APPLICATIONS AND ADVANCEMENTS OF THE PROGRESS-INDEX GUIDED SAMPLING METHOD IN MOLECULAR DYNAMICS SIMULATIONS

Marco Bacci
University of Zürich
Department of biochemistry

02 July 2018
Basel, Switzerland
Conventional sampling in molecular dynamics

“Biology today is moving in the direction of chemistry. Much of what is understood in the field is based on the structure of molecules and the properties of molecules in relation to their structure. If you have that basis, then biology isn’t just a collection of disconnected facts.” - Linus Pauling, 1986

Molecular dynamics simulate the motion of atoms by integrating Newton’s equations of motion

High resolution in space and time. Limited by:
1. Accuracy of models
2. Statistical Precision
   - Parallel computations on HPCs.
   - Advanced sampling methods

Advanced sampling in computer simulations of biological systems

There are 10s of enhanced sampling methods based on, e.g., potential energy sculpting, external forces, knowledge-based, resolution exchange, ... None of those will be optimal for all problems.
Progress-Index Guided Sampling (PIGS) algorithm: Goals

- Enhance the sampling with respect to brute-force MD.
- As easy to use as vanilla MD (or almost...).
- Scalable to modern hardware architecture and synergistic.
- Avoid guidance by a collective variable.
- Maintain pathways information, *i.e.*, sampling should be vastly canonical.
- Able to focus the sampling enhancement to specific degrees of freedom.

PIGS is a parallel replica simulation protocol based on successive reseeding.

The main point is how do we define the heuristic that informs the reseeding.

Benefits / Novelties
- Increased sampling. Pathways are preserved.
- Synergistic. Scalable. Focused explorations.

Caveats
- Ensembles are thermodynamically biased → MSM used to alleviate the bias.
1. We use a data mining technique (Progress-Index) to order all the snapshots from $N_r$ stretches each having $n_o$ snapshots.

$$\text{Pairwise distances: } a_{ab}, a_{bc}, a_{cd}, a_{de}$$
PIGS protocol: Progress-Index method and heuristic reseedings

1. We use a data mining technique (Progress-Index) to order all the snapshots from $N_r$ stretches each having $n_o$ snapshots\(^1\)

\[\text{Pairwise distances}\]
\[
\begin{array}{cccc}
\text{a} & \text{b} & \text{c} & \text{d} \\
\text{b} & q_{ab} & & \\
\text{c} & q_{ac} & q_{bc} & \\
\text{d} & q_{ad} & q_{bd} & q_{cd} \\
\text{e} & q_{ae} & q_{be} & q_{ce} & q_{de} \\
\end{array}
\]

\[\text{Progress Index}\]

PIGS protocol: Progress-Index method and heuristic reseedings

1. We use a data mining technique (Progress-Index) to order all the snapshots from $N_r$ stretches each having $n_o$ snapshots.

$$\text{Pairwise distances} \quad a \quad b \quad c \quad d$$

$$\begin{align*}
q_{ab} & \quad q_{ac} & \quad q_{ad} \\
q_{bc} & \quad q_{cd} \\
q_{ac} & \quad q_{bd}
\end{align*}$$

1. We use a data mining technique (Progress-Index) to order all the snapshots from $N_r$ stretches each having $n_o$ snapshots\(^1\)

\[\begin{align*}
\text{Pairwise distances} & \\
\text{a b c d} & \\
\text{b c q_{ab} q_{bc}} & \\
\text{c d q_{cd}} & \\
\text{d e q_{de}} & \\
\text{e a q_{ae} q_{be} q_{ce} q_{de}} & \\
\end{align*}\]

1. We use a data mining technique (Progress-Index) to order all the snapshots from $N_r$ stretches each having $n_o$ snapshots$^1$

PIGS protocol: Progress-Index method and heuristic reseedings

1. We use a data mining technique (Progress-Index) to order all the snapshots from $N_r$ stretches each having $n_o$ snapshots

