



*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 zusätzlich zu seiner Forschungsgruppe bis 2013 auch die Leitung der Systemadministration und des Rechenzentrums 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.*

*Er ist Mitglied des Lenkungsausschusses 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 darum, 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 revidierten Publikationsstrategie wieder, da wir nur noch Konferenzen und Zeitschriften berücksichtigen werden, welche Vorabdrucke wissenschaftlicher Arbeiten auf sogenannten preprint-Servern (etwa arxiv oder bioarxiv) 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 Sequenzierung 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 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.

## ERGEBNISSE UND ERFOLGE

Ein wichtiges Ereignis im Jahr 2018 war der Release von RAxML-Next Generation (RaxML-NG). RAxML-NG stellt eine komplette Reimplementierung unserer weltweit genutzten Software RAxML zur Berechnung phylogenetischer Stammbäume basierend auf dem Maximum Likelihood Verfahren dar. Diese Reimplementierung trägt auch unseren eigenen Ansprüchen an die Softwarequalität Rechnung, da der alte Code, der von A. Stamatakis über 15 Jahre lang entwickelt wurde, nicht mehr wartbar war.

Darüber hinaus haben wir ein neues Tool namens ParGenes entwickelt und zur Verfügung gestellt, welches die parallele Berechnung sogenannter Genstammbäume auf Rechnerclustern effizient bewerkstelligen kann und eine nahezu optimal Lastbalancierung erreicht. Ziel ist es, zunächst für jedes Gen aus den Genomen der Spezies, welche man analysieren möchte, einen separaten Stammbaum zu berechnen. Diese Genstammbäume werden dann in einem zweiten Schritt zu einem Speziesstammbaum zusammengefasst.

## AUSGEWÄHLTE PUBLIKATIONEN

A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis: RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *BioRxiv*, 447110, 2018.

P. Barbera, A. M. Kozlov, L. Czech, B. Morel, D. Darriba, T. Flouri, A. Stamatakis: EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Systematic Biology*, advance on-line publication, 2018.

B. Morel, A. Kozlov, A. Stamatakis: ParGenes: a tool for massively parallel model selection and phylogenetic tree inference on thousands of genes, *Bioinformatics*, advance on-line publication, 2018.

## MITARBEITERINNEN UND MITARBEITER

### Verwaltungspersonal

Simone Meinhart

### Wissenschaftliches Personal

Ivo Baar

Pierre Barbera (extern)

Lucas Czech (extern)

Dr. Alexey Kozlov (extern)

Sarah Lutteropp (extern)

Benoit Morel (extern)

Johanna Wegener (extern)

Adrian Zapletal (extern)