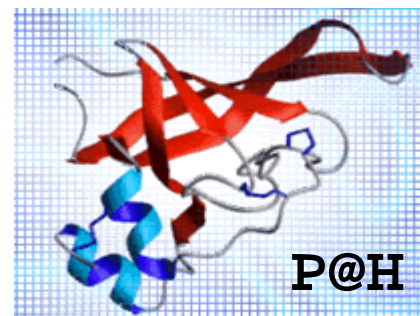


Homogenous Redundancy: a Technique to Ensure Integrity of Molecular Simulations Results on Global Computing

Michela Taufer, David Anderson,
Pietro Cicotti, Charles L. Brooks III



Global Computing Paradigm

- Heterogeneous distributed computing based on PCs volunteered by the public and connected to the Internet
- Volunteers donate unused cycles by installing clients on their PCs and attaching one or more projects to them
- Example of global computing frameworks:
 - [BOINC](#)
 - [XtremWeb](#)
- Example of projects:
 - [SETI@home](#): Analysis of radio telescope data from Arecibo
 - [Predictor@home](#): Prediction of protein structures



Computations from volunteers PCs have non-negligible [error rates](#) and [result variations](#)

Computational Errors and Variations in Computed Results

- Computational errors or variations in computed results are due to:
 - Hardware malfunctions
 - Incorrect software modifications
 - Malicious attacks
 - Differences in floating-point hardware
 - Differences in libraries and compilers
- Results validation is needed to ensure overall correctness

Homogenous Redundancy (HR) policy: a technique to assure result integrity of Monte Carlo (MC) and Molecular Dynamics (MD) simulations on global computing systems

Outline

- Redundant computing and fuzzy comparisons
- Limits of fuzzy comparisons for molecular simulations
- Simulation cases:
 - Monte Carlo protein conformation search
 - Molecular Dynamics protein refinement
- Homogenous Redundancy (HR) policy
- Integration of HR in BOINC
- Result validation using HR on Predictor@home
- Conclusion

Redundant Computing

- Redundant computing is a general technique to deal with errors or result variations
 - Same computation is performed on different PCs
 - If results “agree”, they are correct
- Result agreement techniques:
 - Bit-to-bit identical results
 - Fuzzy direct comparison → results must agree within application-specific tolerances
 - Fuzzy indirect comparison → a quantity, e.g., energy, derived from the results and easy to compute is within a certain threshold

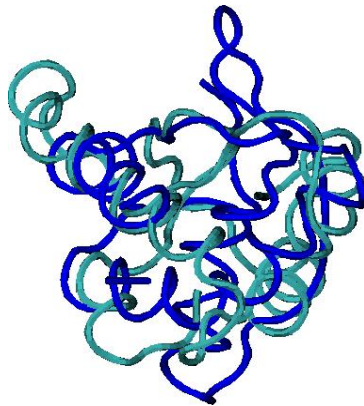
Simulation Cases

- Molecular simulations based on MC and MD produce different outcomes for replicas of the same computation on global computing systems

MC protein conformational search
→ MFold

MD all-atom protein refinement
→ CHARMM

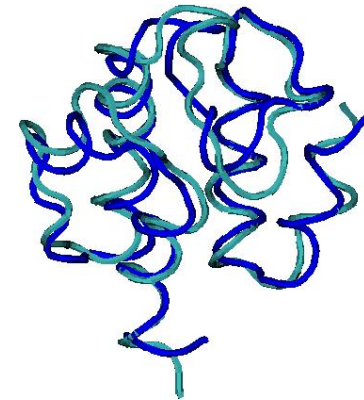
$\Delta\text{energy}=2.5\%$



Target: t0243

- Linux OS, GNU FORTRAN comp.
- Windows OS, Intel FORTRAN comp.

