Hochleistungsrechner und der Stammbaum des Lebens

Eine wahre Flut von DNA-Daten ermöglicht inzwischen immer präzisere Rekonstruktionen von Stammbäumen – im Prinzip jedenfalls. In der Praxis überfordert die Suche nach der optimalen Lösung auch die leistungsfähigsten Computer. Die Herausforderung heißt deshalb, die Effizienz der Programme für Näherungslösungen zu steigern.

Von Alexandros Stamatakis

Die computergestützte Berechnung von Stammbäumen, welche die Verwandtschaftsverhältnisse zwischen Organismen wiedergeben, ist eine verhältnismäßig junge Disziplin. Doch reichen ihre Anfänge immerhin bis in die 1960er Jahre zurück. Für jeden Organismus beziehungsweise jede Spezies, deren Position im Stammbaum ermittelt werden soll, liegen typischerweise DNA-Daten oder Angaben zu morphologischen Merkmalen vor – etwa über die Knochenform. Bei Bakterien kann es sich auch um chemische Eigenschaften handeln, die für die jeweilige Spezies charakteristisch sind.

Das Ziel besteht darin, anhand geeigneter Modelle denjenigen Stammbaum zu rekonstruieren, der am besten zu den vorliegenden Daten passt. Mathematisch gesehen, handelt es sich also um ein Optimierungsproblem. Dahinter steckt die stillschweigende Annahme oder Hoffnung, dass der »optimale« Stammbaum auch der wahre ist. An seinen Blättern befinden sich die Organismen, für welche DNA-Daten vorliegen. Die inneren Knoten – sprich: Verzweigungen – repräsentieren hypothetische gemeinsame Vorfahren.

Von diesen existieren in der Regel keine DNA-Daten, weil sich normalerweise nur aus lebenden Organismen Erbsubstanz gewinnen lässt. Allerdings gab es in letzter Zeit bedeutende Fortschritte bei der Sequenzierung alter DNA; dadurch ist es insbesondere der Gruppe um Svante Pääbo vom Max-Planck-Institut für evolutionäre Anthropologie in Leipzig gelungen, das Neandertalergenom zu entziffern.

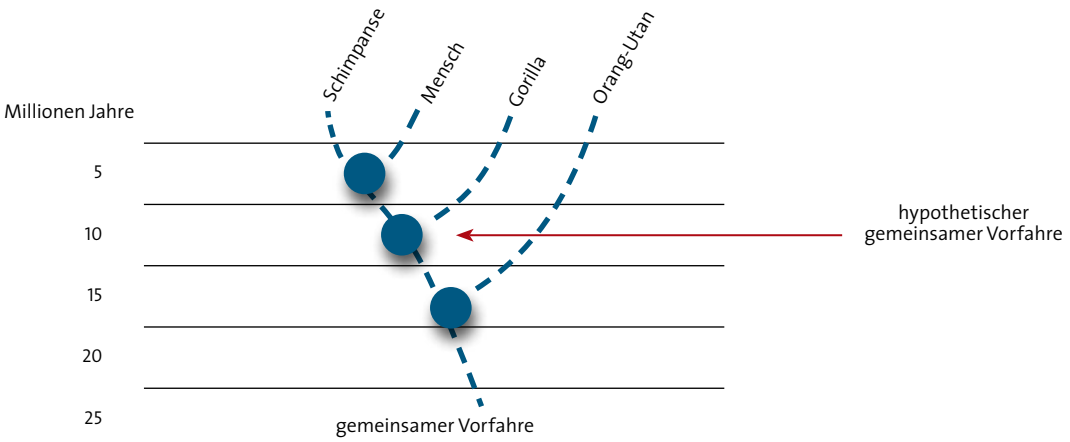
Betrachten wir ein klassisches Beispiel: den Stammbaum von Mensch, Schimpanse, Gorilla und Orang-Utan. Der auf DNA-Sequenzen beruhende Eingabedatensatz könnte, grob vereinfacht, dann so aussehen:

Mensch	AAACCCCGTTTTT
Gorilla	AAACTTTAAGGGT
Schimpanse	AAGATTCGTTTTT
Orang-Utan	AGAATCCGTTTGT

Dabei stehen die Buchstaben für die Basen Adenin, Thymin, Cytosin und Guanin, die das genetische Alphabet ausmachen. Ein möglicher Stammbaum für diese Daten ist im Kasten unten gezeigt. Dabei bleibt offen, wo der gemeinsame Vorfahr aller Menschenaffen, das heißt die Wurzel des Baums, liegt. Diese wird zur Vereinfachung der mathematischen Modelle üblicherweise weggelassen.

