Biologie et Sciences de la vie

SdD et Bioinformatique

C Etchebest, B Villoutreix,
P Tuffery, AC Camproux

Point1: Expliquer le besoin scientifique et la dimension big data (AC Camproux)

- Big biological data (dimension big data)
- Pb spécifique du big data en biologie
- Big data bioinfomatique

Point2 : Applications interdisciplinaires sur SPC (C. Etchebest, B.Villoutreix)

- Omique
- Bioinformatique structurale: dynamique moléculaire
- Molécules chimiques
- Exemple des peptide

Points 3: Besoins en infrastructure (stockage, cpu) (P. Tuffery)

Points 6 et 7: Participation grands programmes nationaux/ Liens industriels, Européens, Internationaux

Point 5 : Laboratoires et/ou autres structures concernées (C. Etchebest)

Point 4 : Liens avec la formation

Point 1 : Scientific need and big data dimension

Big Biological Data:

- Technologies for capturing bio data are becoming cheaper and more effective (such as automated genome sequencers).

Size of a single sequenced human genome is approximately 200 gigabytes

== > New high-flow technologies in molecular biology can deliver multiple gigabytes of data /day.

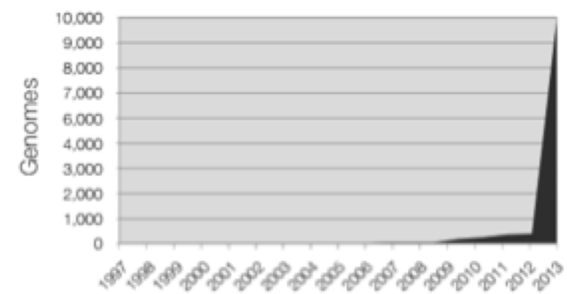
== > Increasingly accumulated large volumes of information about human, animals, plants or microbe,..

Life sciences today need more robust, computable, quantitative, accurate and precise ways to handle the big data

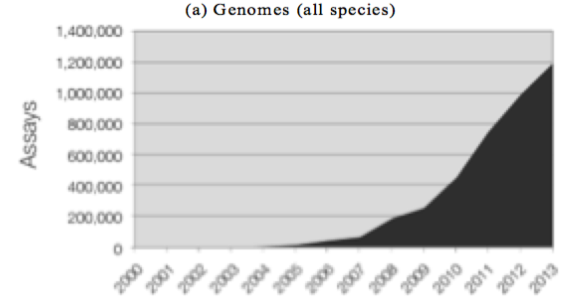
== > **central roles of bioinformatics** in the future research of the biological and biomedical fields

== > need to develop Big Data bioinformatics strategy for data management, analysis and accessibility

European Bioinformatics Institute (EBI), one of the largest biology-data repositories :
≈ 18 petabytes about genes, proteins, molecules data in 2013 versus 40 petabytes in 2014



(a) Genomes (all species)



(b) Gene expression data

Quantity of data stored by EBI over the years [8]

Total storage size doubling every year.
EBI Hinxton data center cluster: 17,000 cores and 74 terabytes of RAM

Biology Big Data specificity

Big data has 4 important features 4V's: volume of data, velocity of processing the data, variability of data sources and veracity of the data quality.

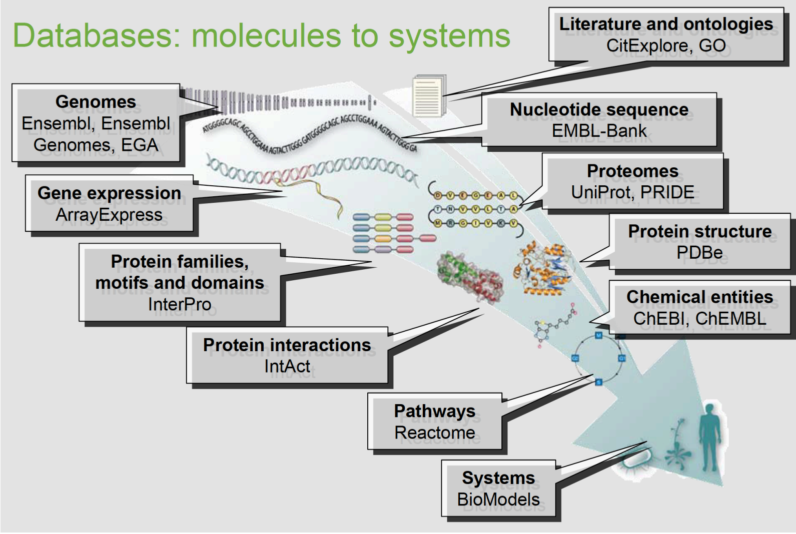
+ incremental data: new data dynamically added to the big data lake from time to time

Biological data

- **Highly heterogeneous:**

- same types of data are represented in different forms generated using different methods from genetics, physiology, pathology to imaging
- simultaneously recorded from over thousands of cells or more, replicats

EMBL database
european molecular biology laboratory



- **Complexity and hierarchy**

*generated at different levels ranging from molecules, cells, tissues to systems

*dynamics: biological processes or states change with conditions and over time

*Structured : existing intrinsic structures determined by various biological principles and/or experiment designs

Biology Big Data specificity

- **Geographically distributed:**

bioinformatics data can be **geographically distributed all over the world**.