$\Delta\text{energy}=1\%$

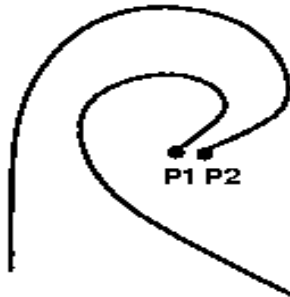


Target: t0221

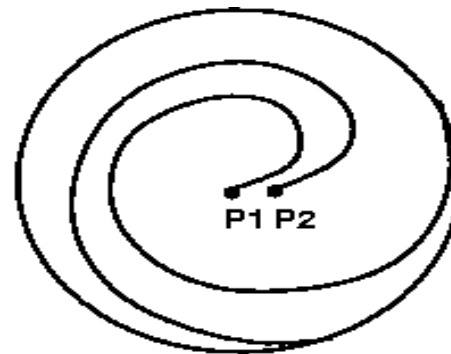
- Intel processor
- AMD processor

Limits of Fuzzy Comparison

- MD and MC simulations are subject to **positive** Lyapunov exponents → Computation results are highly sensitive to initial simulation states and might **diverge**



positive Lyapunov exponent



negative Lyapunov exponent

P1 and P2 are
initial simulation
states

- Fuzzy comparisons are not applicable to MD and MC replicas:
 - Direct comparison: not possible to define a priori a threshold on the deviation of the two results
 - Indirect comparison: derived quantities such as energy do not capture malicious attacks on the returned structures

Homogenous Redundancy Policy

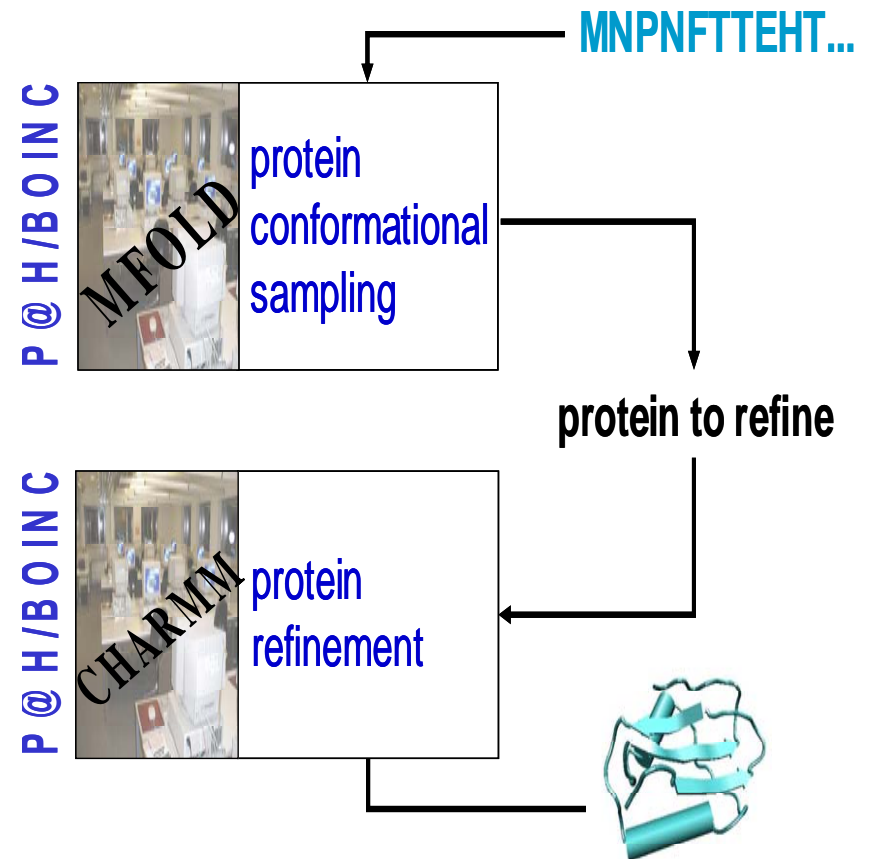
- To ensure result integrity for MC and MD simulations → **Homogeneous Redundancy (HR) policy**
 - Initial simulation states include machine architecture, operating system, specific compiler and compiler flags
 - Distribution of replicas of the same computation only among "numerically equivalent" PCs, i.e., that compute identical floating-point results
- Once an instance of a task is sent to a client, the replicas of the same task are sent to numerically equivalent clients
- HR allows strict equality to compare redundant results
 - Bit-to-bit validation

