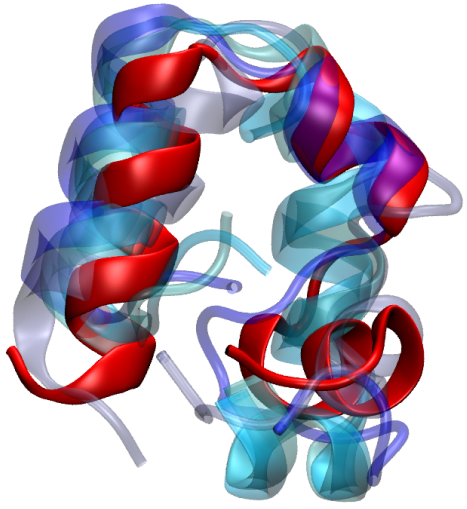
COST School - Trieste 15-18 sep 2008

Bias-exchange metadynamics and the GRID



Riccardo Di Meo

ICTP EUIndia-GRID, Trieste

Fabio Pietrucci

SISSA-ISAS, Statistical and Biological Physics Sector, Trieste

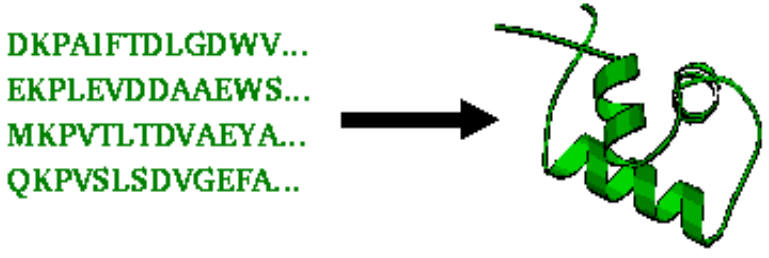
Outline

1. Protein folding
2. Bias-exchange Metadynamics
3. Porting to GRID
4. Results

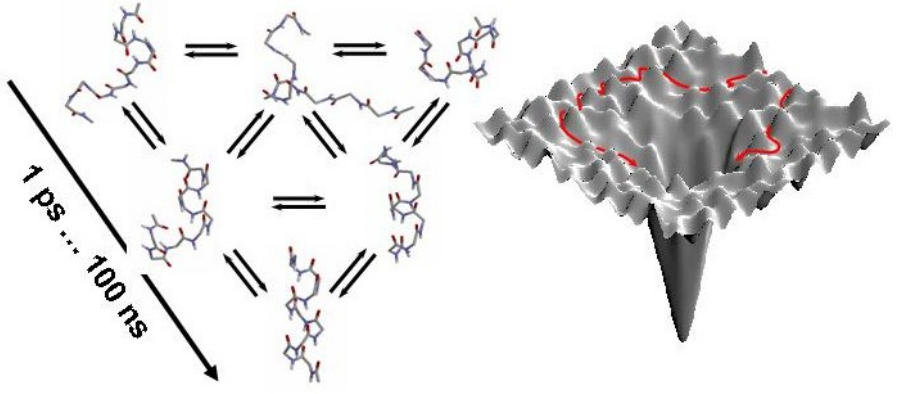
The physical problem: protein folding

a major challenge for any sampling method

- obtain 3D structure from linear sequence of amino acids



- thermodynamic stability of folded state and intermediates
- kinetic pathways of folding



We can simulate protein folding by molecular dynamics

(evolve the atom positions by Newton laws)

IF:

- the potential is accurate enough
- the simulation is long enough to explore

  a very complex free-energy surface

Protein folding is a "rare event":

folded state is separated from random coil by free energy barriers

Grid computing?

One day of CPU time → 1 ns of simulation time

small proteins fold in 10^4 – 10^6 ns

so we need 100–10000 years of CPU time !!!

Folding@Home approach:

equilibrium MD, distributed computing on 10^4 home PCs connected through internet

V. S. Pande et al., Biopolymers 68, 91 (2003)

individual folding trajectories are obtained

the folded state must be known in advance (!)

stability of intermediates?

