Putting ExTASY in charge of an arduous computational challenge

Extensible Tools for Advanced Sampling and Analysis

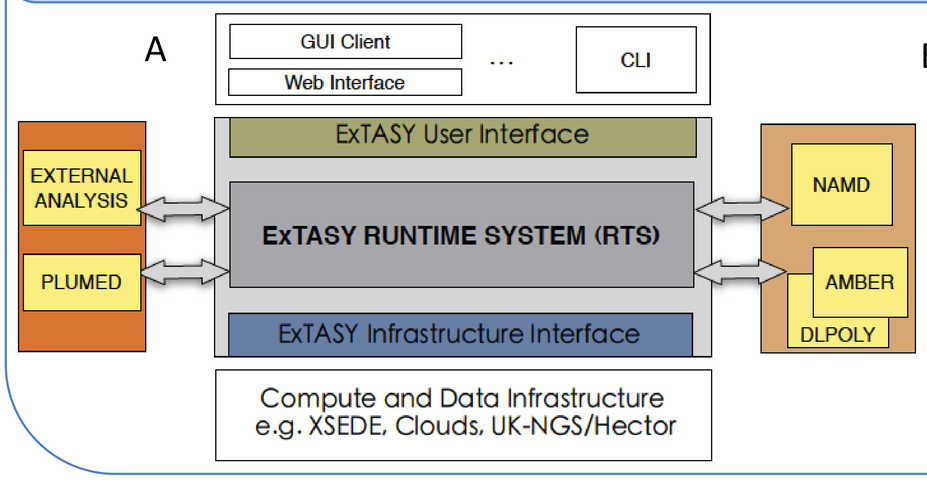
UNITED KINGDOM · CHINA · MALAYSIA

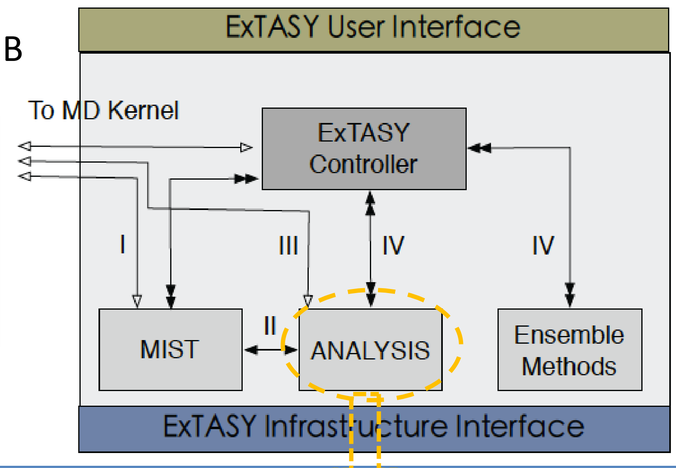
Ardita Shkurti[1], Charles Laughton[1], Ramon Goñi[2,3], Iain Bethune[4], Elena Breitmoser[4], Shantenu Jha[5], Cecilia Clementi[6,7], Ben Leimkuhler[8], Panos Parpas[9], Mauro Maggioni[10]

[1] *School of Pharmacy, Centre for Biomolecular Sciences, The University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom*
[2] *Department of Life Sciences, Barcelona Supercomputing Center, J. Girona 29, Barcelona, 08034, Spain*
[3] *Joint BSC-CRG-IRB Program in Computational Biology, Barcelona, Spain*
[4] *Edinburgh Parallel Computing Centre (EPCC), The University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom*
[5] *Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey, NJ 08854, USA*
[6] *Department of Chemistry, Rice University, Houston, TX 77005, USA*
[7] *Center for Theoretical Biological Physics, Rice University, Houston, TX 77005, USA*
[8] *School of Mathematics , The University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, United Kingdom*
[9] *Department of Computing, Imperial College, 180 Queen's Gate, London, SW7 2AZ, United Kingdom*
[10] *Department of Mathematics, Computer Science, and Electrical and Computer Engineering, Duke University, Durham, NC 27708-0320, USA*

ExTASY

ExTASY - **Ex**tensible **T**ools for **A**dvanced **S**ampling and anal**Y**sis – is a project to provide the biomolecular simulation community with a **flexible** and **extensible** software **toolkit** of advanced sampling methods for **molecular simulation**, targeted primarily at High Performance Computing (HPC) infrastructures (www.extasy-project.org).





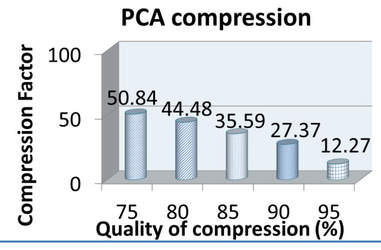
(A) Design of ExTASY. The modular elements that are developed or adopted are: **the ExTASY Runtime**, the **ExTASY interfaces** to the user and the infrastructure, and **plugins to community codes** and **analysis routines**.