Grundlage für die Optimierung ist eine abstrakte Funktion f, eine Rechenvorschrift, die zu einem gegebenen Stammbaum und zu gegebenen DNA-Daten einen Zahlenwert liefert: die »Plausibilität« *(likelihood)*. Je höher dieser Wert, desto besser ist der Stammbaum mit den Daten vereinbar. Wenn man also drei

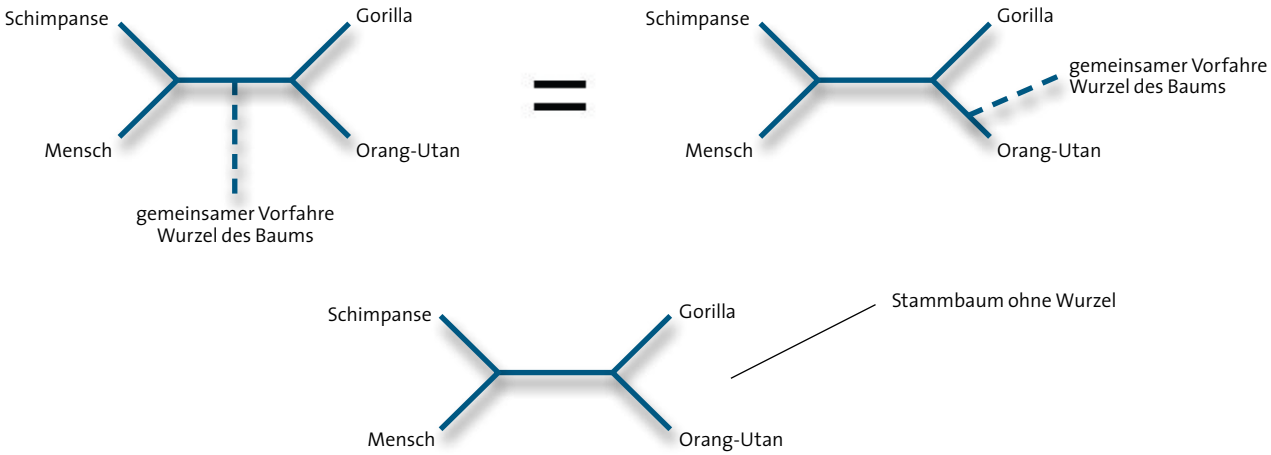
Der DNA-Stammbaum der Menschenaffen



ALLE ABBILDUNGEN DIESES ARTIKELS: ALEXANDROS STAMATAKIS

Stammbäume ohne Wurzel

Anhand von DNA-Daten der Menschenaffen allein lässt sich keine Aussage über die Wurzel des Stammbaums machen. Sie kann an den verschiedensten Stellen liegen (oben). Dem trägt die Darstellung ohne Wurzel Rechnung (unten).



Stammbäume in Betracht zieht, muss man die Funktion für alle drei berechnen. Der optimale Baum ist dann derjenige, für den der größte Wert herauskommt.

In unserem Beispiel mit den Menschenaffen lässt sich dieses Problem leicht lösen, weil für vier Organismen nur drei unterschiedliche wurzellose Bäume existieren (Kasten oben). Dabei erweist sich derjenige, bei dem der Mensch mit dem Schimpansen näher verwandt ist als beide mit dem Gorilla und dem Orang-Utan, als plausibelste Lösung. Doch wie sieht die Funktion f aus? In der Praxis benutzt man dafür statistische Modelle, die auf Schätzungen beruhen, wie wahrscheinlich Mutationen sind, bei denen eine der vier Basen durch eine andere ersetzt wird.

Das grundsätzliche Problem bei diesem Verfahren ist, dass die Anzahl der möglichen Bäume extrem stark mit der Anzahl der enthaltenen Spezies zunimmt. So beläuft sie sich bei 50 Arten, was heutzutage noch eine relativ kleine Zahl ist, bereits auf $2{,}84 \cdot 10^{76}$ Kandidaten. Für jeden von ihnen müsste der Wert der Funktion f berechnet werden, denn es gibt keinen Trick, einen Großteil davon von vornherein auszuschließen. Unter der optimistischen Annahme, dass diese Berechnung für einen Baum mit 50 Organismen eine Sekunde Rechenzeit benötigt, würde die Evaluierung aller Bäume auf einem einzelnen Prozessor $9 \cdot 10^{68}$ Jahre dauern. Selbst mit der gesamten Rechenkraft auf der Erde wäre diese Aufgabe vermutlich nicht innerhalb eines vernünftigen Zeitraums zu schaffen.