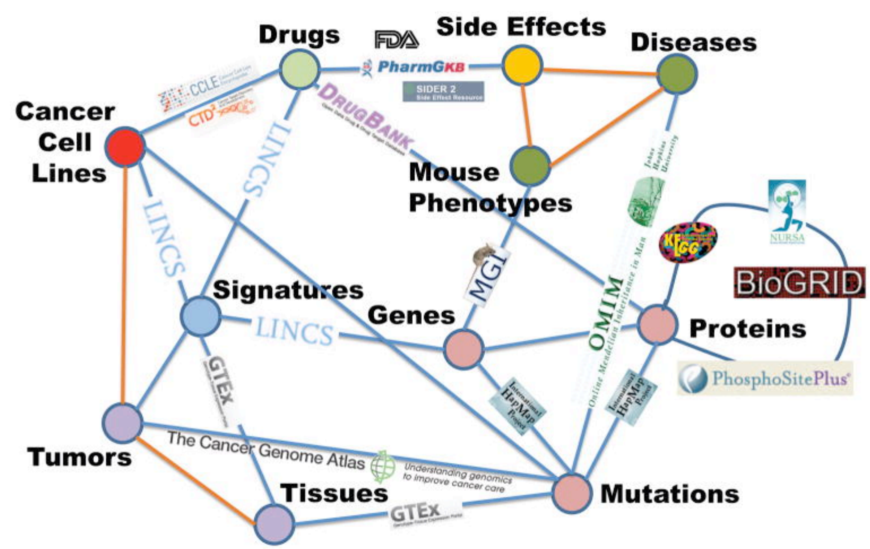
== > difficulty to transfer due to their cost, privacy and other ethical issues …

- **Hypothesis-driven study: a key for big biological data mining**

4-V features == > **association or correlation** rather than causal relationships

For deciphering the mechanisms of biological processes and diseases, need **to know causal relationship** among biological elements (genes, proteins, and pathways) which **form complex biological systems**

Hypothesis-driven study: possible ways to identify causal relationship among biological molecules



Bioinformatics analysis

Bioinformatics traditionally organized and developed around 4 skills:

- Genomic bioinformatics: DNA and proteins expertise (genetics, genomics, transcriptomics, metagenomics and métagénétique)

- Structural bioinformatics: modeling of molecules and macromolecules (drug design, proteomics, immunology)

- Bioinformatics image: apply the methods of signal processing in medical and biological imaging (CT / MRI, microscopy, microarray)

- Biostatistics : process needs in biology statistics (population genetics, toxicology, etc.)

==> curation of data in current bioinformatics analysis : 60% timing work

Big Data Bioinformatic strategy

New era of bioinformatics for big data

- Management, curation and connection of biological big data

- Integration, comparison and relationship
- To issue new hypotheses to produce new models and compare them to experimentation

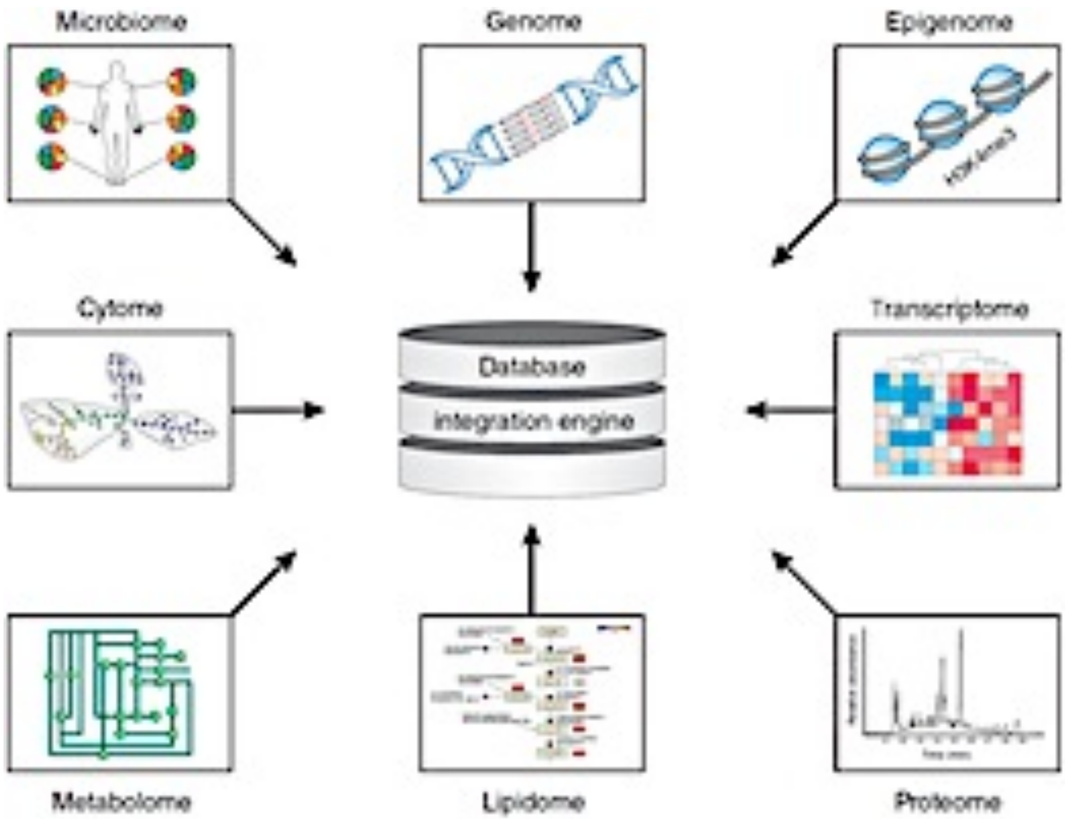
Adapted algorithms and methods fast, large scale, distributed, optimised for iterative and complexe bioinformatics problems but also fault tolerant, robust to missing data, unlabeled data, redundancy, variables selection…

== > Unsupervised and supervised machine learning methods/ Graph Theory

== > Computational systems biology: to understand essential mechanisms of biological systems

Possibilités d'applications interdisciplinaires connues

From « Omics » to System Biology



petabyte (PB) even exabyte (EB).

Questions and Challenges:

- Sequence Analysis
- Microarrays Analysis: Adding Time variable=> Dynamics
- Gene-gene network Analysis: a complex and highly iterative problem
- Protein-Protein Network Analysis
- Evolutionary research:
- Pathway analysis
- Disease network analysis

MAIN CHALLENGE: from association study to causality study.