(B) Architecture for the **ExTASY Runtime**, which defines the main components and the control flow between them. Closed double-headed arrows represent communication between components internal to ExTASY; arrows with open-heads represent communication from an ExTASY component to the MD Engine. ExTASY uses SAGA-based Pilot-abstractions to provide fundamental support for ensembles, and SAGA to implement the infrastructure interface to different back-end systems.

PCA

Principal **C**omponent **A**nalysis (PCA) applied to molecular simulation data:

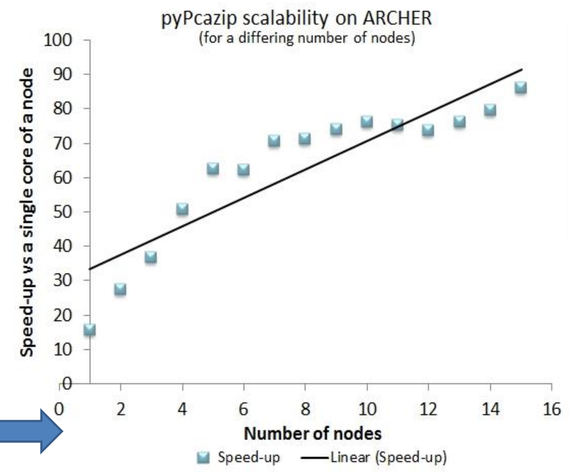
- Reduces sampling data dimensionality in order to capture the dominant modes of motion of the molecular system;
- Gives insight into structural and dynamical behaviour of molecules;
- Enables highly compressed data storage of simulation trajectory files.



pyPcazip

A **Python-based PCA analysis tool** (pyPcazip) has been developed, *amplifying the analysis functionalities* of its Fortran and C-based predecessors [1], including:

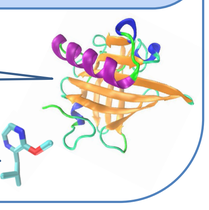
- *Better handling of memory issues* when dealing with *very large data sets*;
- *On-the-fly selection of subsets of atoms of interest* for the PCA analysis from the available data sets;
- *Flexible support* for the simultaneous analysis of *multi-trajectory datasets* that vary in their molecular *topology* and *number of atoms*;
- *MPI* support for input processing and internal calculations;
- Compliance with *HPC supercomputing* architectures such as ARCHER.



Results from the application of pyPcazip to the investigation of ligand-induced changes in molecular flexibility in the Major Urinary Protein (MUP) are shown, extending our previous work [2].

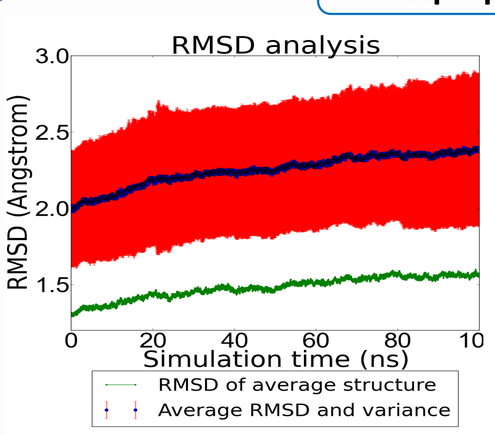
MUP secondary structure

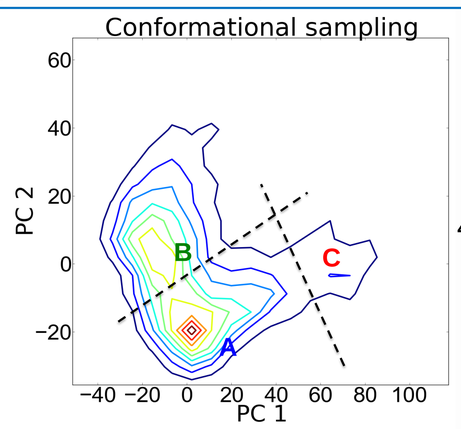
Structure of pheromone isobutylmethoxypyrazine

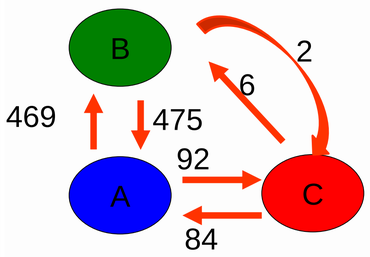


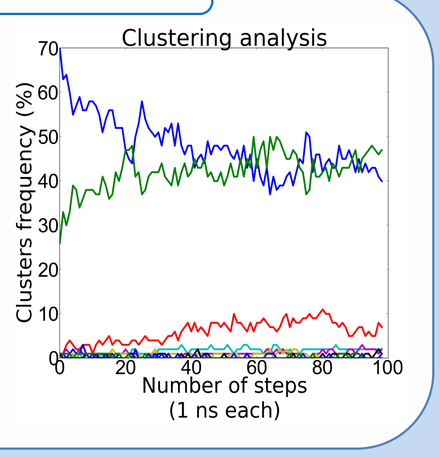
Results

Apo-protein: 100 independent simulations; 100 ns per simulation.