![Diagram of pairwise distances and Progress Index](image)

PIGS protocol: Progress-Index method and heuristic reseedings

1. We use a data mining technique (Progress-Index) to order all the snapshots from $N_r$ stretches each having $n_o$ snapshots\(^1\)

---

1. We use a data mining technique (Progress-Index) to order all the snapshots from $N_r$ stretches each having $n_o$ snapshots\(^1\)

2. Heuristic: Final snapshots of all replicas are ranked according to (greater is better):
   - Position in the progress index
   - Distance to parent snapshot
   - Minimum distance in the progress-index to any other final snapshot

---

1. CAMPARI (campari.sourceforge.net) as propagator and data-miner (no I/O costs, slower propagation)
2. GROMACS (www.gromacs.org) as propagator ⇔ python wrapper ⇔ CAMPARI as data-miner (I/O costs)
1. Cost of FILE HANDLING, CAMPARI set-up and read-in is $\sim 10\%$ total time, regardless of # replicas
2. Cost of OMP-only data-mining is $\sim 10\%$ with a few tens of copies but $\sim 30\%$ with a few hundred copies
## PIGS: Focused conformational sampling - Overview

<table>
<thead>
<tr>
<th></th>
<th>ATAD2A</th>
<th>BAZ2A</th>
<th>BRPF1B</th>
<th>CREBBP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sampler</strong></td>
<td>CS ZA PIGS</td>
<td>CS ZA PIGS</td>
<td>CS ZA PIGS</td>
<td>CS ZA PIGS</td>
</tr>
<tr>
<td><strong>Representation</strong></td>
<td>- 13Ψs + 1χ₁</td>
<td>5Ψs + 1χ₁ + 1χ₂</td>
<td>- 13Ψs + 1χ₁</td>
<td>- 14Ψs + 1χ₁ + 1χ₂</td>
</tr>
<tr>
<td><strong>Total Sampling</strong></td>
<td>8.16 5.76</td>
<td>8.16 5.76</td>
<td>8.16 5.76</td>
<td>8.16 5.68</td>
</tr>
<tr>
<td>time in µs</td>
<td>6.22</td>
<td>5.76</td>
<td>5.76</td>
<td>5.68</td>
</tr>
<tr>
<td><strong># Replicas</strong></td>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Representative bromodomain fold**

**Representative structures ZA PIGS**

**Representative structures BC PIGS**

**Formulas:**

$$\Psi_s + \chi_1 + \chi_2$$
PIGS: Focused conformational sampling - Low dimensional projections (PCA)

- PCA on sine and cosine values of the dihedral angles of ZA and BC loop residues (all snapshots together)
- Densities reflect the steady state of underlying MSMs
- ZA and BC PIGS explore different areas of phase space with enhanced rates. Their overlap is limited to the space sampled by CS

PIGS: Focused conformational sampling - Discovery of states (torsional assignments on 3-residues-long segments)

By scanning 3-residues-long segments along the sequence of bromodomains, we count how many unique torsional states\(^1\) are found in a data set (i.e. by ZA PIGS, BC PIGS and CS separately)

\(^1\)Vitalis, A. and Pappu, R.V., 2014. A simple molecular mechanics integrator in mixed rigid body and dihedral angle space. JCP.
PIGS: Focused conformational sampling - Discovery of states (torsional assignments) and $\alpha$-content

- Focused explorations: the PIGS enhancement allows discovering more states for precisely the residues in the PIGS representation.