Optimierungsprobleme, für die der Bedarf an Rechenzeit derart schnell zunimmt, kommen in vielen Bereichen der Informatik vor und heißen NP-vollständig. Peter Gritzmann und René Brandenberg beschreiben sie in ihrem populärwissenschaftlichen Buch »Das Geheimnis des kürzesten Weges« auf für Laien gut verständliche Art und Weise.

Automatische Suchverfahren

Da das Problem nicht exakt lösbar ist, behilft man sich mit so genannten heuristischen Suchverfahren, die zwar nicht die beste, aber zumindest eine ziemlich gute Lösung liefern. Leider gibt es bei der Berechnung von Stammbäumen keine Möglichkeit, mit Sicherheit zu sagen, wie weit das Ergebnis einer solchen approximativen Suche vom Optimum entfernt ist. Deshalb ist es unerlässlich, dass Biologen den gefundenen Baum anhand ihres Wissens auf Plausibilität prüfen.

Man kann das Suchverfahren auch an sehr schnell evolvierenden Organismen wie etwa Viren testen, deren Stammbaum über die letzten Jahre bis Jahrzehnte bekannt ist. Auch im Erfolgsfall bietet das jedoch keine Gewähr dafür, dass die Methode bei Lebewesen, die sich im Verlauf von Jahrmillionen entwickelt haben, genauso gut funktioniert.

Ein weiterer Unsicherheitsfaktor ist die Programmverifikation. Selbst wenn die Evolutionsmodelle und Rekonstruktionsverfahren perfekt sind, heißt das nicht, dass sie auch korrekt auf dem Computer umgesetzt wurden. Durch die starke Zunahme rechnerbasierter Datenanalysen in der Biologie haben Fehler in Veröffentlichungen, die auf Programmierfehlern beruhen, in jüngster Zeit stark zugenommen. Zusammen mit meinem Doktoranden Fernando Izquierdo-Carrasco habe ich die Probleme der Verifikation von Stammbäumen und von Programmen zu deren Berechnung kürzlich ausführlich dargelegt *(Briefings in Bioinformatics 12, S. 270)*.

Trotz solcher Schwierigkeiten und Unsicherheiten kommen Verfahren zur Rekonstruktion von Stammbäumen in der medizinischen und biologischen Forschung heute routinemäßig zum Einsatz. So dienen sie etwa dazu, den Ursprung von Virusepidemien zu ermitteln oder die bakterielle Zusammensetzung der Darmflora zu analysieren. Um das berühmte Zitat des russischen Genetikers Theodosius Dobzhansky (1900 – 1975) zu bemühen: »In der Biologie macht nichts Sinn, außer im Licht der Evolution.«

Was sind die aktuellen Entwicklungen und Herausforderungen auf dem Gebiet der Stammbaumberechnung? Zuallererst ist die Revolution bei der DNA-Sequenzierung zu nennen. Die Analyse des Erbguts wurde durch bahnbrechende Fortschritte in den letzten fünf bis sechs Jahren wesentlich vereinfacht und beschleunigt, so dass zugleich die Kosten

dramatisch gesunken sind. Dadurch lassen sich inzwischen auch komplette Genome einer Spezies sehr viel leichter entziffern. Während vor zehn Jahren die Sequenzierung des menschlichen Erbguts noch Schlagzeilen machte, nehmen heute selbst Biologen eher gelangweilt zur Kenntnis, dass schon wieder irgendein Genom entschlüsselt wurde.

Die Herausforderung verlagert sich daher zunehmend vom Labor zur Datenverarbeitung. Das Hauptproblem besteht darin, dass die Menge der DNA-Daten wesentlich schneller zunimmt als die Rechengeschwindigkeit der Computer oder Prozessoren zu ihrer Analyse. Das betrifft sowohl die Bioinformatik als auch ihre Teildisziplin, die rechnergestützte Ermittlung von Stammbäumen. Die Computerwissenschaftler stehen deshalb vor der schwierigen Aufgabe, immer effizientere Programme und Methoden zur Datenspeicherung und -analyse bereitzustellen.