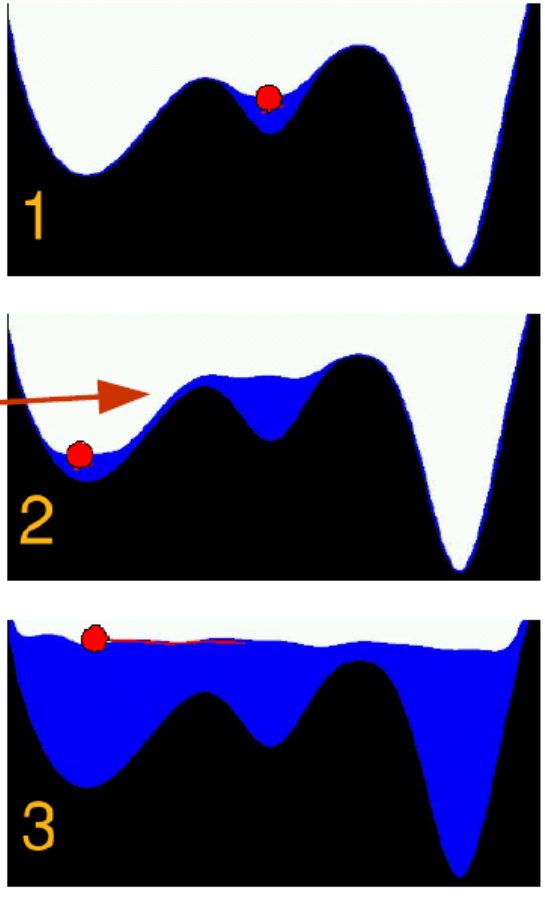
impressive computational cost

Metadynamics

A. Laio and M. Parrinello, PNAS, **99**,12562 (2002)

- Choose a reaction coordinate *s*
- Put a "small" Gaussian on *s*
- The normal dynamics brings to the closest local minimum of *F(s)* plus the sum of Gaussians

The walker finds first the lowest transition state



for large t :

$$\sum_{t'<t} W \exp\left(-\frac{(s-s(t'))^2}{2\sigma^2}\right) \rightarrow -F(s)$$

Limitations of metadynamics:

- difficult to “know” in advance all the relevant variables
- if one is forgotten → histeresis !!!
- even if you know all:

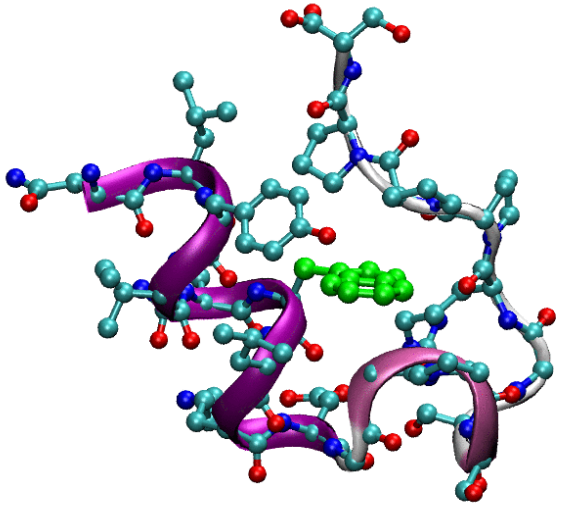
  the filling speed falls exponentially with dimensionality !

this applies not only to metadynamics

but to all methods in which one (or a few) variables are biased,

such as thermodyn. integration, umbrella sampling, WHAM, …

Protein folding: why is it so difficult?



- At least two relevant torsional angles for each residue
- Thousands of water molecules

Several possible “general” collective variables:

Gyration radius

Backbone-backbone H-bonds

Hydrophobic contacts

Fraction of α helix

Fraction of β sheet

Correlation between dihedrals

Contact order

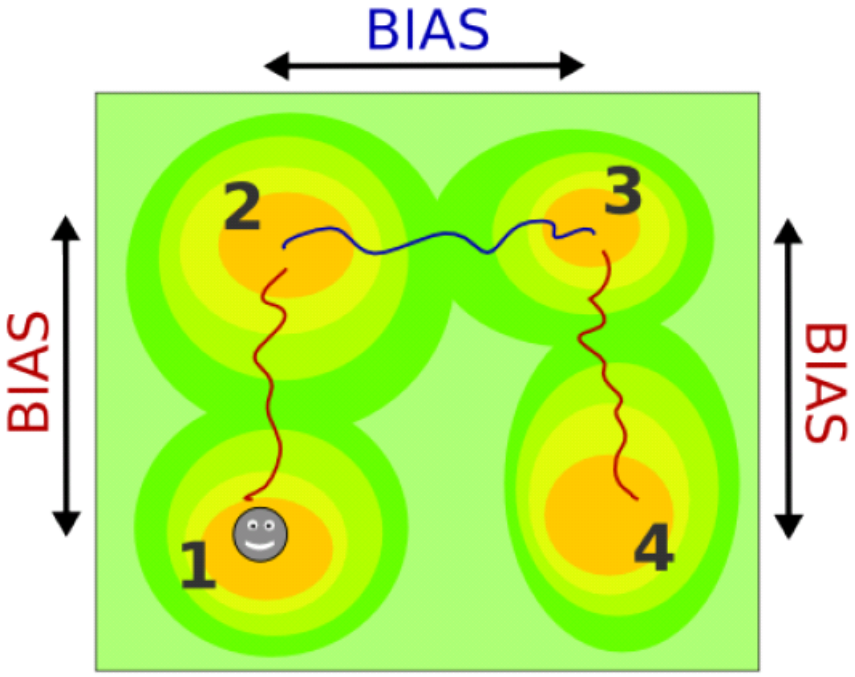
Number of salt bridges

…..