« Omics » and System Biology

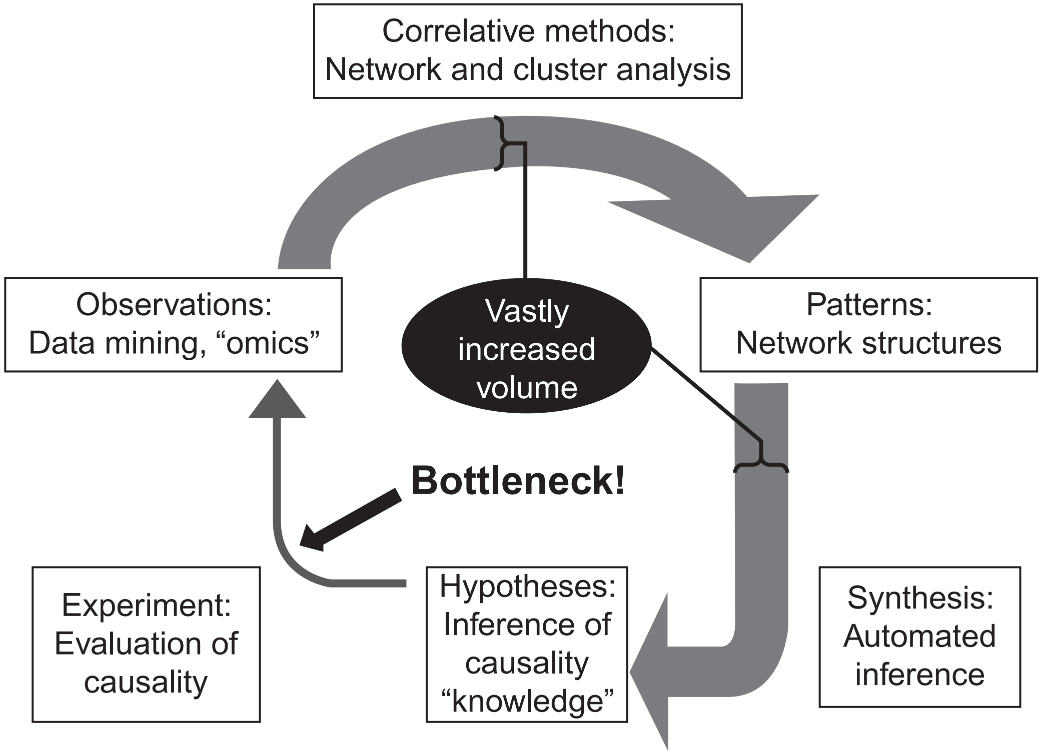
- New technologies and methods for analyses: Hardware and Software.
  - Many new companies:

  (e.g. **BioDatomics,**

  400 tools for analyzing

  genomic data running on a cluster)

- Data driven Hypothesis? :
  - Example: Deep learning methods



Applications in Structural Bioinformatics Molecular Dynamics Simulations

- Study of systems behaviour as a function of time:
  - Example:
    - Conformational Transitions
    - Transport Mechanisms
    - Recognition Mechanisms : protein-protein, proten-ligand.
    - Folding of Biological macromolecules

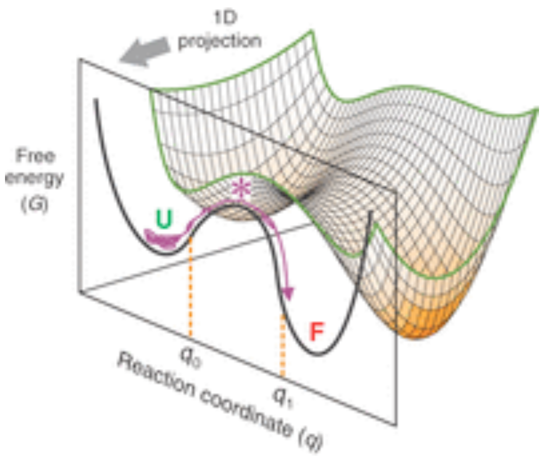
- Statistical Sampling :
  - Example:
    - Evaluation of thermodynamical quantities
    - Refinement of structures obtained from biophysical data: NMR, X-Ray

Stakes, challenges and bottlenecks

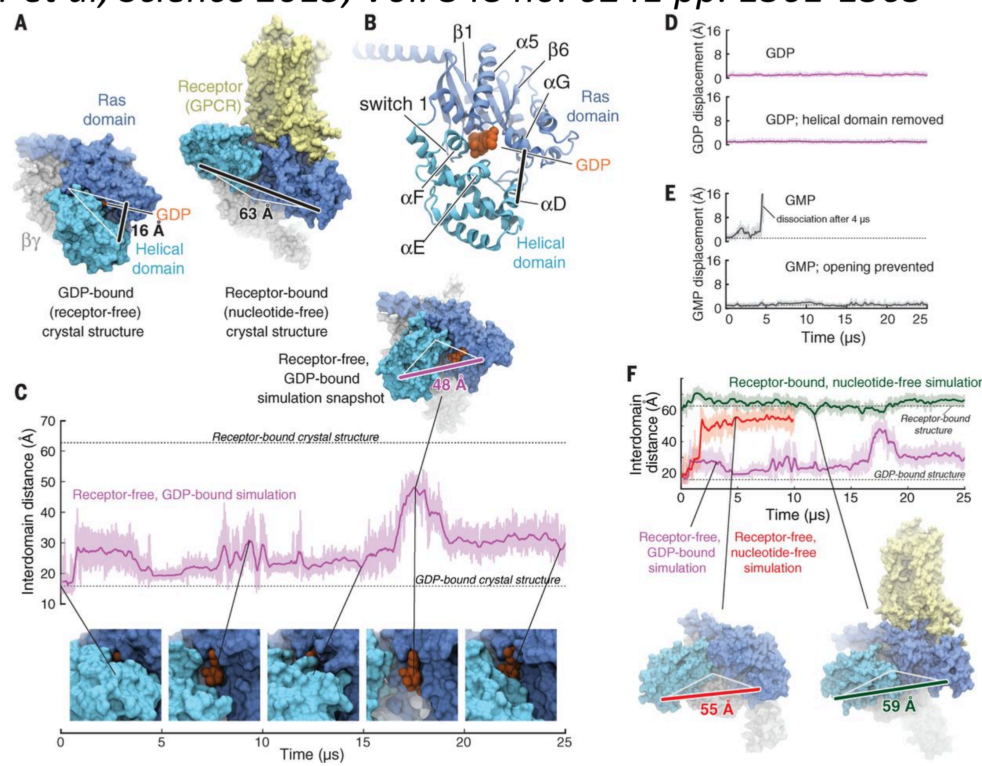
- Challenges: Examples

Elucidation of functional mechanisms of the most important drug targets: GPCR receptor, G protein and mechanisms of nucleotide release through internal rearrangement of G protein