PIGS: Focused conformational sampling - Mean First Passage Times

- Clustering on either BC loop or ZA loop torsional angles
- MFPTs to the node representative of PDB structure
- 3 MFPT curves per representation, e.g., BC loop MFPTs for BC PIGS, ZA PIGS, and CS
- MFPTs are much larger when the enhanced and the analyzed segments match
- PIGS MFPTs exceed CS values for 70%-90% of the conformational space
- States discovered by PIGS span time scales inaccessible by CS
PIGS: Focused conformational sampling - Life-time of metastable states

- RMSD (Helices+Loops) to initial states. Exit events measured at crossing of lower-threshold + 1.0 Å. Lower threshold is varied between 0.5 and 3.0 Å. Life-times are (under)estimated by MLE from EXP-fits (no corrections)

- The metastability of states discovered by PIGS is \( \sim \) to the metastability of crystal structures (\( \tau \in [10:60] \text{ns} \))
Summary: Key properties of PIGS (2014-2018)

- Small sensitivity to parameters and remarkable increase of phase-space coverage

1-Dimensional Toy Model

- Length of stretch ($f_p$) varies
- Number of saved snapshots ($n_0$) varies
- Temperature ($T$) varies

ClD

 φ-PIGS

r-PIGS

REX

Cluster fractional population

Marco Bacci @ PASC 2018
Summary: Key properties of PIGS (2014-2018)

- Small sensitivity to parameters and remarkable increase of phase-space coverage
- Continuous unperturbed stretches up to tens of ns are available despite reseedings

![Graph showing cumulative distribution function]
Summary: Key properties of PIGS (2014-2018)

- Small sensitivity to parameters and remarkable increase of phase-space coverage
- Continuous unperturbed stretches up to tens of ns are available despite reseedings
- Pathways information is substantially preserved (kinetics and mechanics)
Summary: Key properties of PIGS (2014-2018)

- Small sensitivity to parameters and remarkable increase of phase-space coverage
- Continuous unperturbed stretches up to tens of ns are available despite reseedings
- Pathways information is substantially preserved (kinetics and mechanics)
- The thermodynamic bias can be large (not easy to verify) → reweighting strategies open a new opportunity of research

Markov State Models

![Graphs showing normalized frequency and probability density distributions of different models.](image)
**Summary: Key properties of PIGS (2014-2018)**

- Small sensitivity to parameters and remarkable increase of phase-space coverage
- Continuous unperturbed stretches up to tens of ns are available despite reseedings
- Pathways information is substantially preserved (kinetics and mechanics)
- The thermodynamic bias can be large (not easy to verify) → reweighting strategies open a new opportunity of research
- A greater number of replicas implies a faster exploration: rational usage of computational resources
Summary: Key properties of PIGS (2014-2018)

- Small sensitivity to parameters and remarkable increase of phase-space coverage
- Continuous unperturbed stretches up to tens of ns are available despite reseedings
- Pathways information is substantially preserved (kinetics and mechanics)
- The thermodynamic bias can be large (not easy to verify) → reweighting strategies open a new opportunity of research
- A greater number of replicas implies a faster exploration: rational usage of computational resources
- The choice of the representation of the system can be tuned to answer specific questions


**Introduction**

The bromodomain-peptide (un)binding network is essential for many processes in biology, including gene regulation and protein-protein interactions. In recent years, computational methods have been developed to predict and analyze these interactions, providing insights into the mechanisms of action of bromodomain-containing proteins.

**Progress - Index Guided Sampling**

Index guided sampling is a method that uses a Markov chain Monte Carlo approach to sample the conformational space of a protein. This method has been applied to predict the binding modes of bromodomain-peptide complexes, providing insights into the selectivity and specificity of bromodomain interactions.

**Applications**

Recent studies have shown that index guided sampling can be used to identify metastable states in high-dimensional time-series data, which can be used to predict the long-term behavior of complex systems. This method has potential applications in various fields, including drug discovery, molecular dynamics, and systems biology.

**Closing**

The presentation concludes by highlighting the potential of index guided sampling in predicting metastable states and suggests future directions for research in this area.
Acknowledgements

Many thanks to
- Amedeo Caflisch
- Andreas Vitalis
- Cassiano Langini
- Francesco Cocina
- Davide Garolini
- Nicolas Blöchliger
- Caflisch group
- Organizers PASC18
- All of you for your attention

Funding
- SNSF
- PASC
- CSCS (computing time)