Bias-Exchange Metadynamics

S. Piana and A. Laio, J. Phys. Chem. B **111**, 4553 (2007)

- Run several metadynamics at the same T,
  each biasing different collective variables
- Try to exchange the bias potential at fixed time intervals (Metropolis)

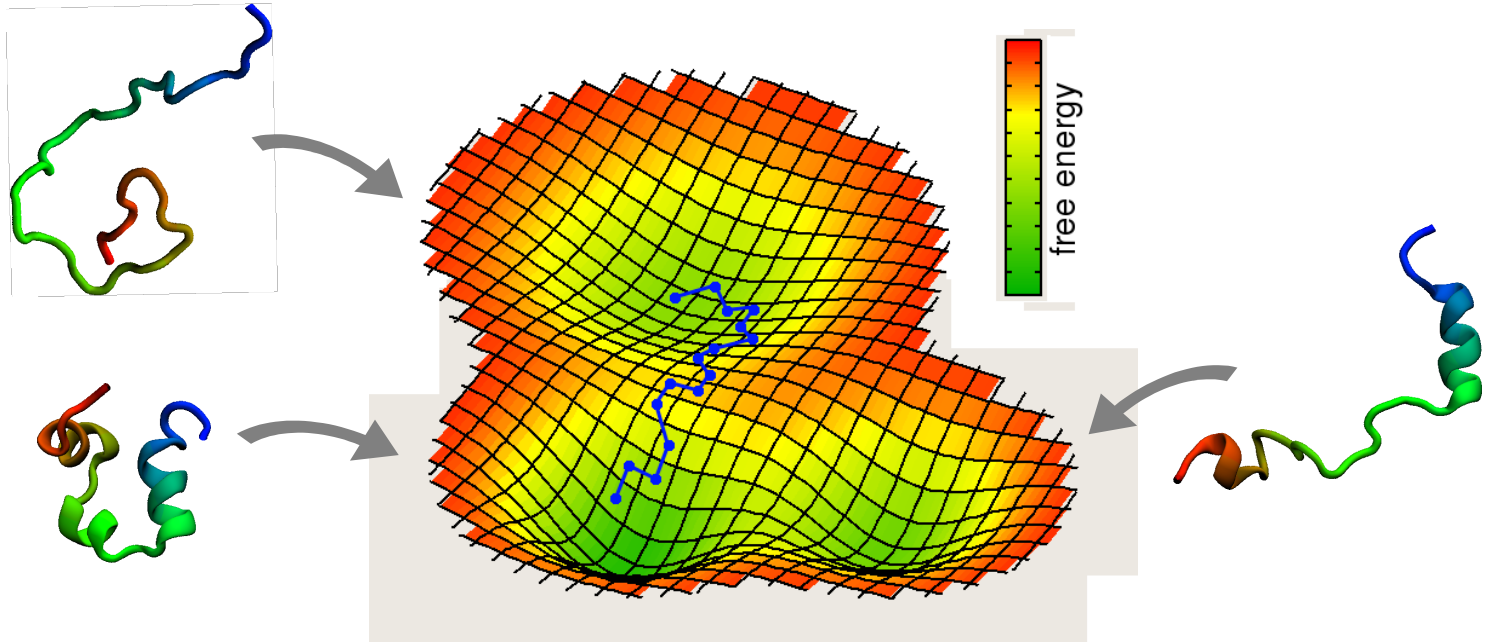


Changing the direction of
a low-dimensional bias
enables to explore
a high-dimensional space

cartoon: 2D FES explored applying 1D biases

outcome of bias-exchange simulation:

- Reconstruction of a multidimensional free-energy surface discretized on a grid of cells
- Calculation of transition rates among neighboring cells from the diffusion coefficient



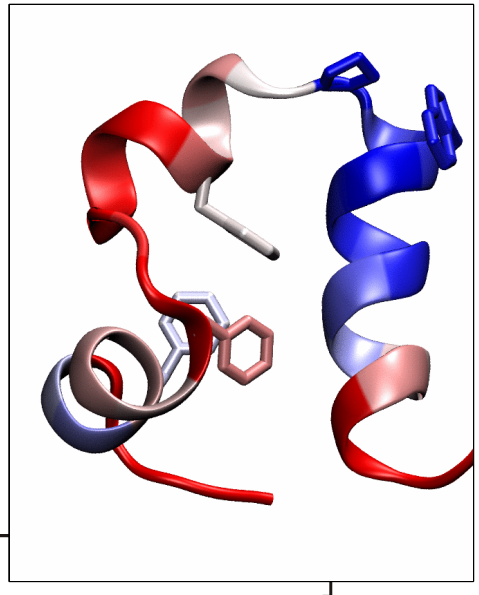
Finally, the long-time kinetics of the protein is simulated by generating cheap kinetic monte carlo trajectories → comparison with experiments

Marinelli, Pietrucci, Laio & Piana (to be submitted)