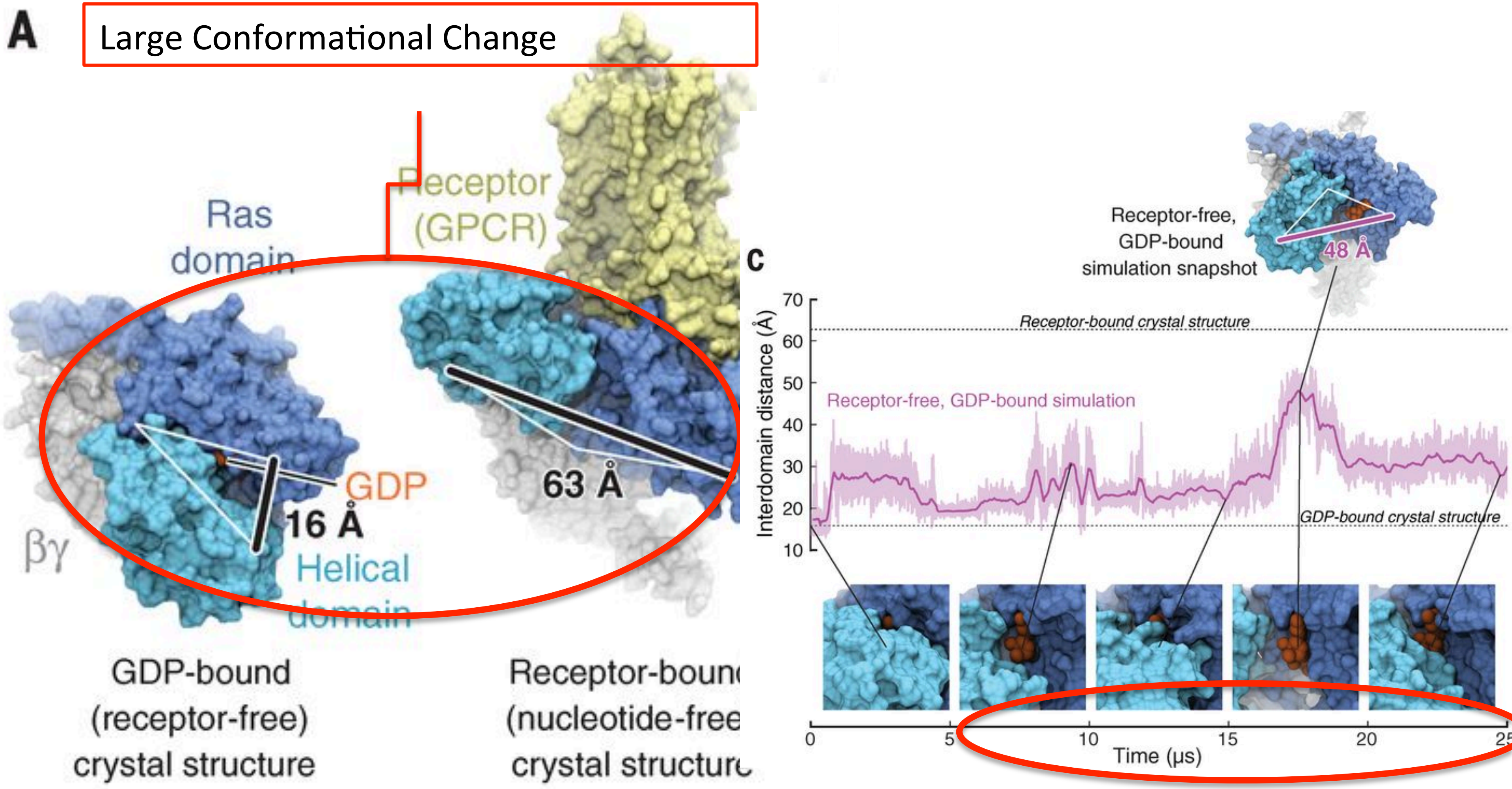
Dror et al, Science 2015, Vol. 348 no. 6241 pp. 1361-1365



Structural origin of slow diffusion in protein folding:
Chung et al., *Science 2015, Vol. 349 no. 6255 pp. 1504-1510*



