

Red española de e-Ciencia

Bioinformatics Applications in the Spanish Network for e-Science

Ignacio Blanquer on behalf of the Spanish
Network for e-Science Biomed Community

Acción financiada por:



GOBIERNO
DE ESPAÑA

MINISTERIO
DE CIENCIA
E INNOVACION

Entidad Coordinadora:



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

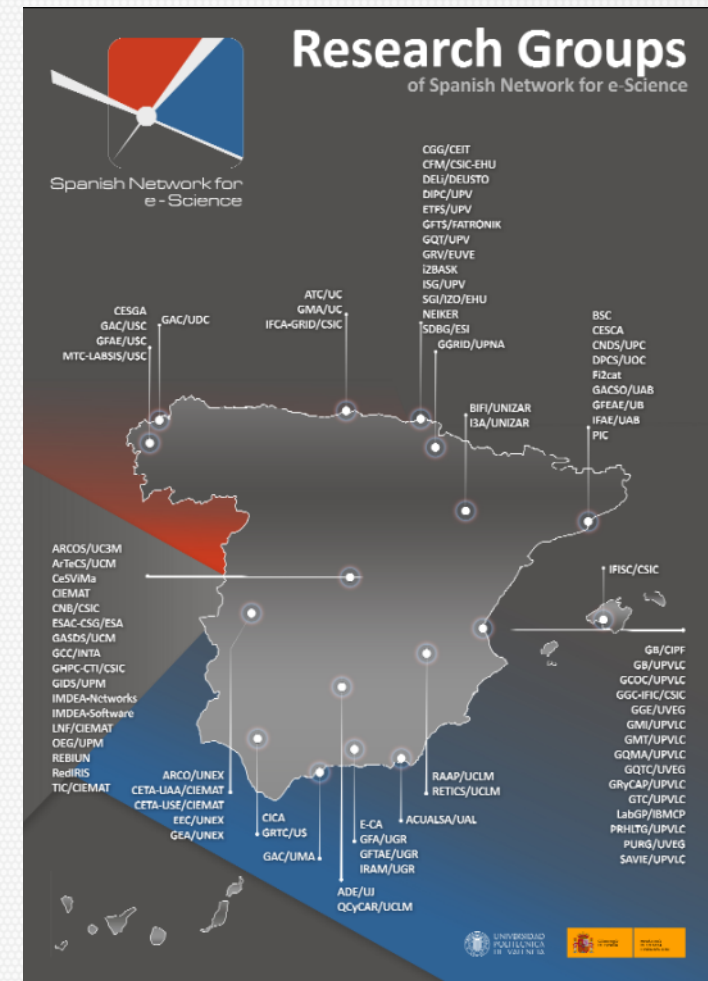
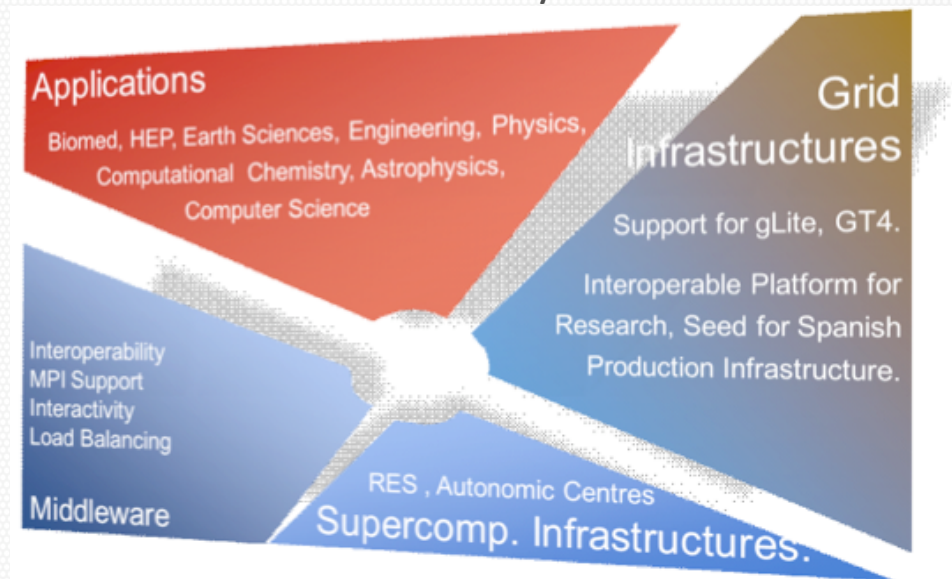
Outline



