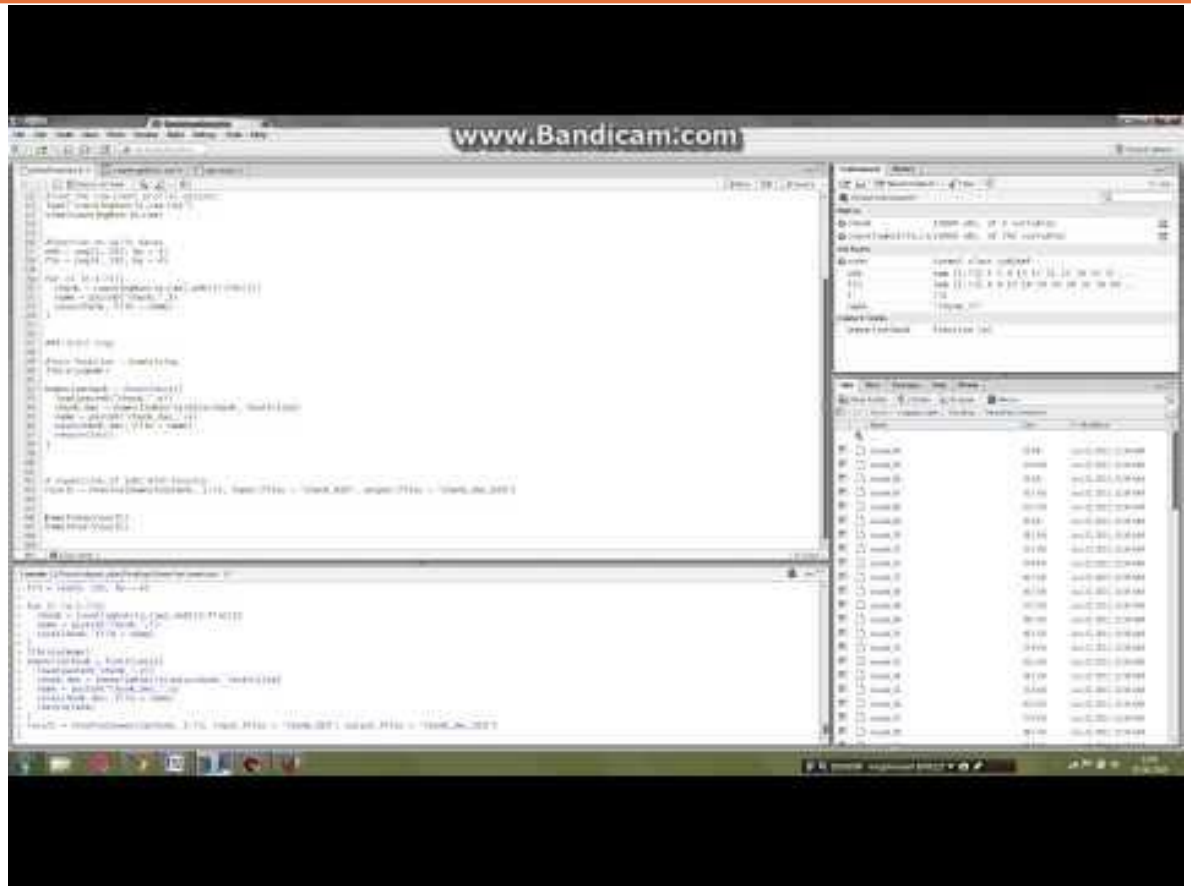
Identify clinically relevant groups

* phases HPC

Démonstration : downsizing matrix



Démo: Downsizing d'une matrice



Conclusion

PAR Connector permet de:

- Distribuer des calculs « facilement » (`PASolve`)
- Rester sur l'environnement R
- Traiter des données de type « Big Data »

Remerciements

MetaGenoPoliS

Nicolas Pons
Edi Prifti
Emmanuelle Le Chatelier
Magali Berland
Anne-Sophie Alvarez
Pierre Léonard
Amine Ghozlane
Dusko Ehrlich
Ndeye Aram Gaye
Jean-Michel Batto

ActiveEon

Brian Amedro
Youri Bonnaffé
Denis Caromel
Laurent Pellegrino
Fabien Viale


Thanks for your attention!
MACH
ACTIVEEON
SCALE BEYOND LIMITS
metagenopolis
mgps.eu