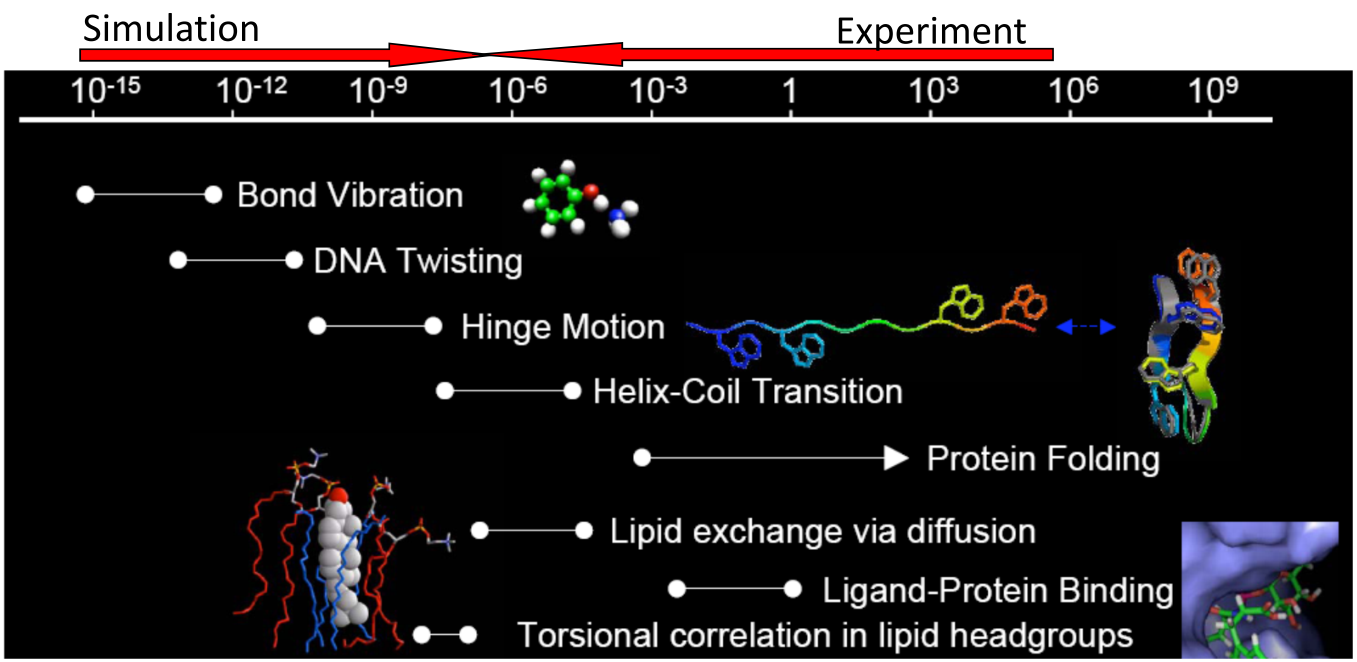
Large Conformational Change



Time Scale

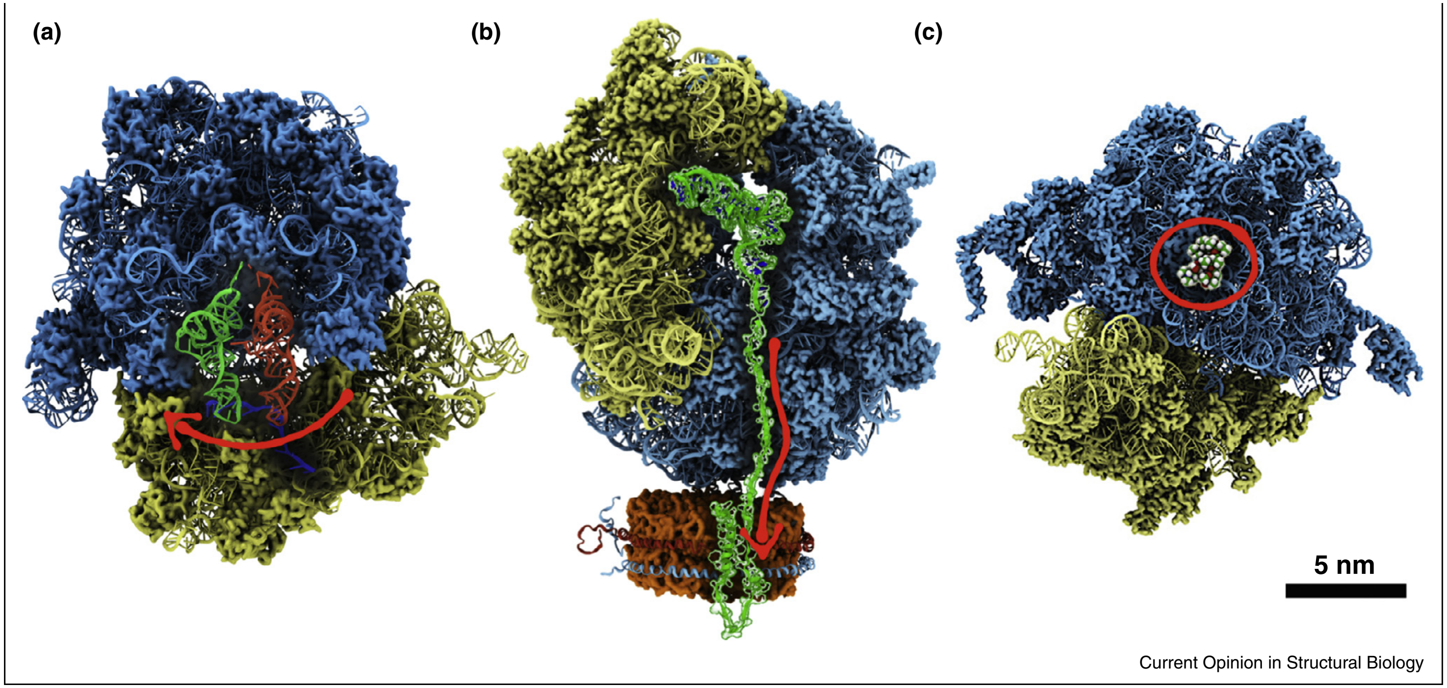
Stakes, challenges and bottlenecks

- Time Scale in biology: μs, **ms, s**
  - Integration Step: **fs**. => 10[9-15] calculations



Stakes, challenges and bottlenecks

- System Size: billions of interacting particules .



(a) Translocating ribosome at the pretranslocation state with an A-site tRNA (red) and a P-site tRNA (green) [49••]. A red arrow shows the direction of tRNA's traversal motion. **(b)** Insertion of a nascent protein by the ribosome into a nanodisc [50] membrane working with the SecYE translocon [51]. The nascent protein and P-site tRNA are shown in green. A red arrow shows the direction of the nascent protein's insertion motion. C. Bacterial ribosome with the antibiotic drug *erythromycin* (in red circle) shown at its binding site inside the ribosome [16•].

Stakes, challenges and bottlenecks

- Data Storage:
  - Obligation to neglect time steps that could be crucial for understanding mechanisms
  - Obligation to neglect some particles.