Ohne Hoch- und Höchstleistungsrechner, in denen mehrere Einzelrechner (Prozessoren) gleichzeitig an einem Problem arbeiten, lässt sich die Datenflut vielfach nicht mehr bewältigen. Zur Rekonstruktion von Stammbäumen standen noch vor zehn Jahren lediglich die Sequenzen von ein oder zwei Genen zur Verfügung, die jeweils etwa 1000 Basenpaare umfassten. Inzwischen liegen immer öfter die weitaus umfangreicheren kompletten Genome vor. So besteht das Erbgut des Menschen aus etwa 20 000 bis 25 000 Genen; nach einigen Schätzungen sind es sogar bis zu 75 000.

Diese Datenflut stellt die Informatiker vor enorme Probleme. Das gilt insbesondere für den Speicherplatzbedarf der Programme zur Stammbaumrekonstruktion, da zur Berechnung der Bewertungsfunktion f zunehmend komplette Genome für 50 oder 100 Spezies im Arbeitsspeicher gehalten werden müssen.

Ziel: Effiziente Bewertung der Güte eines Stammbaums

Solche Programme verbringen bis zu 99 Prozent ihrer Gesamtlaufzeit damit, die Funktion f für verschiedene denkbare Bäume auszuwerten (Kasten unten). Deshalb besteht eines der Hauptziele der von mir geleiteten Scientific Computing Group am Heidelberger Institut für Theoretische Studien darin, die Zeit und den Speicherplatzbedarf für diese Aufgabe so weit wie möglich zu reduzieren.

Über die vergangenen zehn Jahre haben wir das frei verfügbare Programm RAxML *(Randomized Accelerated Maximum Likelihood)* entwickelt. Statt die Menge aller Stammbäume erschöpfend abzuarbeiten – was aussichtslos wäre –, konstruiert das Programm zu Beginn eine Anzahl von Bäumen, indem es Blatt für Blatt in zufälliger Reihenfolge an jeweils optimaler Stelle einfügt. Es versucht diese Bäume zu verbessern, indem es ganze Äste abschneidet und an anderer Stelle wieder einsetzt, das Ganze im Rahmen eines kombinatorischen Optimierungsverfahrens namens *simulated annealing*. RAxML gehört zu den fünf bis sechs weltweit am meisten benutzten Programmen zur Stammbaumrekonstruktion.

Der Webserver http://phylobench.vital-it.ch/raxml-bb/ bietet auch interessierten Laien die Möglichkeit, es auszuprobieren; ein kleiner Testdatensatz findet sich unter www.exelixis-lab.org/dna.phy.

Zur Beschleunigung der Rechnung verfolgen wir verschiedene Ansätze. So sind wir auf der Suche nach Tricks, um redundante Berechnungen zu vermeiden und Speicherplatz zu sparen. Ausgangspunkt hierfür ist die mathematische Beschreibung der Wahrscheinlichkeitsberechnungen: Wir bemühen uns, die Funktion f so zu transformieren, dass sie bei geringerem Speicherbedarf und weniger Rechenoperationen genau das gleiche Ergebnis liefert. Von großer Bedeutung ist auch, das Programm an moderne Rechnerarchitekturen anzupassen. Dadurch lassen sich die Ressourcen der eingesetzten Prozessoren besser nutzen. Das ermöglicht einen höheren Datendurchsatz und steigert so die Anzahl der evaluierten Bäume pro Sekunde.

Wir gehen allerdings auch den umgekehrten Weg und fragen uns, wie die ideale Rechnerarchitektur für unser Programm aussehen würde. In diesem Teilprojekt entwerfen wir optimale Schaltkreise zur Berechnung der Wahrscheinlichkeitsfunktion f. Zum Testen und Verifizieren unserer Architekturen benutzen wir so genannte Field Programmable Gate Arrays, bei denen es sich um eine Art programmierbare Hardware handelt. Sie bestehen aus vielen elektronischen Grundbausteinen (»Gattern«), die sich mittels einer Hardware-Beschreibungssprache dynamisch miteinander verbinden lassen, um die vorgegebene Schaltung nachzubilden.

