



[Prof. Dr. Alexandros Stamatakis // Hochleistungsrechnen in den Lebenswissenschaften]

Alexandros Stamatakis studierte von 1995–2001 an der TU München Informatik mit Studienaufenthalten in Athen, Paris, Lyon und Madrid. Er promovierte dort im Jahr 2004 über verteilte und parallele Algorithmen zur Berechnung großer Stammbäume.

Nach seiner Tätigkeit als Postdoc am Institute for Computer Science in Heraklion, Griechenland arbeitete er als Postdoc an der ETH Lausanne.

Im Jahr 2008 kehrte er nach Deutschland zurück und leitete bis 2010 eine Emmy Noether-Nachwuchsforschungsgruppe, zunächst an der LMU und dann an der TU München.

Im Jahr 2010 übernahm er die Leitung einer Forschungsgruppe am Heidelberger Institut für Theoretische Studien (HITS). Das Institut ist eines der wenigen privat finanzierten Forschungsinstitute in Deutschland und seine Organisationsstruktur entspricht in etwa der eines Max-Planck-Instituts. Im Jahr 2012 wurde er zusätzlich zu seiner Funktion am HITS auf die Professur für Hochleistungsrechnen in den Lebenswissenschaften am Karlsruher Institut für Technologie (KIT) berufen. Seit April 2021 ist er Affiliated Scientist des Evolutionary Genetics and Paleogenomics (EGP) lab am Institute of Molecular Biology and Biotechnology der Foundation for Research and Technology Hellas in Heraklion, Kreta. Er ist Mitglied des Lenkugausschusses des Höchstleistungsrechners am Leibniz-Rechenzentrum in München. Sein Forschungsschwerpunkt ist die Entwicklung skalierbarer Software, Modelle und Algorithmen für die Evolutionsbiologie.

// Überblick und Allgemeines

Unsere Forschungsgruppe arbeitet an der Schnittstelle von Algorithmik, statistischer Modellierung, Parallelverarbeitung und Evolutionsbiologie. Anders als in der reinen Informatik geht es nicht primär um Proof-of-Concept Implementierungen, sondern Programme, die von Biologinnen und Biologen weltweit zur Datenanalyse eingesetzt werden können, zur Verfügung zu stellen. Das übergeordnete Ziel ist es, Forschung in der Evolutionsbiologie durch entsprechende Algorithmen und Implementierungen zu ermöglichen.

Besonders wichtig ist uns die freie Verfügbarkeit unserer Programme für die Wissenschaft in Form von open source codes und kostenlos benutzbarer Web-Services. Dieses spiegelt sich auch in unserer Publikationsstrategie wieder, da wir nur noch Konferenzen und Zeitschriften berücksichtigen, welche Vorabdrucke wissenschaftlicher Arbeiten auf sogenannten preprint-Servern (etwa arXiv oder bioRxiv) zulassen. Damit ist sichergestellt, dass jeder freien Zugang zu unseren Forschungsergebnissen hat. Eine besondere Herausforderung, auch für das Hochleistungsrechnen, stellt die zunehmende molekulare Datenflut dar. Aufgrund bahnbrechender und kontinuierlicher technologischer Fortschritte im Bereich der DNA-Sequenzierung seit etwa 2007 sinken die Kosten zur Se-

quenzierung von Genomen gegenwärtig wesentlich schneller als die entsprechenden Datenanalysekosten, basierend auf Moores Gesetz.

Darüber hinaus ändern sich mit jeder neuen und günstigeren Sequenzierungstechnologie auch die Charakteristika der zu analysierenden Daten. Daher stehen wir vor zwei grundlegenden Herausforderungen: der Bewältigung der Datenflut, z. B. die Analyse aller sequenzierten SARS-CoV-2 Genome, durch effizientere Algorithmen und den Einsatz von Hoch- und Höchstleistungsrechnern sowie der Anpassung existierender Methoden an die sich permanent verändernden Eigenschaften der Ausgabedaten von DNA-Sequenzierern.

Eine weitere große Herausforderung liegt in der Validierung wissenschaftlicher Software im Bereich der Evolutionsbiologie. Aufgrund der zunehmenden Quantifizierung in der Biologie und des stärkeren Fokus auf die Datenanalyse, statt wie bis vor einigen Jahren auf die Datenakquise, wird für die meisten biologischen Arbeiten eine Vielzahl von Programmen eingesetzt. Die Softwarequalität dieser Codes, auf denen die meisten biologischen Schlussfolgerungen basieren ist gegenwärtig aber allenfalls mittelmäßig. Insofern dürfen und sollten auch die Ergebnisse dieser quantitativen Analysen in Frage gestellt werden. Ein weiteres unserer Ziele ist es daher Lösungen für dieses Problem aufzuzeigen.

// **Neueste Ergebnisse**

Die Beobachtung, dass die Berechnungen von Gewinnwahrscheinlichkeiten bei Eliminationsturnieren (z. B. Fußball) algorithmisch große Ähnlichkeit mit den Wahrscheinlichkeitsberechnungen auf-

weist, welche wir üblicherweise auf phylogenetischen Bäumen durchführen, haben zur Entwicklung eines neuen effizienten Verfahrens zur exakten Berechnung der Turniergewinnwahrscheinlichkeiten eines jeden teilnehmenden Teams geführt. Die effiziente und exakte Berechnung dieser Wahrscheinlichkeiten ermöglicht nun die Abschätzung der Varianz dieser Gewinnwahrscheinlichkeiten wenn man diese paarweisen leicht variiert (z. B., wie wahrscheinlich es ist, dass England gegen Deutschland gewinnt).

// **Ausgewählte Publikationen**

B Morel, P Schade, S Lutteropp, TA Williams, GJ Szöllösi, A Stamatakis: SpeciesRax: a tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *Molecular Biology and Evolution*, 2022.

S Lutteropp, C Scornavacca, AM Kozlov, B Morel, Alexandros Stamatakis: NetRAX: Accurate and Fast Maximum Likelihood Phylogenetic Network Inference. *Bioinformatics*, 2022.

J Haag, D Höhler, B Bettisworth, A Stamatakis: From Easy to Hopeless-Predicting the Difficulty of Phylogenetic Analyses. *bioRxiv preprint*, 2022.

// **Mitarbeiterinnen und Mitarbeiter**

Verwaltungspersonal

Simone Meinhart

Wissenschaftliches Personal

Ben Bettisworth (extern)

Julia Haag (extern)

Dimitri Höhler (extern)

Lukas Hübner

Dr. Alexey Kozlov (extern)

Dr. Benoit Morel

Anastasis Togkousidis (extern)

// **Website**

www.exelixis-lab.org