- Data analysis :
  - Determination of dynamical interaction networks
  - Transition Pathways

- Switching between scales: (QM/MM/CG)

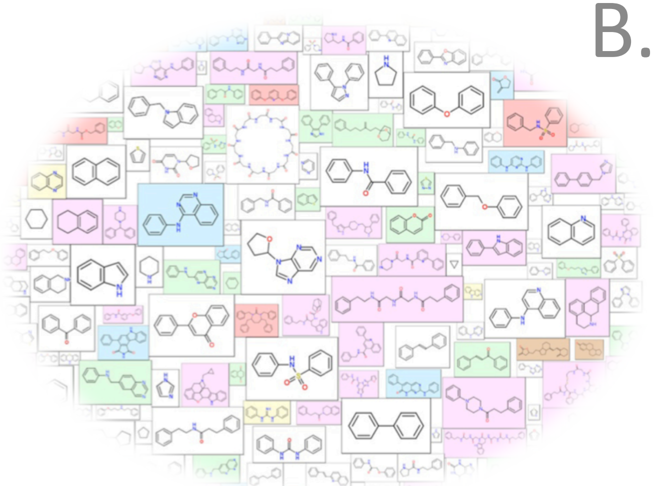
NEEDS for NEW ANALYSIS PARADIGMS

Possibilités d'applications interdisciplinaires sur SPC

““Big data”” in the area of “<u>chemistry - drugs</u>”: overview

B. Villoutreix, MTi

UMRS 973



USPC

Nov 2015

Epochs in the field

- **Empirical – up until 1960's**
- **"Rational" – 1960's to 1990's: "lock & key"**
- **Big Experiment – 1990's to 2000's**
  – High throughput screening…Human genome
- **Big Data – 2010's onwards**
  – Informatics-driven drug discovery and biology
  – Diseases are more complex than anticipated

Big Data in the public domain: 28 nov 2015

Some data in chemistry/drugs - biochemistry, e.g.,:

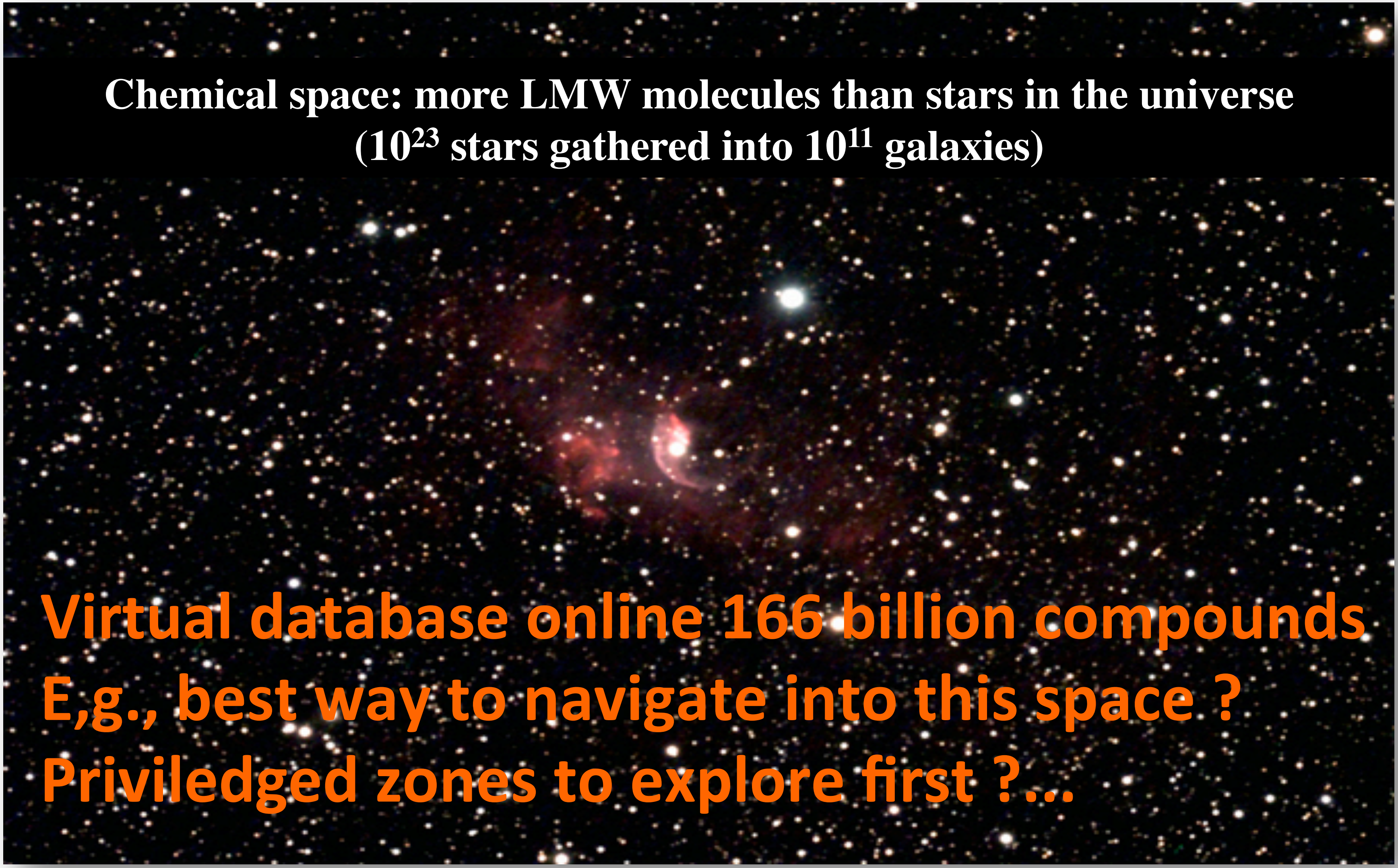
- **Pubchem**

  Compounds: 61,025,551; Tested Compounds: 2,091,562; Protein Targets: 9,954
- **PDB**: 113,971
- **Chembl**