Folding advillin with BEM

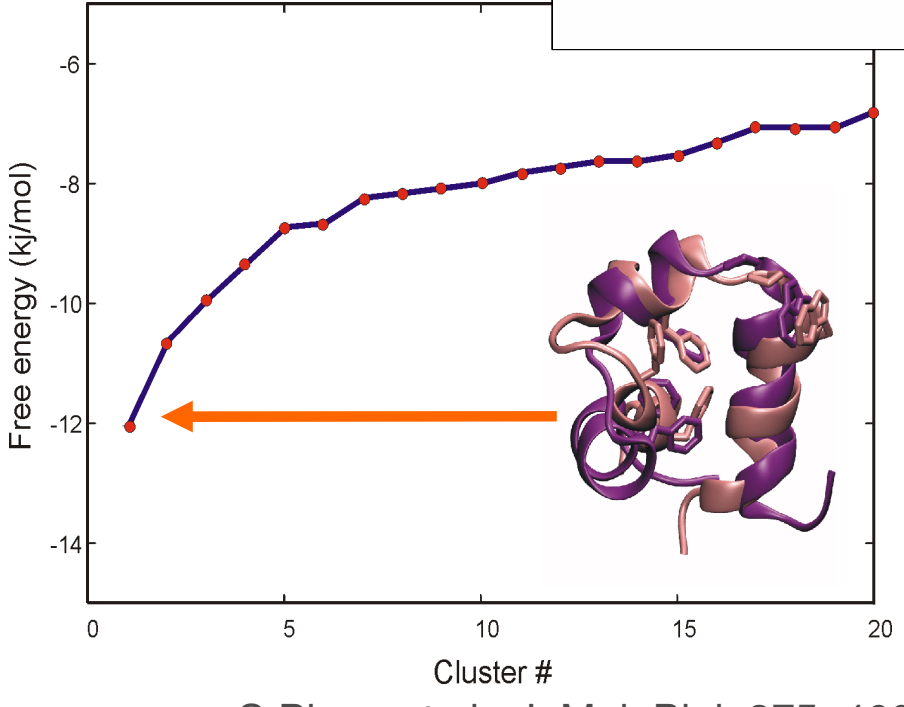
MPI: 8 procs x 40 ns

- 36 AA three helix bundle.
- Folding rate: 5-14 μs
- Amber force field, PME, NPT, T = 323 K, 3633 waters

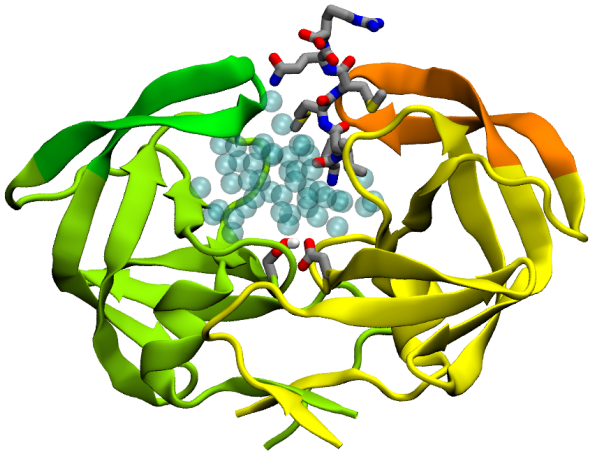


COLLECTIVE VARIABLES:

1: number of backbone H-bonds
2: number of salt bridges
3: number of hydrophobic contacts
4 (5): number of α/β residues
6 (7): dihedral correlation
8: neutral walker



S.Piana et al., J. Mol. Biol. 375, 460 (2008)

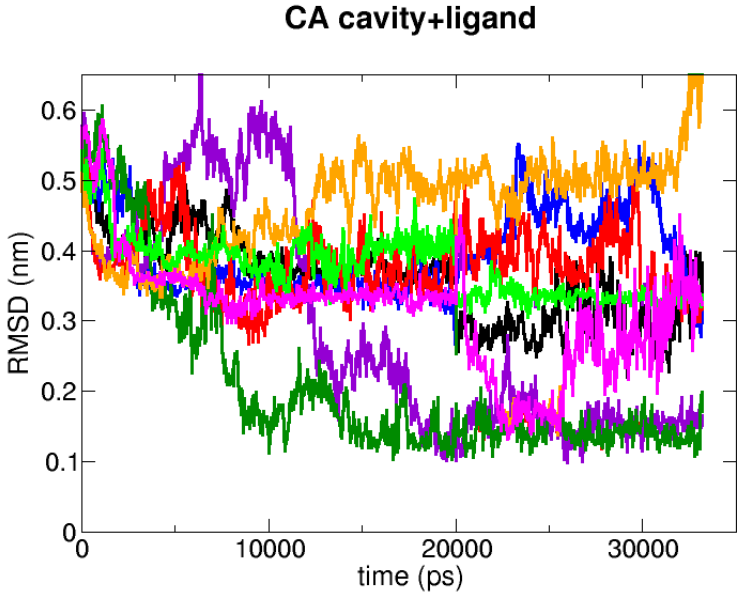


Binding mechanism of HIV-1 protease

enzyme necessary for replication of HIV virus, it cuts long proteins. Target for anti-AIDS therapy

we use 8 walkers and **7 CVs:**

- cavity-ligand hydrophobic contacts
- " hydrogen bonds
- " distance
- distance between flap tips
- number of interfacial waters



