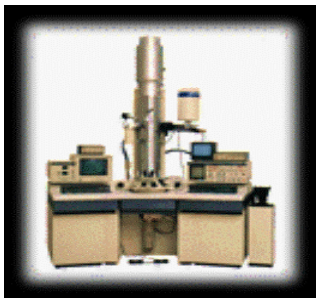
Bringing 3D-EM to the Grid

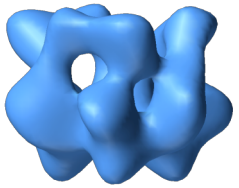
Geneva September 2006

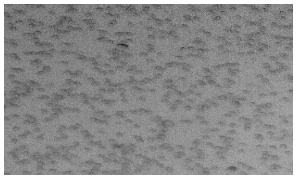
Germán Carrera
David García
Alfredo Solano
José Ramón Valverde
José M. Carazo

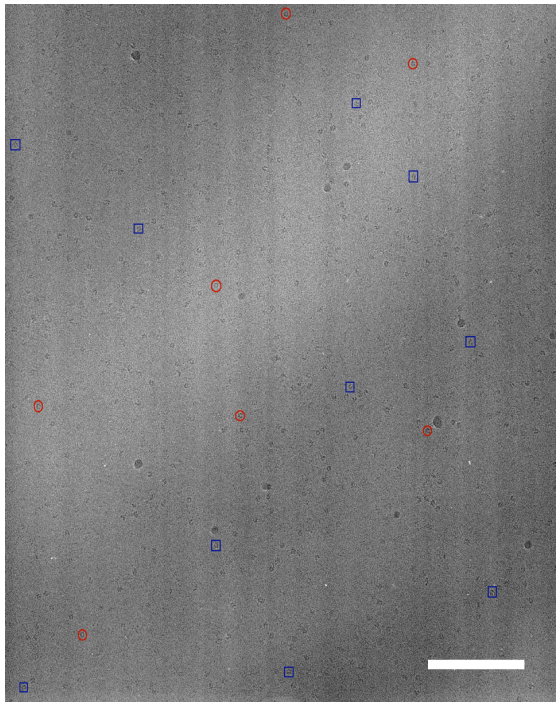
What is 3DE: 3D-EM in context (I)

SIMILAR TO A BIOMEDICAL SCAN; BUT AT THE NANOMETER SCALE: It studies "macromolecular nanomachines"



