HPsCreen

High Performance Molecular Screening at Massive Scale

Programm / Ausschreibung	IKT der Zukunft, EuroHPC, IKT der Zukunft - EuroHPC 2019	**Status**	abgeschlossen
Projektstart	01.11.2022	**Projektende**	31.10.2023
Zeitraum	2022 - 2023	**Projektlaufzeit**	12 Monate
Keywords	pharmacology; high performance computing; virtual screening; scheduling; performance optimization		

Projektbeschreibung

LigandScout ist ein ausgereiftes Softwarepaket für das Moleküldesign in der Frühphase pharmazeutischer Forschung, das u.a. zwei wichtige Methoden zur Identifizierung potenziell bioaktiver Moleküle mittels in-silico-Methoden beinhaltet: (1) Konformerberechnung von Molekülstrukturen und (2) virtuelles Screening (VS) von Moleküldatenbanken durch Alignment-Experimente der generierten Konformationsmodelle auf funktionsbasierte Pharmakophormodelle.
LigandScout beinhaltet bereits effiziente Implementierungen für beide Methoden und unterstützt zudem elastisches Cloud Computing und die Nutzung von Compute- Clustern. Obwohl LigandScout bereits eine grundlegende Unterstützung für die Nutzung paralleler Maschinen bietet, wollen wir die Grenzen des virtuellen Screenings erweitern, um die Nachhaltigkeit des Wirkstoffentdeckungsprozesses zu erhöhen: Je mehr Verbindungen in einem bestimmten Zeitraum effektiv gescreent werden, desto mehr kostenintensive in-vitro Experimente können vermieden werden.
Das Ziel des geplanten Projekts ist es deshalb, LigandScout auf massiv-paralleler Hardware so ausführen zu können, dass der Gesamtdurchsatz beim virtuellen Screening optimiert wird. Wenngleich der Berechnungsprozess beim virtuellen Screening recht einfach parallelisierbar ist, müssen drei Herausforderungen gemeistert werden, um LigandScout auf Großrechnern effektiv laufen zu lassen. Zuerst muss eine entsprechende Skalierbarkeit bei der Ausführung des Programms sichergestellt werden, d.h. alle Recheneinheiten müssen effizient, ohne großen Kommunikationsoverhead, genutzt werden. Es gilt überdies, einzelne Teile der Moleküldatenbanken so an die Recheneinheiten zu senden, dass eine gute Lastbalancierung und eine sichere Übertragung der Daten gewährleistet werden kann. Zuletzt muss die Software so angepasst werden, dass Berechnungen nicht direkt ausgeführt werden, sondern als "Job" in einer von vielen Nutzern geteilten Warteschlange eines Großrechners verweilen kann.
Um die genannten Ziele zu erreichen, werden wir in einem ersten Schritt, die Ausführung von LigandScout auf Großrechnern (z.B. dem Vienna Scientific Cluster) ermöglichen und im Anschluss dessen parallele Skalierbarkeit untersuchen. Dazu werden wir geeignete Profiling- und Tracing-Werkzeuge einsetzen. Mit Hilfe dieser Analysen können wir Flaschenhälse bei der Programmausführung charakterisieren und entsprechend Lösungen zur Vermeidung finden. Ein zentrales Problem ist hierbei das Scheduling (die Arbeitsaufteilung) der Moleküle auf die einzelnen Recheneinheiten. Im letzten Schritt wenden wir uns der einfachen Nutzbarkeit von HPC-Ressourcen zu. Unser Ziel ist es, den Nutzer:innen eine Vorhersage zu liefern, wie lang ein Rechenjob (für ein virtuelles Screening) mit unterschiedlich vielen Recheneinheiten dauern wird und wie hoch die

zugehörigen Gesamtkosten sein werden (inkl. des CO2-Verbrauchs).

Abstract

LigandScout is an advanced molecular design software package that supports two important methods for identifying potentially bio-active molecules via in-silico methods in early pharmaceutical research: (1) conformer generation of molecular structures and (2) virtual screening of molecules using alignment experiments using the generated conformational models and 3D chemical feature-based pharmacophore models. Although LigandScout already has basic support for using parallel machines and elastic cloud computing, we want to push the limits of virtual screening in order to increase the sustainability of the drug discovery process, because the more compounds can be screened effectively in a time period, the more cost-intensive in-vitro experiments can be avoided.
For that reason, the goal of the proposed project is to allow LigandScout to be applied at massive scale, aiming to optimize the overall throughput of the virtual screening process. Although the virtual screening process can efficiently be parallelized, we face three main challenges when targeting large-scale, public supercomputers. First, we need to ensure that a high parallel efficiency can be maintained in order to utilize the computational resources in the best way possible, without introducing a large communication overhead. Second, effective data management is key for high performance, i.e., we need to transfer the right number of molecules to each worker process to guarantee a good load balancing of the computational work. Third, we need to consider the possibility of longer queuing times when running jobs on large-scale supercomputers, as they are equipped with batch schedulers (e.g., SLURM) to fairly share the resources among the users. Thus, the uncertainty of the completion time of jobs has to be taken into account.
In the presented project, we will address each of the challenges mentioned above. In a first step, we will analyze the parallel scalability of LigandScout on current supercomputers such as the VSC. To that end, we can apply typical performance analysis tools for the profiling and tracing of HPC codes. In a second step, we will examine how to optimize the scheduling of individual tasks that form a virtual screening process, i.e., how many molecules from the main database need to be sent to each worker node in the system. In a last step, we will investigate how the completion time of a screening jobs and its associated costs (incl. CO2 consumption) can be estimated, considering the current queue length of the actual batch scheduler, i.e., evaluating the trade-offs between running on a few compute nodes but starting earlier or waiting for more resources while gaining a larger overall throughput. In particular, the last goal of giving users a predicted completion time is novel and very challenging, but a successful implementation will significantly foster the utilization of HPC resources in the traditional, computational workflows.

Projektkoordinator

- Technische Universität Wien Institute of Computer Engineering

Projektpartner

- Inte:Ligand Software-Entwicklungs- und Consulting GmbH