Homogenous Redundancy in BOINC

- BOINC (Berkeley Open Infrastructure for Network Computing):
 - Global computing open-source framework based on public-resources
 - Built-in support for distributed computing on heterogeneous PCs connected to the Internet.
- So far BOINC provided built-in support for redundant computing
- We have extended BOINC to support our HR policy:
 - Empirical recognition of numerical equivalent clients
 - Information on processor manufacturer and OS is provided by the BOINC client to the server

Predictor@Home (P@H)

- P@H → a global computing project for prediction of unknown protein structures starting from amino acid sequences
- Each prediction is the result of a code pipeline:
 - MC protein conformational search → MFold
 - MD protein refinement → CHARMM
- P@H uses BOINC in-built support extended with HR for distributing conformational samples and proteins refinements



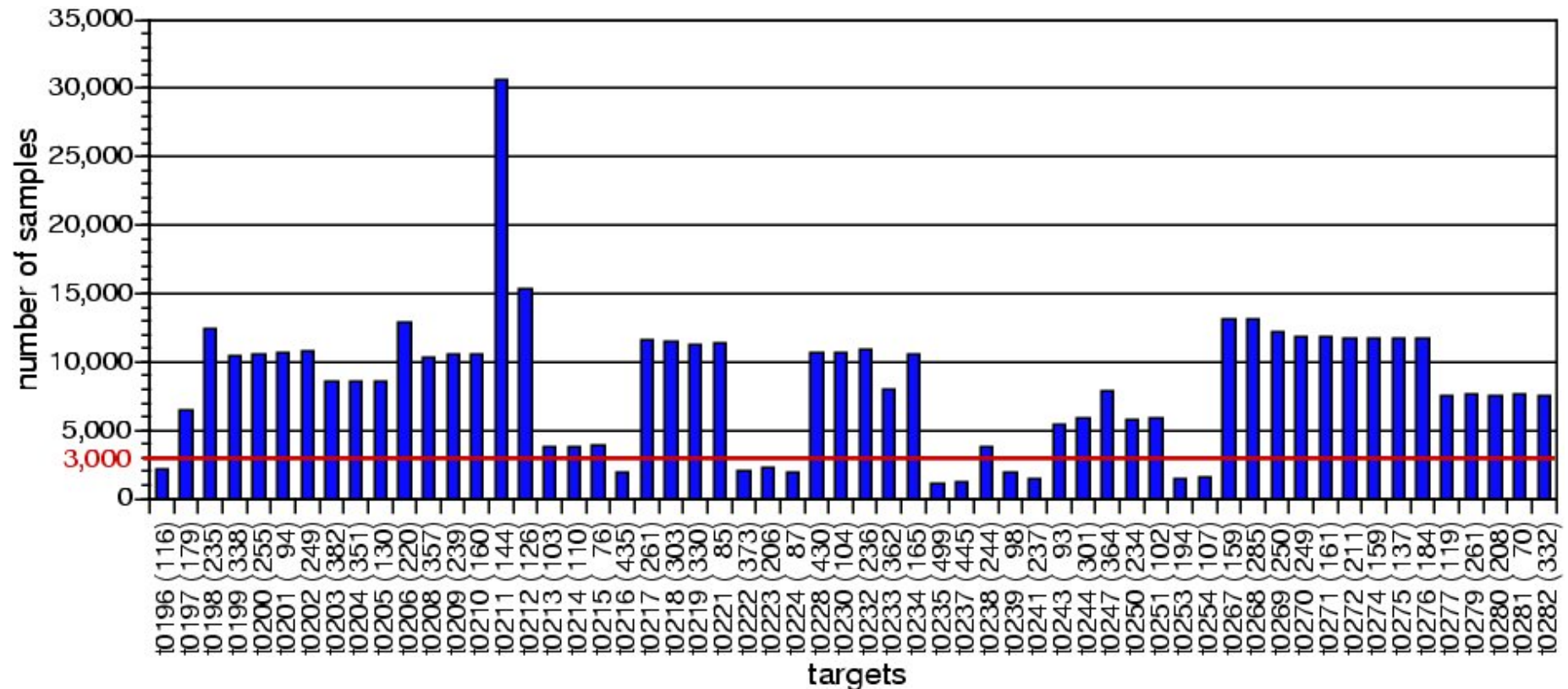
Results from Predictor@Home

- Results collected over two overlapping interval of times:
 - MFold: from August 17. to August 30.
 - CHARMM: from August 26. to August 30.
- For each prediction we ran at least **3 replicas**
- Farm of heterogeneous PCs:
 - Number of PCs: ~12,000
 - Number of users: ~ 6,000

Replicas	MFold	CHARMM
days	14	4
total	402,002	138,591
valid	380,269	99,352
	94.6%	71.8%
invalid	11,690	12,139
	2.9%	8.7%
error	10,043	27,100
	2.5%	19.5%

- CHARMM is characterized by higher number of flops than MFold
→ more vulnerable to crashes, computational errors or variations in computed results

Predictor@Home in Numbers



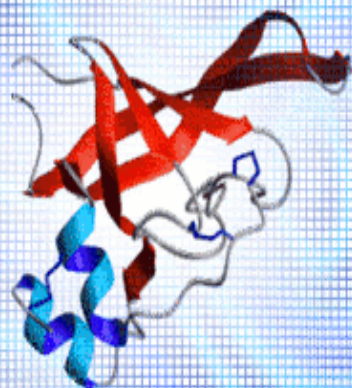
From Jun 1. to Aug 31. → 398,499 **HR-validated** protein predictions of 58 amino acid sequences for a total of 380 years of computing time

Conclusions

- Simulations using global computing based on public resources have a non-negligible error rate
- Replication and fuzzy comparisons are common techniques for result validation