198 + 6 amino acids (protease+ligand)
26000 atoms, 7700 TIP3P waters
Amber force field, total time 1000 ns

in 35 ns 4 binding events are observed

Cα RMSD < 2 A from experiments

Porting bias-exchange metadynamics to GRID

Characteristics of the algorithm:

- Computationally intensive
- Loosely coupled
- Non I/O intensive

The porting we could have done (and that we didn't)

- The program came with a working MPI interface already
- Doing the porting would have been a trivial matter of writing a simple JDL
- The EGEE infrastructure already supports a MPI flavor
  - specifically, MPICH with the device p4 compiled

Good reasons to rule MPI out

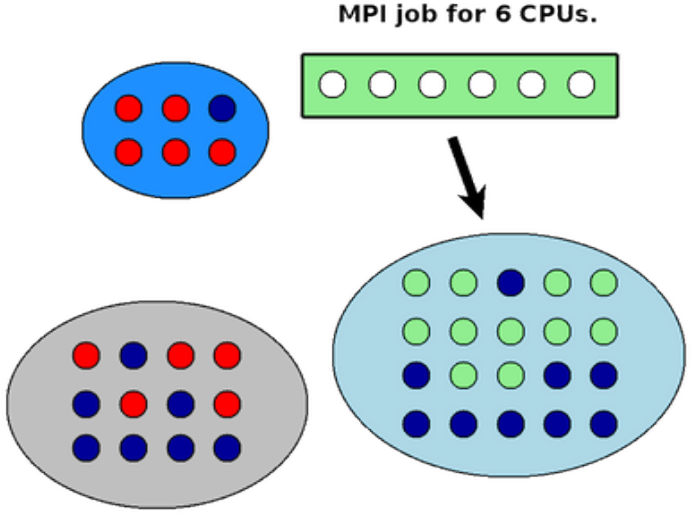
- Only a subset of the CEs are supporting it
- Negatively affects the scheduling of jobs
- MPI does allow just parallel execution within CEs and not among them
- Does not allow new dynamic recruiting

Our approach

- Standard client-server approach
  - Allows to easily keep track of the simulation
  - Bridges data between different CEs
- Asynchronous recruitment of the resources
- Complete rewriting of the network code

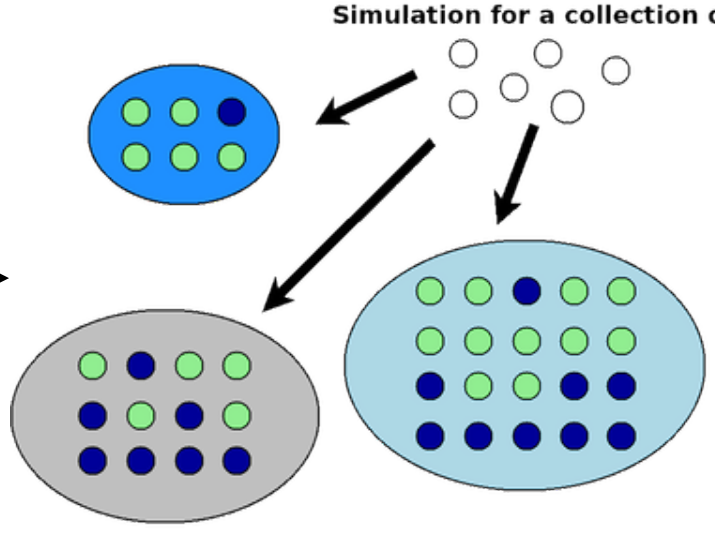
MPI vs. Custom I/O

Normal MPI submission:
only one CE available for every job!



- Slots already owned (17 of 38)
- Slots suitable for the MPI job (11 of 38)
  (but really only 1 slot for the job available!)
- Free slots where our job will not fit (10 of 38)

Custom networking:
all free resources are available!



- Slots already owned (17 of 38)
- Slots suitable for the execution (21 of 38)
  (3 simulations can be executed at once, or a single one for 21 processors can be launched!!)
- Free slots where our job will not fit (0 of 38)

GRID-related issues

- The behavior of the algorithm with a reduced frequency of exchanges?
- How the asynchronous scheduling would have affected the execution?
- What would have been the contribution of different CPUs from mixed CEs?