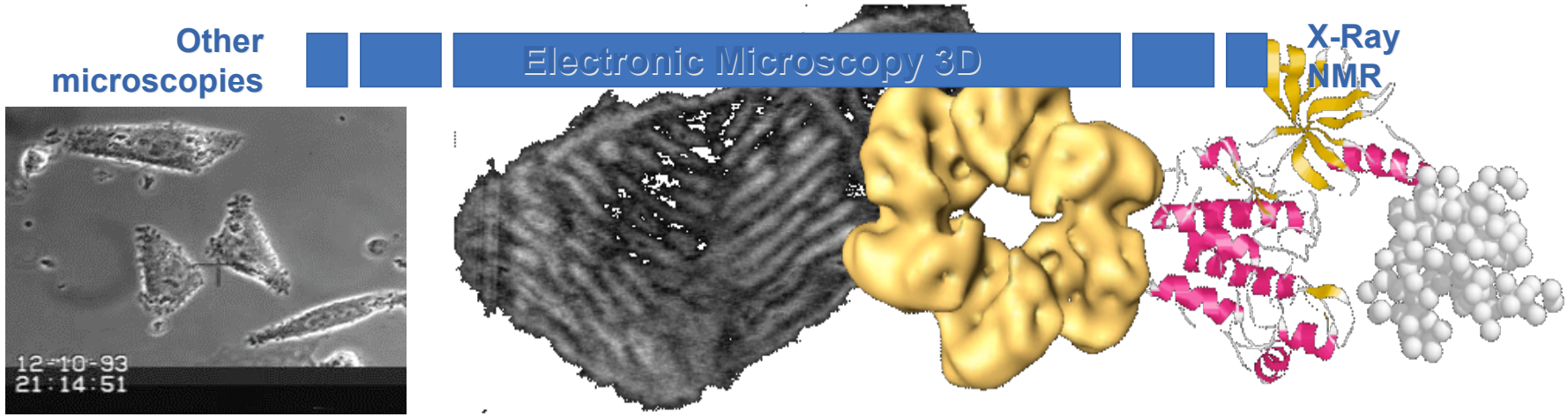


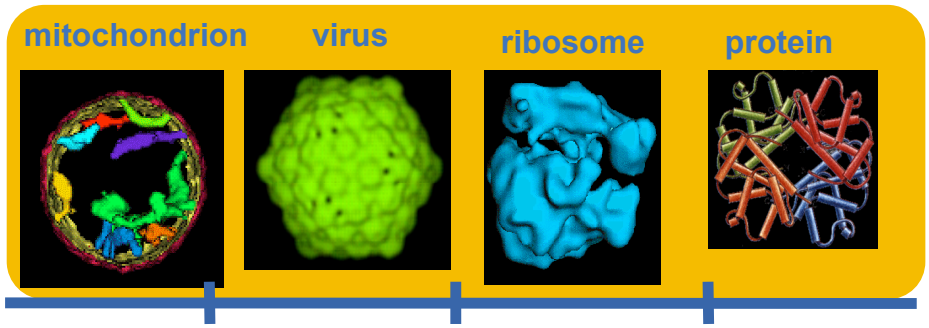




- pH 7.6
- 0.1 mM ATP

Which are the specimen: 3D-EM in context (II), nanomachines





3DEM: Opportunities for parallelization

- **3D EM is a complex research field of key biological importance with applications that offer various opportunities for parallelization**
  - Large data sets
    - might be processed independently
    - might be distributed for efficiency
  - Some steps are highly computer intensive
  - Various approaches are possible
    - Networks Of Workstations ⎫
    - Clusters ⎬
    - Supercomputing ⎬
    - Grid ⎭

… BUT, it also posed serious challenges for Grid porting.

Some of the challenges to the grid are:

1 Core Basic Grid Problem: High volume of data, we need a strategy to publish all the information efficiently

Typically we have to distribute 100.000 images!!!!

2 Applications with intrinsic data locality needs: How to optimize this in the Grid?

3 In summary: Need to pay close attention to efficient data distribution in the Grid. How to better handle the Resource Broker for it?

A 3DEM application case study, ML2D and 3D

- Classification into structurally homogeneous subsets is a complicated problem, because:
- the data are very noisy (SNR<1/10!)
- the projection directions of the experimental images are not known
- Our apporach to this problem is a Maximum Likelihood based refinement protocol of multiple 3D-references

Computational cost: Previous experience

- The classification of 91,000 ribosome projections into 4 classes took more than 2500 CPU hours using the resources of the MareNostrum supercomputer at the Barcelona Supercomuting Center.

- As not all groups may have access to such supercomputing resources

On the bases of its high computational cost and key biological relevance WE PROPOSE THE USE OF THE GRID FOR THIS APPLICATION

In collaboration with the Network of Excellence in 3DEM

Tests so far

- “SMALL” (2 clock days) RUN OF ML2D were launched on the Grid
- using 3 different resource brokers
- using 3 different storage elements
- BUT
- 5% of the submissions never started
- a few jobs run forever
- when an unreliable storage element was used, files could not be deleted or overwritten…
- THERE IS STILL ROOM FOR IMPROVEMENT IN THE MIDDLEWARE.

But...

- **But...**
  - Better strategies are possible
    - **Pre-publish (store data and software replicated directly on the Grid)**
    - Better ways to handle “iterative” procedures (better ways to maximize new data distribution in the context of synchronizing the Data and Processing Elements for the next iteration)
    - **“The total is more than the collection of the individual components” (i.e.; the real application should be more than blindly sending again one iteration after the other”)**

Still a major conceptual challenge

Make it transparent to the biological user

Put bluntly

- **Users need to work with large data sets**
  - Efficient publishing, paying attention to replication and data locality
- **We need to harmonize local work and Grid work**
  - A typical user Is going to combine Grid applications with local applications in his/her daily workflow
  - Easy of use by the average structural biologists (e.g. authentication)
- **Ideally we'd like**
  - A Grid that pays close attention to large volumes of data
  - A Grid easier to work for the scientific community at large (and even for me, a technical developer!)