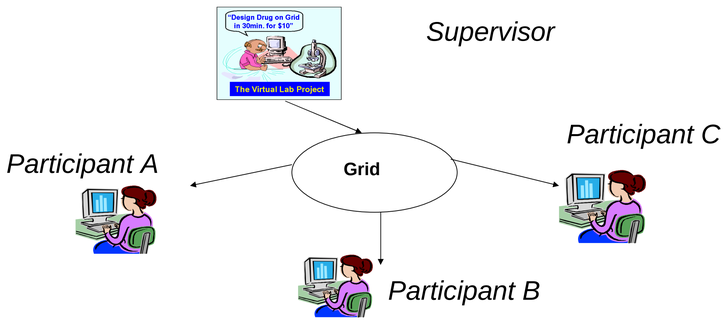
Uncheatable Grid Computing and Its Application in Drug Discovery

Mummoorthy Murugesan (Purdue University) and Wenliang (Kevin) Du (Syracuse University)

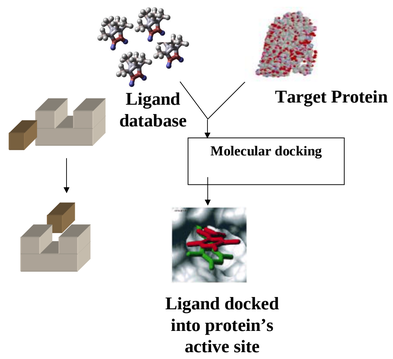
Grid Computing

- Pervasive access to Computational Resources such as computer time, storage, data, etc.
- Supervisor-Participant architecture
- SETI@home (5 Million users, 15 Teraflops) , FightAIDS@Home(HIV drug)



Drug Discovery

- Find a 'drug' molecule to modulate disease
- Lab experiments – for billion of molecules? Impossible
- Molecular Modeling – also known as 'Ligand Docking'
  - Model the 2 molecules (Protein Data Bank)
  - Match & Score each matching configuration
  - Find a good matching – lead molecules for lab experiments
- Total time : 12-15 Years & Spending : 350 Million dollars
  Internet gird : FightAIDS@HOME



Motivation

200$

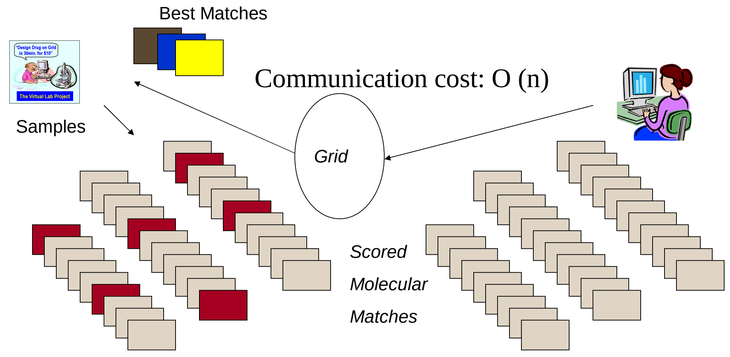
2 Molecule Models. Do matching of 100 Million Configurations. Give me the Top 100

Here is the TOP 100

Did you do all 100 Million matching?

How to verify the results?

- reduce communication cost and re-computation cost



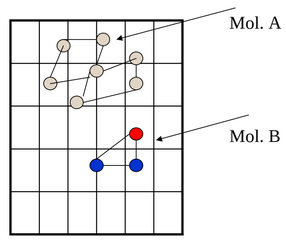
Improvements to the CBS [Du, et al. 2004]

- remove the interaction (Non-Interactive CBS)
- reduce the storage requirements

Implementation of CBS in two Molecular Docking tools

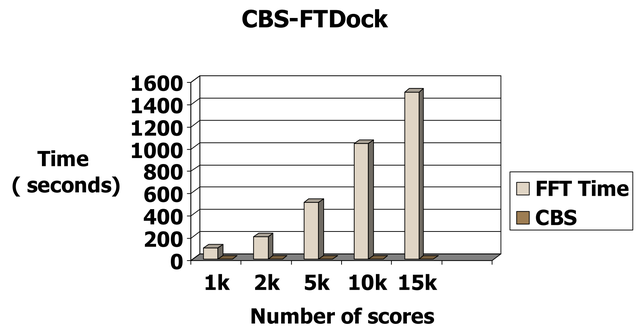
1. FTDock (FFT based Molecular docking)