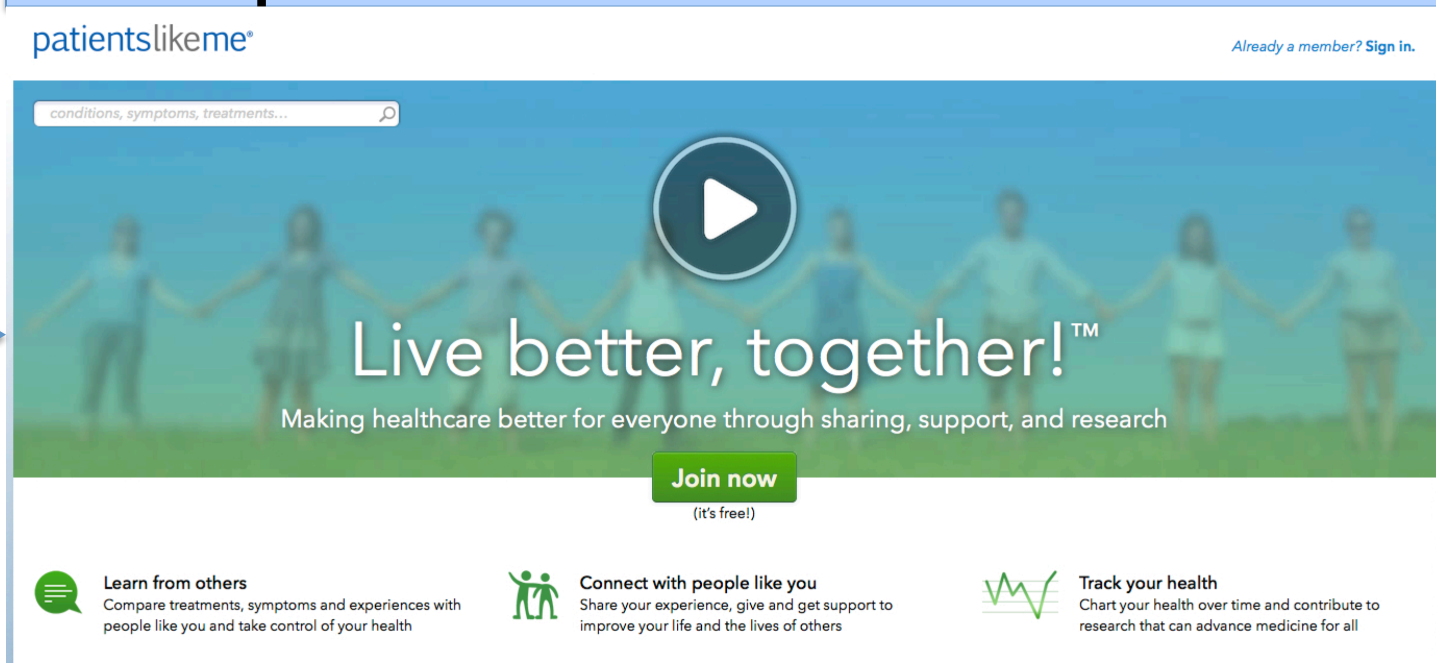
  Compound records: 1,715,667; Activities: 13,520,737

Critical: **relationships between these entities (cmpds, genes, targets..)**

Big Data in the public domain:
28 nov 2015



Involving patients, a lot of additional data, eg: 350,000 people, 28 million data points about disease… idem in EU



+ Literature Extraction…..
+personalized medicine…

Some challenges

- Data storage, data curation, data integration, data sharing, data processing, data visualization…

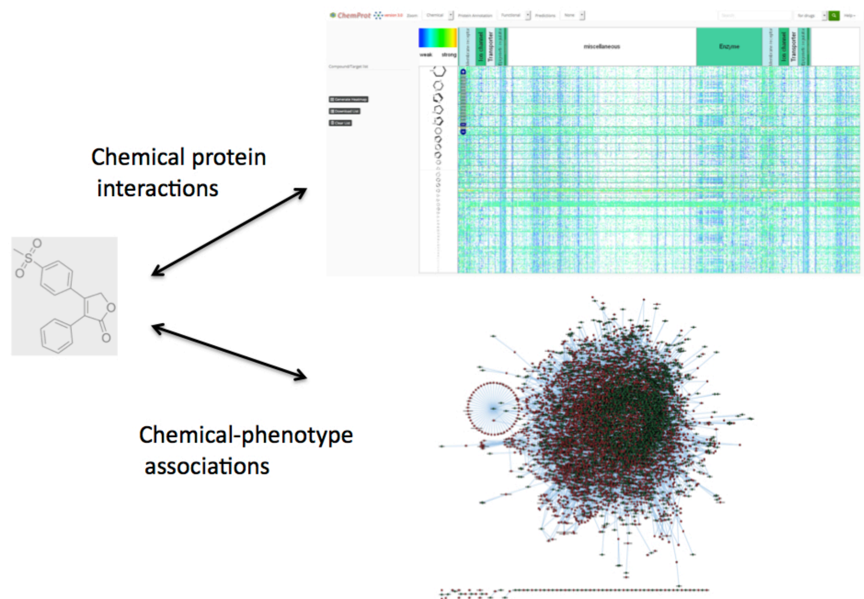
e.g., Data-driven decision making to assist drug discovery…
New drugs for new diseases…

USPC “chemistry-drugs”: some examples

- One project that could be done: Large scale 3D Virtual ligand screening in MTi + biostat + chemoinfo, tools and skills are present – “forces” in 3D in USPC → help to design new drugs or understand targets…

- One ongoing project: the Chemprot database (can be linked to the project above)

O Taboureau et al

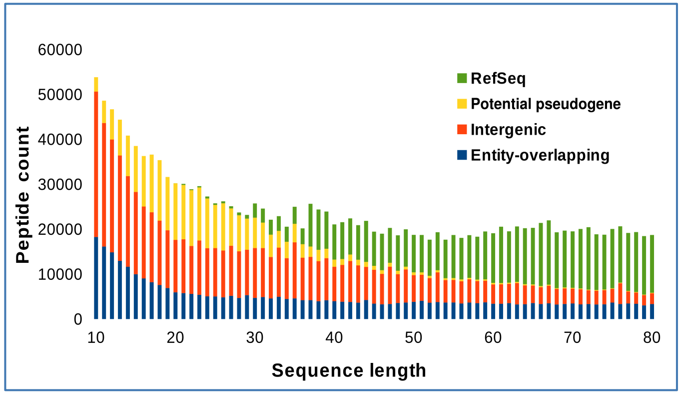


- “Chemistry-drugs data generated in USPC:…. ADMET, adverse drug reactions..etc, etc… by chemists, biochemists, clinicians… integrated with the help of computer scientists, mathematicians, bio-chemoinfo teams….” This will have to be coordinated and linked to national and international projects

USPC : Prokaryote candidate peptidome
Bactpepdb.rpbs.univ-paris-diderot.fr

- Identify candidate peptides from large survey: Built over **re-analysis of all prokaryote genomes** for short ORFs (+ RBS) 10-80 amino-acids. Over 2,000,000 candidate peptides from over 2300 genomes.