- The Spanish Network for e-Science
 - Structure and link with the Spanish NGI.
- Biomedical applications in the Spanish Network for e-Science.
- Challenges for Biomedicine on the Grid.

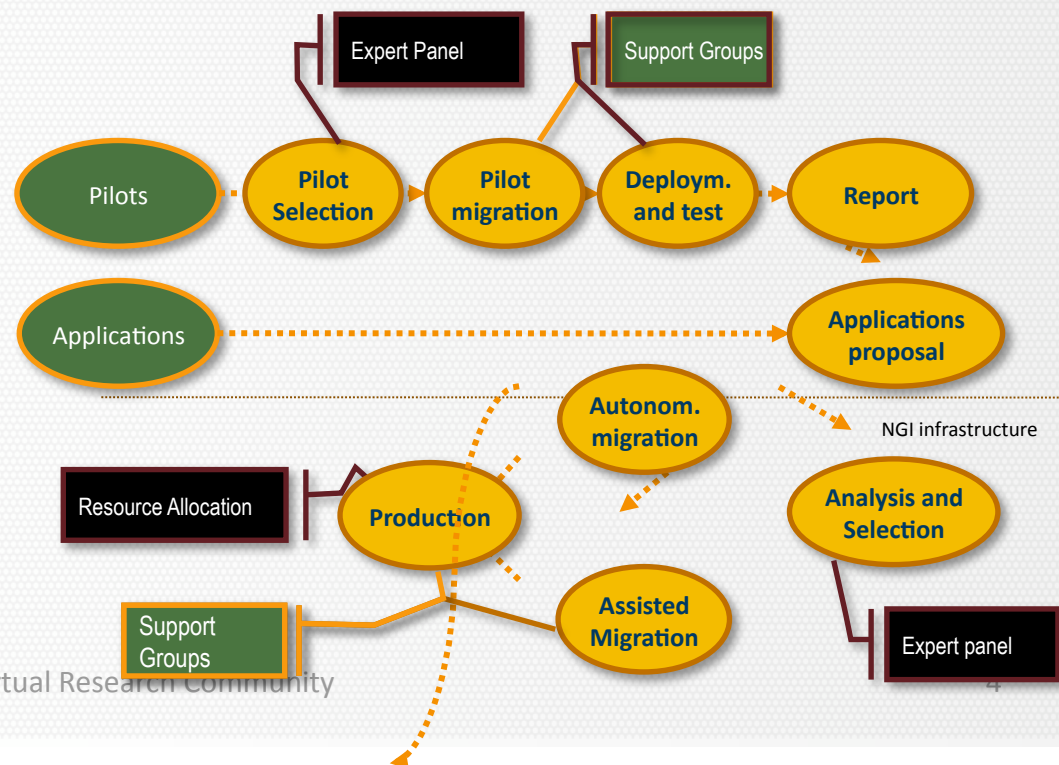
Structure and Participants

- More than 50 different institutions and 97 Groups.
- More than 1000 researchers.
- Dynamic Structure
 - 28 Groups have been incorporated activity.
- Structured in Four Activity Areas



Applications

- 3 Roles are identified
 - Mature applications aiming at a challenging experiment.
 - Pilots that require intensive porting and a feasibility study.
 - Support groups with experience on porting applications.
- Pilots, Applications and Support Groups are certified by an expert board.
- An internal call for projects was set up.