- N*N*N 3-D Grid for Molecule A & B
- Straightforward 3-D pattern matching
  Molecular A: fa – Surface 1 , Core –ve value
  Molecular B: fb - Surface 1
  or each translation
  $fc=\sum\sum\sum fa*fb$
- Use Discrete Fourier Transform to find correlation fc
- Generate all the correlations using FFT



Verification:

Total Scores: 10^10 , Sample size: 1000
Communication Cost
Naïve Sampling - 20 GB CBS - 0.53MB
Supervisor work:
1 in 10^7 computation

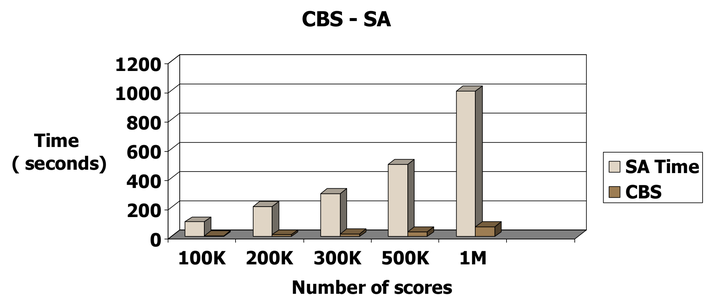


2. AutoDOCK (Simulated Annealing based Molecular Docking)

- Protein A & Molecule B
- Keep A static and Molecule B is randomly configured
- Calculate 'Interaction Energy' at each step
  If better then previous, Accept it
  Else Accept on some probability
- Repeat

Verification:

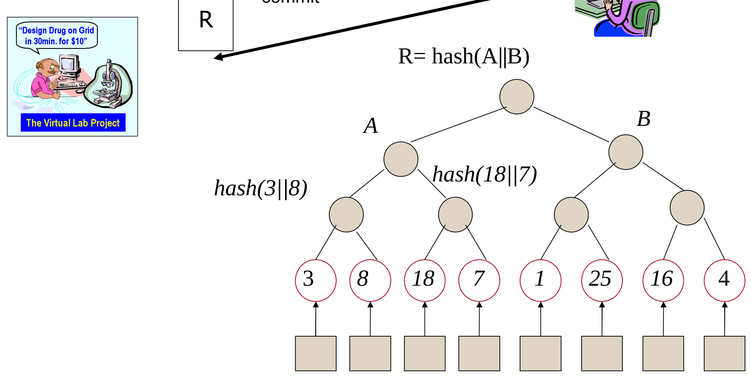
- Use the modified version of CBS for semi-sequential computations
- Total Energy calculations: 10^8 (approx) Sample Size: 1000
  Communication Cost
  Naïve Sampling: 4.8 GB CBS – 4MB
  Supervisor work
  1 in 10^5 computations

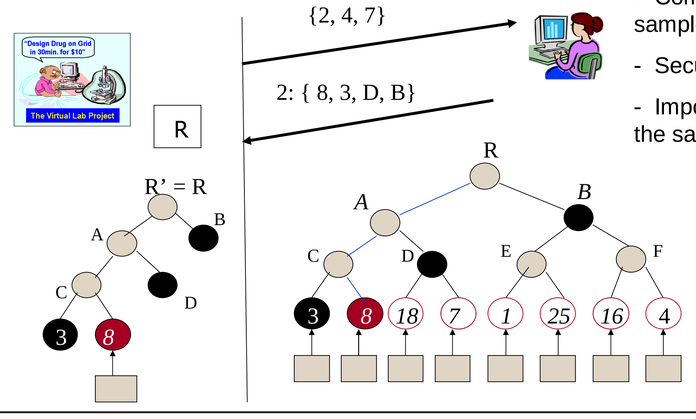


Solution:

Commitment Based Sampling Scheme [Du, et al. 2004]

1. Commit all the results using Merkle-Hash tree
2. Randomly sample and verify





- Communication cost: O(m log n) (m – number of samples)
- Security is based on the underlying hash function
- Impossible to find values other than {8,3,D,B} to get the same R value

PURDUE UNIVERSITY

CERIAS®

Discovery Park e-Enterprise Center