- For molecular simulations, replica results might significantly diverge → fuzzy comparison is not applicable
- Homogeneous Redundancy (HR) is based on the distribution of replicas among numerically equivalent machines
- HR policy has been implemented within the BOINC framework to support result validation in [Predictor@home](#), a world-community effort to predict unknown protein structure from amino acids sequence.

Acknowledgments

- Charlie L Brooks III (TSRI)
- David Anderson and Rom Walton (BOINC)
- Pietro Cicotti and Chris Wildman (CSE-UCSD)
- Chahm An and Andre Kerstens (TSRI)
- The Predictor@home and BOINC community of volunteers



PREDICTOR @ home

powered by 



predictor.scripps.edu

What is Predictor@home?

Predictor@home is a world-community experiment and effort to use distributed world-wide-web volunteer resources to assemble a supercomputer able to *predict protein structure from protein sequence*. Our work is aimed at testing and evaluating new algorithms and methods of protein structure prediction. We recently performed such tests in the context of the Sixth Biannual [CASP](#) (Critical Assessment of Techniques for Protein Structure Prediction) experiment, and now need to continue this development and testing with applications to real biological targets. Our goal is to utilize these approaches together with the immense computer power that can be harnessed through the internet and volunteers all over the world (you!) to address critical biomedical questions of protein-related diseases. **Predictor@home** is a pilot project of the Berkeley Open Infrastructure for Network Computing ([BOINC](#))

12/21/04 - An update from Professor Charles L. Brooks, III: [Update](#)

Predictor Status

server:	up
results in queue:	9990
successful results yesterday:	14649
unique hosts yesterday:	5090
last update:	18:15

Account Creation

Open! Welcome to Predictor. PLEASE NOTE: YOU MUST TURN OFF YOUR SPAM FILTER TO GET THE CONFIRMATION EMAIL WHEN YOU CREATE YOUR ACCOUNT!