Technical issues

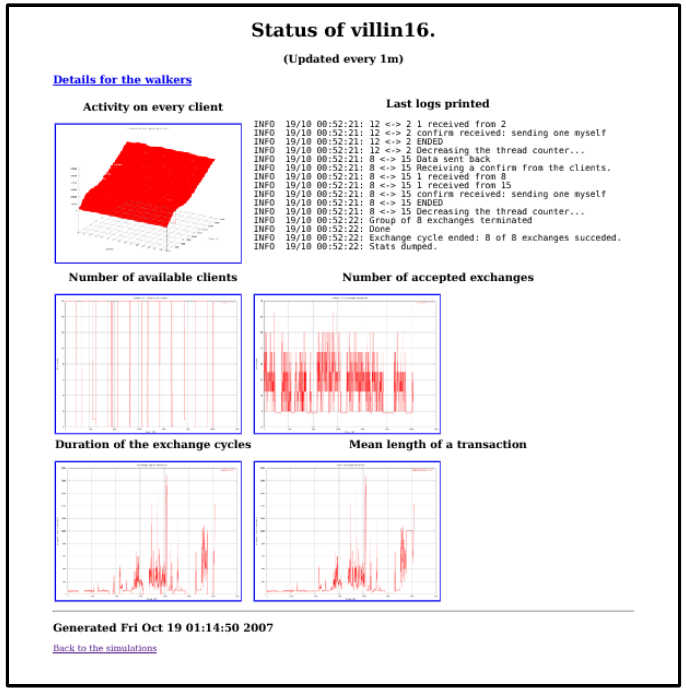
- Very limited changes to the original code are allowed.
- WNs do not have inbound connectivity.
- The communication is not allowed among WNs belonging to different CEs.
- The simulation needs a synchronization source.

Our implementation

- To minimize the impact on the original code, a client wraps it and handles the network IO.
- A server (usually on an UI) receives and handles the requests, bridging the data between different WNs and synchronizing the simulation.
- Everything is written in Python 2.5 using plain sockets.

Extra features

Small webserver



User Interface

```
Environment setup
*****************
* Compile gromacs
* Compile python (for the WNs)
* Check gromacs and python
* Check/Setup the environment
* Upload/Copy the support files
* Back


--------------------
Back to the main Menu
```

Real time feedback

```
INFO   11/10 23:17:56: Sending the data
INFO   11/10 23:17:56: The size of the data is 2488077
INFO   11/10 23:17:57: Receiving the data
INFO   11/10 23:18:09: Exchange refused.
INFO   11/10 23:18:09: Sending a confirm to the server.
INFO   11/10 23:18:09: Receiving a confirm from the server.
INFO   11/10 23:18:09: Exchange correctly completed!
INFO   11/10 23:18:09: ener.edr found!
INFO   11/10 23:18:09: traj.xtc found!
INFO   11/10 23:18:09: traj.trr found!
INFO   11/10 23:18:09: Copying traj.trr to checkpoint_dir/traj.trr
INFO   11/10 23:18:10: Copying HILLS.0 to checkpoint_dir/HILLS.0
INFO   11/10 23:18:13: Copying ener.edr to checkpoint_dir/ener.edr
INFO   11/10 23:18:13: Copying META_INP.0 to checkpoint_dir/META_INP.0
INFO   11/10 23:18:15: Copying traj.xtc to checkpoint_dir/traj.xtc
INFO   11/10 23:18:15: Copying villin.tpr to checkpoint_dir/villin.tpr
INFO   11/10 23:18:17: Copying COLVAR.0 to checkpoint_dir/COLVAR.0
```

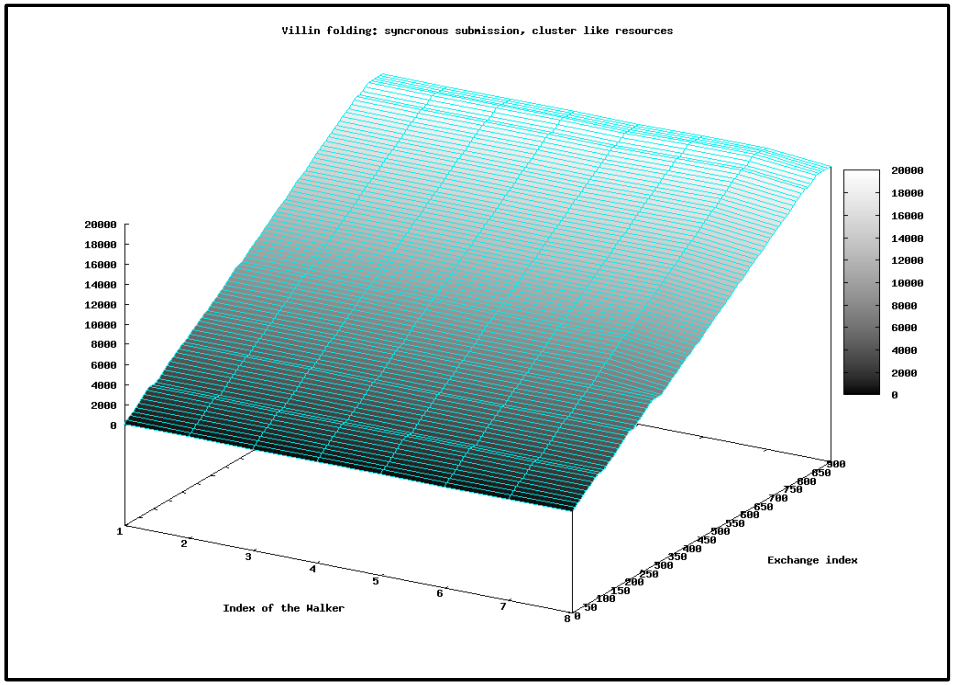
Tests performed (1)