Overview of the Bioinformatics Applications



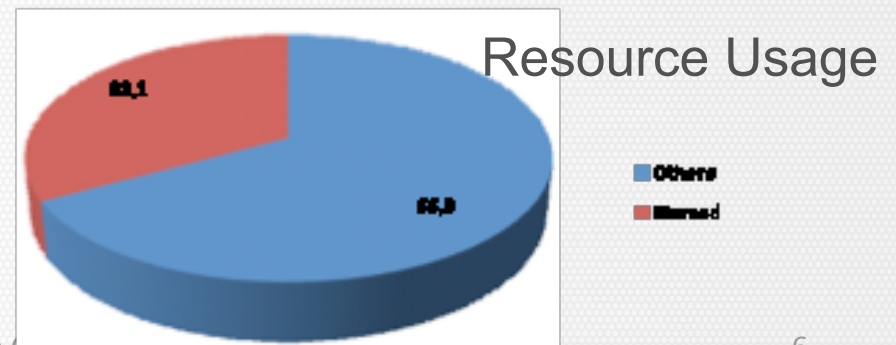
- Consolidated Use
 - Work on current bioinformatic databases to analyse quality, improve annotation or increase the usability
 - Alignment and Annotation in large databases and metagenomes.
 - CD-HIT.
 - Phylogenetic analysis.
- Emerging Use
 - Port new applications on the Grid for providing new services
 - Docking (Gfrodock).
 - Assembly (G-MIRA).
 - Transnational sharing of Medical Images on the area of Mammography.



Current Status



- 4 Projects already have a VO created (vo.odthpiv.es-ngi.eu, vo.blast.es-ngi.eu, vo.filogen.es-ngi.eu and vo.frodock.es-ngi.eu).
 - Currently on the process of consolidating a single vo for life sciences (already in place).
- 3 Projects (GBLAST, FILOGEN, and g-MIRA), have been granted with resources for porting through an internal project call.
- 33% of the resources have been consumed by the biomed applications.

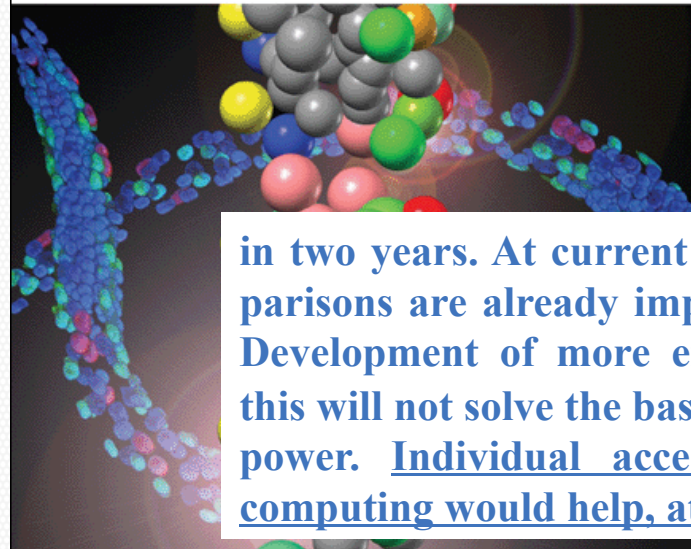


The perception of the need for resources is clear

September 2009 | volume 6 | number 9

nature | **methods**

www.nature.com/naturemethods Techniques for life scientists and chemists



in two years. At current database sizes all-versus-all comparisons are already impossible without a supercomputer. Development of more efficient algorithms will help, but this will not solve the basic problem of too little computing power. Individual access to supercomputers or cloud computing would help, at least temporarily.

- A digital atlas of the worm
- Tools for metagenomics
- A reporter line for erythroid differentiation
- BreakDancer
- Addressing the crystallography phase problem

EDITORIAL

Metagenomics versus Moore's law

Metagenomics sprang from advances in sequencing technology, and continued improvements are providing data in quantities unimaginable a few years ago. But without concerted efforts, the amount of data will quickly outpace the ability of scientists to analyze it.

As Craig Venter sails the oceans collecting seawater samples to profile microbial communities by high-throughput sequence analysis, microbiologists around the world are busy collecting their own samples. The diversity of locations—from Antarctic lakes to human armpits—highlights the reality that microscopic organisms represent a significant fraction of the Earth's ecosystem.

