Ministry of Human Resources and Social Security

JÜLICH
FORSCHUNGSZENTRUM

2016 Jülich – OCPC – Programme

for the involvement of postdocs in bilateral collaboration projects

PART A

Title of the project:	A parallel data analysis framework for large scale pattern mining in biomolecular simulations
Jülich's institute:	Jülich Supercomputing Centre (JSC)
Project leader:	Dr. Olav Zimmermann, Simulation Laboratory Biology Email: olav.zimmermann@fz-juelich.de

Web-address:
http://www.fz-juelich.de/SharedDocs/Personen/IAS/JSC/EN/staff/zimmermann_o.html
http://www.fz-juelich.de/ias/jsc/EN/AboutUs/Organisation/ComputationalScience/Simlabs/slbio/_node.html

Description of the project (max. 1 page)[1]**:** see overleaf

Description of existing or sought Chinese collaboration partner institute (max. half page)**:**
The research areas of a best matching collaboration partner institute include protein folding and structure prediction; protein structure analysis, classification and clustering. The institute should also have a perfectly fitting profile with first grade experience in high performance computing and machine learning.

Required qualification of the post-doc:

- PhD in Computer Science, Bioinformatics, Physics or Mathematics
- Experience with C++ required
- Additional skills in Molecular Simulation, Python, Software Engineering, Statistics, Machine Learning, or Geometry are a plus.

PART B

Documents to be provided by the post-doc:

- Detailed description of the interest in joining the project (motivation letter)
- Curriculum vitae, copies of degrees

---

[1] Please add overleaf

- List of publications
- 2 letters of recommendation

PART C

Additional requirements to be fulfilled by the post-doc:

- Max. age of 33 years
- PhD degree not older than 5 years
- Very good command of the English language
- Strong ability to work independently and in a team

Description of the project:

Molecular simulation has become an indispensible tool to study the molecular mechanisms that underlie biological function. To study biological processes that involve large conformational changes like protein folding and peptide aggregation atomistic Markov chain Monte Carlo Methods (MCMC) has recently been demonstrated as a computationally efficient alternative to molecular dynamics (MD) simulations.

In our group ProFASi (Protein Folding and Aggregation Simulator) has been developed, a highly efficient and scalable Open Source software package that implements a range of modern methods for atomistic Monte Carlo simulation of biomolecules. Using the supercomputers at JSC, it has recently been demonstrated to be able to simulate the folding of a protein which has been determined experimentally to take one second [1]. This is orders of magnitude beyond the longest molecular dynamics simulations which rarely exceed 1 ms.

While a range of software packages for analysis of MD trajectories exists, these can typically not be used to analyze MC simulations as the primary order criterion in MD simulations is time. MCMC simulations make large nonphysical simulation steps. Accordingly the samples do not contain any time information. Analysis of large MCMC sample sets without time ordering information involves idiosyncratic computing challenges not addressed by available tools for MD analysis.

The data volumes of MCMC simulations reach several TBytes per simulation making highly parallel implementation of analysis methods and sophisticated indexing techniques imperative. Methods that will be implemented in the framework will need, in particular, to tackle high dimensional data, as the primary simulation output, millions of molecular conformations, each have up to tens of thousands of variables. New distance measures, clustering methods and inference algorithms will have to be developed or adapted to the data characteristics. Among the methods that have been shortlisted to become part of the framework are subspace clustering, spectral clustering, and nonlinear dimension reduction.

The framework will be employed to get insight into the characteristics and causal events of various biological processes. Among these processes for which such knowledge is currently elusive are protein folding and misfolding, the latter being implicated in a range of neurodegenerative diseases. Intrinsically disordered proteins for which experimental data is scarce although they are involved in important biological processes, including cancer, have also been studied successfully using ProFASi [2]. Their apparent lack of a structure defies typical methods of structure analysis.

Long term plans include the proposed project for implementing a user facility for large scale storage and high performance analysis of biomolecular simulations which would enable the ability to compare different simulations strategies, find common structure formation patterns in a diverse protein family or differnces of structure preferences for different mutations of the same protein. The analysis outputs will also fuel current and future machine learning based approaches in our lab for protein structure prediction, force field improvement and synthetic biology.

A small software prototype, largely tailored for the analysis of peptide aggregation information, was recently developed in collaboration with the Computer Science Department at RWTH Aachen, supported by a JARA HPC Seed Fund Grant. This implementation, written in modern C++ could serve as a blueprint for the development of the proposed analysis framework.

[1] Mohanty, S., Meinke, J.H., Zimmermann, O. (2013) Folding of Top7 in unbiased all-atom Monte Carlo simulations. Proteins 81:1446–1456.

[2] Jónsson, S.A., Mohanty, S., Irbäck, A. (2012) Distinct phases of free α-synuclein–a Monte Carlo study. Proteins 80:2169-77.