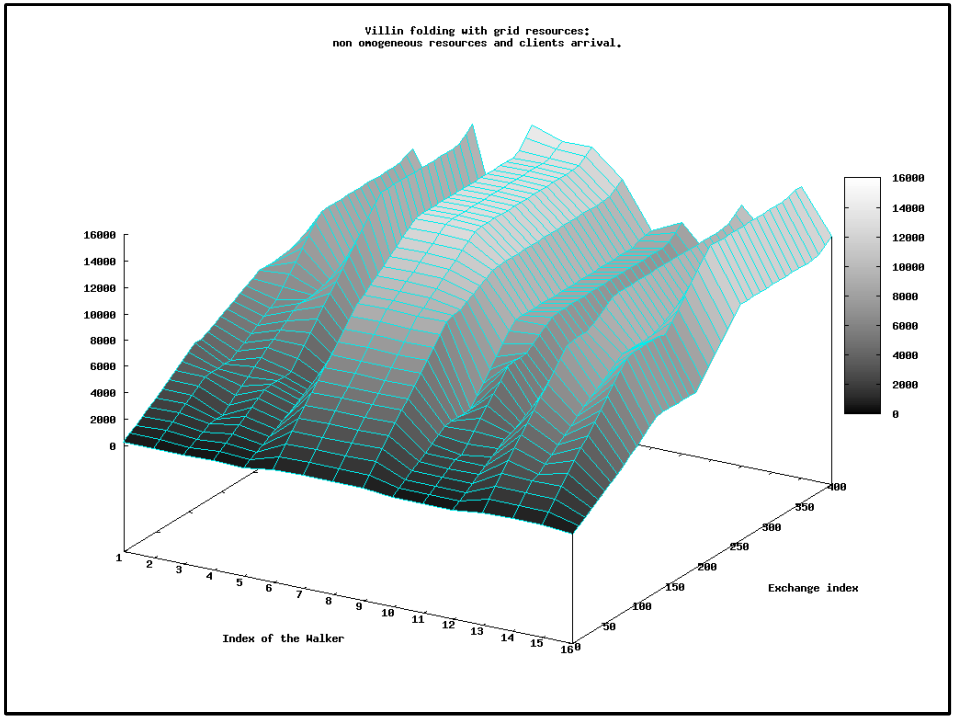
- 8 and 16 CPUs, homogeneous resources, synchronous submission.
- 8 CPUs, homogeneous resources, asynchronous scheduling (3h between each client arrival).
- 16 CPUs heterogeneous hardware and asynchronous scheduling (blind grid-like scheduling).

Tests performed (2)

- Cluster-like simulation

(homogeneous – synchronous)