Some numbers about the database :



- 557 genera
- 1252 species
- 2369 strains
- 1598 plasmids
- 1,834,816 peptides
  - 263,000 with SS bonds
  - 173,000 with TM segments
  - 30,000 with signal peptide
  - 112,000 are conserved across families of an order

- **Challenges**: Large scale 3D modelling, large scale prediction of target-peptide interactions, peptide sequence optimisation for better affinity/specificity.

Infrastructures for structural bioinformatics

PAAS SdV (2016-2018)

IA Infrastructures

RPBS co-coordinates deployments for Structural bioinformatics (with Nancy)

BipBip: Bayesian Inference Paradigm Biology in Processors

IA Bioinformatique

(13 french teams)

HIRAM (H 2020)

France Grille

Genci CT7

Protein (comparative / de novo) modeling, complex modeling, virtual screening (chemicals, peptides)

Needs for infrastructures (storage, cpu)

- **Warning:** Big data centralized storage area NOT EFFICIENT due to network bandwidth limitations.
- **Warning:** better to FAVOR VERSATILE CALCULATION RESSOURCE. Bioinformatics mixes parallel, distributed, sequential calculation, possibly requiring specific banks => heterogeneous resource (big memory nodes, GPU nodes, manycores nodes, …)
- **Warning:** evolution of methods is rapid => need for FLEXIBLE DEPLOYMENT SCHEMES. (E.g. France Grille, Genci do not make easy to deploy own softwares)
- **PAAS:** Docker / slurmm / lustrefs, presently ~960 cores, 3x30 Tb storage.

(e.g. calculation of molecular descriptors for over 20,000,000 compounds took 2 months)

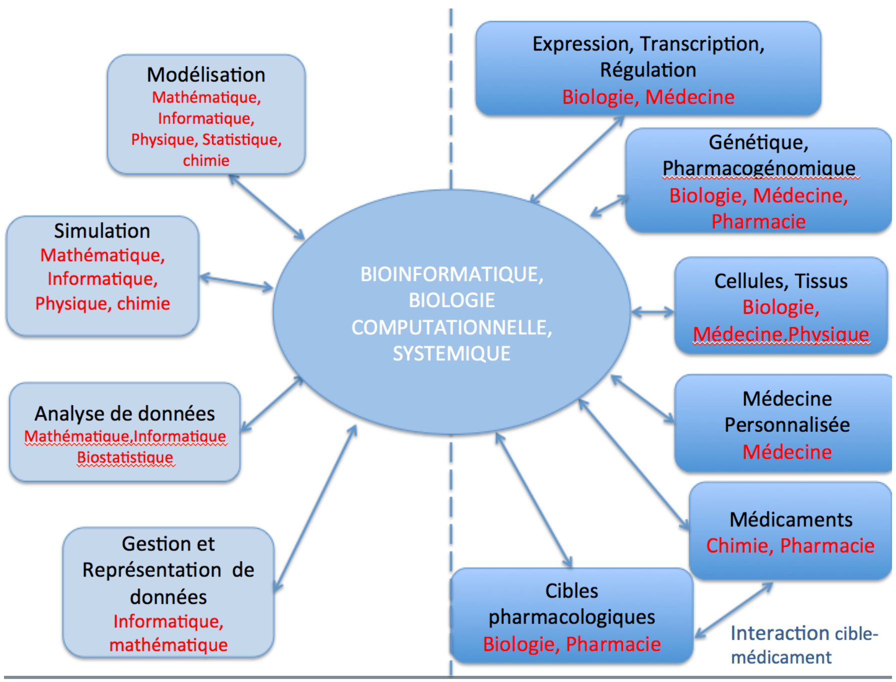
- **Big Data projections:**
  - Increase size of calculation resource (e.g. x 3)
  - Favor GPU when possible (e.g. MD / Gromacs)
  - Expected need for storage (modelome, interactome, screening collections, dynameome,…) up to several hundreds To.

Bioinformatics in SPC

- 88 teams ou people:

From methodologies to biology and medical sciences

- Well-identified bioinformatics teams
- Strengths as well for methodological aspects as applications
- Multidisciplinary collaborations



- Plate-Forme

Extension Big data

Teaching in Bioinformatics in SPC And Science of Data

Programming, Machine Learning and Biostatistics

Two Masters : Strong expertise in Structural Bioinformatics

- Biology-Informatics/Bioinformatics: strong participation of P7 Informatic department
- In Silico Drug Design « IsDD »: link with chemistry department

DU Bioinformatics: to be renewed

10 FC: modules

Modules in Doctoral Schools

- Merci de votre attention

Big Data Bioinformatic strategie

- **Big Biological Data:** remarkable example by Yuan and her colleagues [5].

They found that very diverse outputs are often generated when the same gene expression data is analyzed using different algorithms, *i.e.*, low overlap and substantial false positives. The problem results from the extreme heterogeneousness of gene expression data and there is no guarantee that a pure statistical model will solve it.

A recent effort was made to present a methodology, aimed to circumvent the limitations of pure statistical models and general gene expression data analysis strategy.

The method was based on a simple biological assumption: ''If a number of genes that are conservatively co-expressed emerge as a dynamically-cooperative group across certain biological processes, these genes are most likely functionally closely related with physiological and pathological processes'' [5].

Then, according to this ''hypothesis'', the data mining is just to be converted to finding those gene clusters with strongly cooperative and conservative properties across cancer progression stages