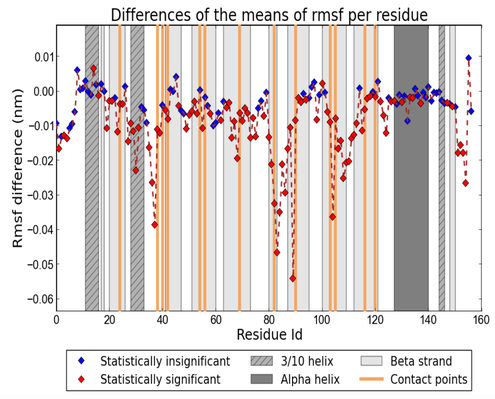




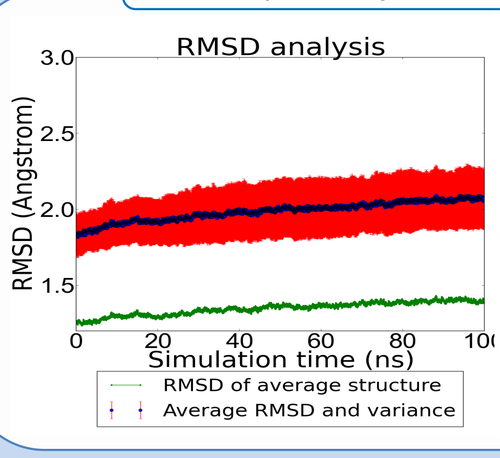


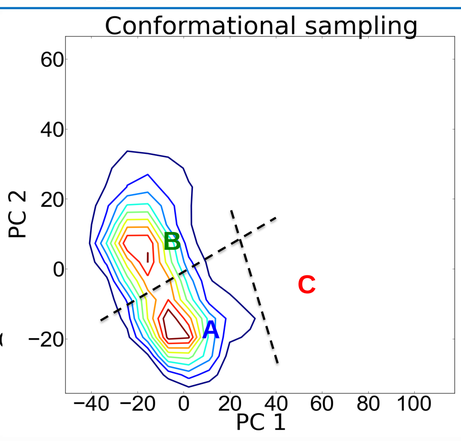


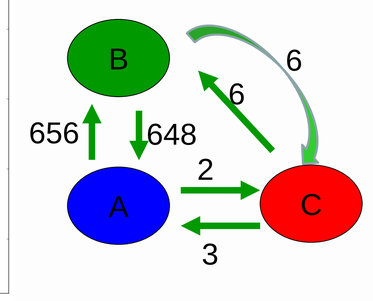
Statistical analysis

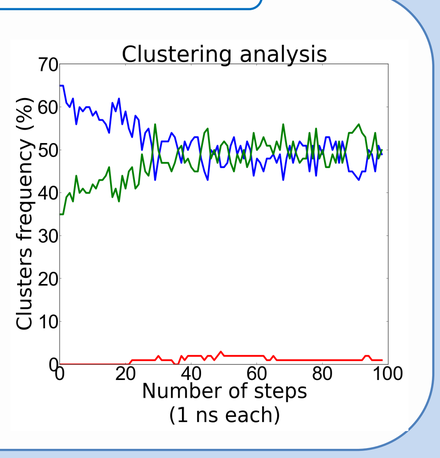


Complex of **protein and ligand**: 100 independent simulations; 100 ns per simulation.









- **Rapid** and **quantitative analysis** of molecular simulation data is shown, especially on **equilibration** and **convergence** (the dotted lines represent the clusters boundaries of the PCA projections);
- Micro-level of ligand-induced residue-specific changes in dynamics is reported: More than **60%** of the **rmsf difference** of residues of the protein is **significant** at a 5% level (contact points represent residues containing atoms less than 3 Angstroms away from any atoms of the ligand).

References

[1] Meyer, T. ; Ferrer-Costa C. ; Perez A. ; Rueda M. ; Bidon-Chanal A. ; F. Luque J. ; Laughton C. A. ; Orozco M. *J. Chem. Theory Comput.*, *2 (2)*, 251–258, **2006**

[2] Roy, J; Laughton, CA; *Biophys J.* ; **99**(1): 218-226, **2010**

Acknowledgements

The University of Nottingham

EPSRC Engineering and Physical Sciences Research Council

Grant code: EP/K039490/1

NSF

CBS Centre for Biomolecular Sciences