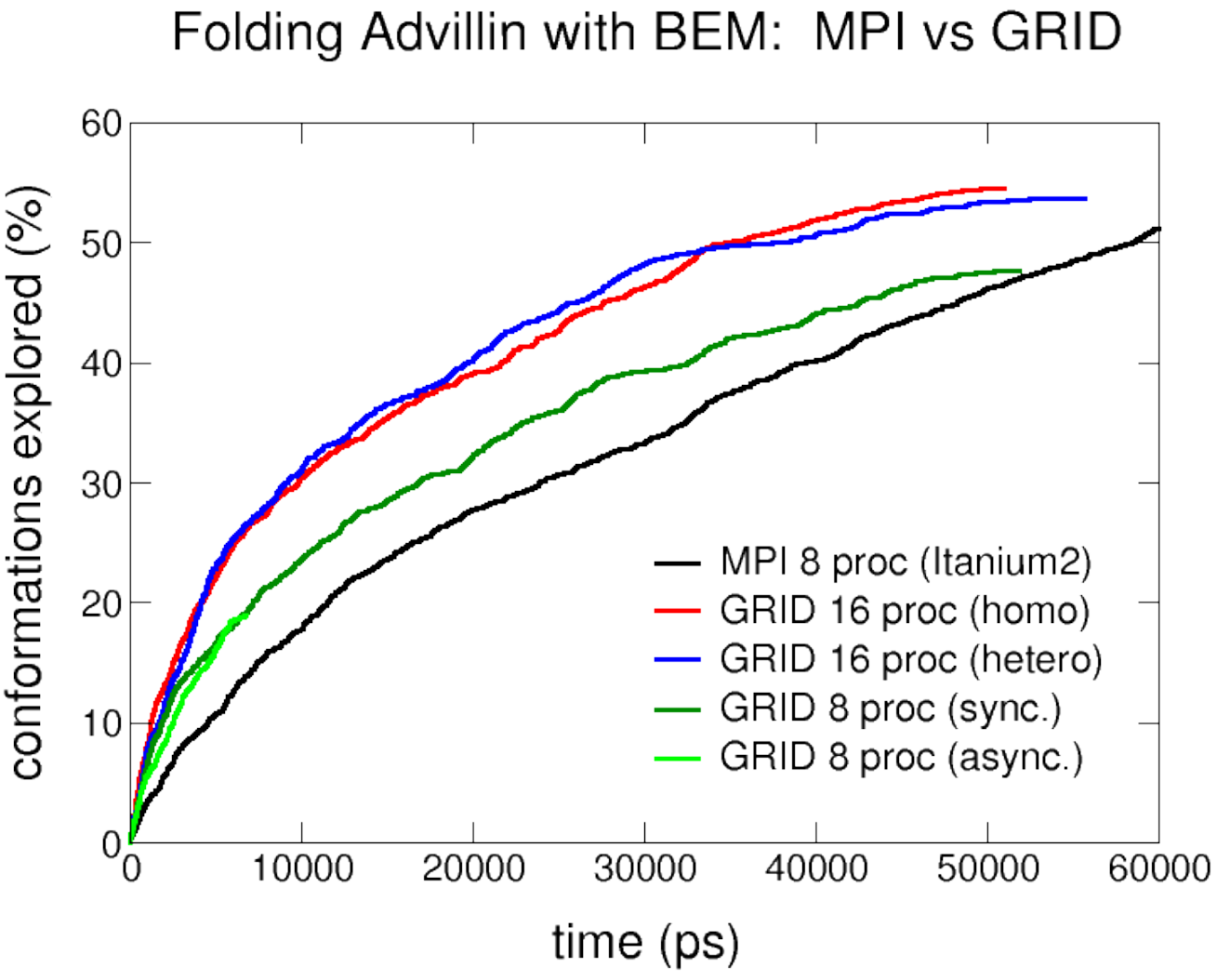




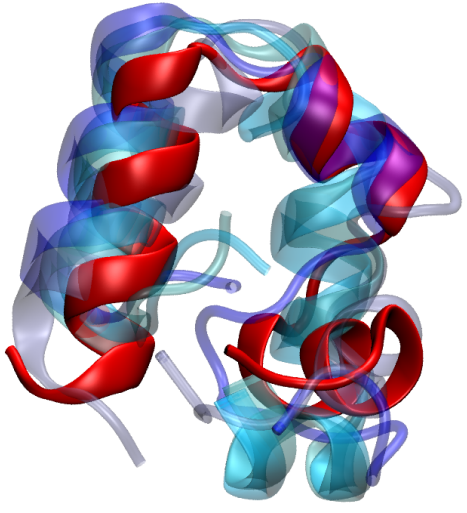
- Grid simulation

(mixed resources, asynchronous)

Results



Our simulations were able to explore the conformations of the protein at a speed comparable to MPI !!!



RMSD < 4 A from exp.

Ongoing development

Thanks to the huge computational resources offered by GRID,

we are able to address complex biological problems

Binding between drugs and HIV-1 protease

(>200 amino acids, all-atom explicit solvent, ~100 procs)

Folding of a 56-amino acids protein in water (src-SH3)

Remarks (1)

- We have been able to use the grid resources with the same efficiency as HPC resources

Key factors:

- Processes in a BEM simulation are loosely coupled

(data exchanged are few Mb/h)

- We discarded the easy to implement MPI approach in favour of a more grid suited one.

Remarks (2)

- For a given simulation time, the number of explored conformations of advillin is similar in MPI and GRID

- We solved the problem of folding a 36-aa protein at a cost (**< 1 year of CPU time**) far lower than other GRID-based techniques (Folding@Home required 1000 years of CPU time)

- We were able to combine the advantages of GRID with the advantages of the Bias-Exchange Metadynamics algorithm

Conclusions

- We successfully ported the BEM algorithm on the EU-India GRID infrastructure

- We proved that GRID environment perfectly fits the computational experiment that can be performed using the BEM algorithm

- BEMuSE is now a working tool ready to tackle important computational challenges

Acknowledgements

Alessandro Laio

SISSA – Statistical and Biological Physics Sector, Trieste

Stefano Cozzini

Democritos National Simulation Center, Trieste

Links

- The EU-IndiaGRID project site at ICTP: **www.euindia.ictp.it/bemuse**
- The homepage of Alessandro Laio: **people.sissa.it/~laio/metadynamics.htm**
- The homepage of Fabio Pietrucci: **people.sissa.it/~pietruc/research_BE.html**

details on Bias-Exchange Metadynamics

Run several metadynamics at the same T, biasing different collective variables

Exchange move: swapping the gaussian potentials V_G^a , V_G^b of the two replicas. Accept the move according to:

$$P_{exch} = \min(1, \exp(-\Delta)) , \quad \Delta = \beta(V_G^a(x^b,t)+V_G^b(x^a,t)-V_G^a(x^a,t)-V_G^b(x^b,t))$$

- ☻ Parallel reconstruction of *F(s)* in a virtually unlimited number of CVs
- ☻ The accuracy of each *F(s)* is greatly enhanced by the jumps in CV space due to the exchanges
- ☻ As the replica are at the same temperature, the normal potential energy cancels out: it works even with a few replicas.

S. Piana and A. Laio, J. Phys. Chem. B **111**, 4553 (2007)