Any population this large is certain to have profound influences on its environment. Yet our knowledge of

40 megabases. Today there are over 4,000 sequenced metagenomes, and their size and number are increasing. Each new pyrosequenced metagenome is 200–500 megabases, and those generated on Illumina platforms are 20–50 gigabases. To analyze these metagenomes using established pipelines would take tens of years on a single processor and weeks to months on machines with up to 1,000 processors. The rate of increase in sequence generation is far outpacing Moore's law, and the cost of analyzing the largest datasets

ly gene
enomes
ting this
n genes.
n genes
all com-
puter.
elp, but
comput-
or cloud

o avoid
available
(subvir-
ic anno-
t it does
to hap-
biology
set sizes
sed sup-

about 50% for pyrosequencing and 90% for Illumina and SOLiD platforms. Most metagenomics analysis pipelines are designed for Sanger sequencing data, so the short read lengths and error profiles of the new methods present challenges for data analysis and interpretation.

Reports on pages 639 and 673 and an accompanying News and Views on page 636 illustrate some of the dangers and challenges involved and describe new algorithms to deal with them. More work is needed to assess the new technologies and develop optimized analysis pipelines, and these efforts are well underway.

But even as these problems are being solved, a larger problem has taken the community off-guard: the exponentially increasing amount of sequence data. Just over three years ago, the first two second-generation sequencing platform-based shotgun metagenomes were reported—each less than

port for data analysis. Secondly, the community needs to decrease computational demands by improving data sharing through standards and centralized coordination and by aggregating computationally intensive operations.

This summer, after discussions at the International Conference on Systems for Intelligent Molecular Biology, community members formed the M5 (metagenomics, metadata, metanalysis, multiscale-models and meta-infrastructure) Consortium under the roof of the Genomics Standards Consortium to devise a solution to the coming gridlock. Their proposed 'M5 Platform'—to be announced later this year—deserves the support of the community, funding agencies and those who hold the keys to the high-performance computing centers. Unless major efforts are taken immediately, researchers will find they have a wealth of data but no way to interpret it.

Need for an European Grid-related interaction



- Do we need a European structure for grid-related Health and LS communities?
 - Regarding resources, it will be necessary for supporting computational peaks that are being faced. Sharing resources with LHC users could be complicated.
 - Regarding cooperation, there is a clear need for avoiding replicating efforts
 - The Bioinformatics community, for example, is highly coordinated and open.
 - The bottleneck is clearly data, and the future is uncertain with NGS and Medical Imaging.
 - At the end of 2010 it will use the 30% of world global storage¹.
 - Costs for processing will be 20 times larger than sequencing².

¹ Fact sheet: Information Storage Trends, IBM.

² Wilkening, et al. IEEE Cluster 09.

Need for an European Grid-related interaction



- What would be its role ?
 - Main point is coordination.
 - Coordination of activities, negotiation of resources, supervision of services, etc.
- What would be the supporting infrastructures ?
 - Based on the Regional level.
 - Regarding computing, In the Spanish and Portuguese NGIs, a general VO for biomed (life.vo.ibergrid.eu) has been created to support Spanish and Portuguese activities.
 - This will ease keeping the support to the international vo.
 - Regarding coordination, in Spain there is a global Network for e-Science (on top of Supercomputing and Grids)
 - Along with many other vertical scientific associations in Health and Bioinformatics.

Need for an European Grid-related interaction



- Who would be the targeted user communities?
 - In Spain and Portugal:
 - Research Centres in Life Sciences (CNIO, CIB, IBMCP, CSISP, CIPF)
 - They already have computing resources but facing new challenges in data and computing.
 - IT Research centres and Universities developing tools in biomedicine (UPV-I3M, CETA-CIEMAT, TIC-CIEMAT, I3A)
 - Developers of applications.
- How would it be organized and how would it interact with the existing initiatives?
 - There is a need for a common international place: HealthGrid is perceived as a good structure.