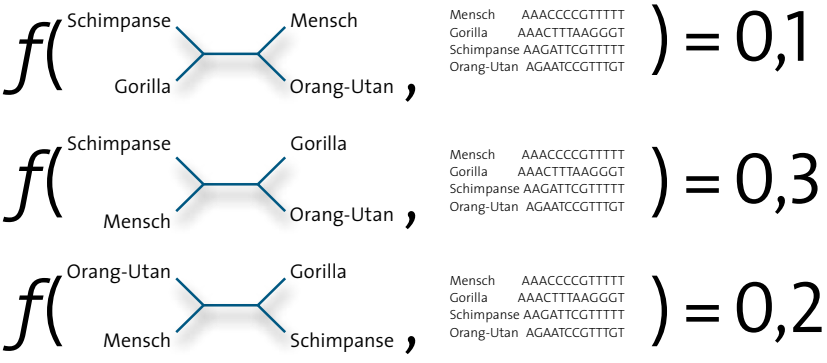
Bei all diesen Versuchen achten wir darauf, dass unsere Ergebnisse nicht nur auf RAxML anwendbar sind, sondern auch auf alle anderen likelihood-basierten Programme zur Stammbaumberechnung. Deren Geschwindigkeit hängt ja gleichfalls entscheidend davon ab, wie effizient die Funktion f auf dem Rechner umgesetzt ist.

Wie erwähnt, lassen sich sehr umfangreiche, speicherintensive Datensätze inzwischen nur noch mit Hochleistungsrechnern verarbeiten. Am HITS steht uns solch ein großer Parallelrechner zur Verfügung. Das System besteht aus 42 Rechenknoten mit je 48 Prozessoren, die durch ein leistungsfähiges Netzwerk miteinander verbunden sind.

Idealerweise gilt es, diese insgesamt 2016 Prozessoren alle gleichzeitig zu beschäftigen.

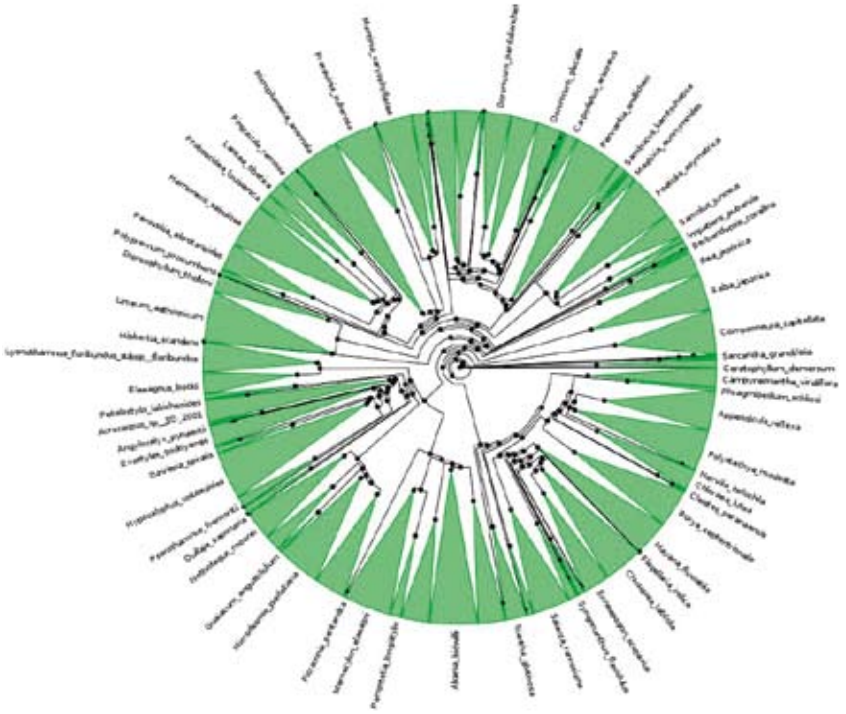
Berechnung des Verwandtschaftsgrads

Für vier Spezies existieren nur drei unterschiedliche wurzellose Stammbäume. Die Funktion f berechnet die Wahrscheinlichkeit, dass der betreffende Baum zu den DNA-Daten passt. Ihre Werte zeigen, dass Mensch und Schimpanse enger miteinander verwandt sind als mit Gorilla und Orang-Utan.



Beispiel eines ausgedehnten DNA-Stammbaums

Diesen Stammbaum für 56 000 Pflanzen errechnete die Gruppe des Autors kürzlich in Zusammenarbeit mit Forschern von der Yale University und der Brown University in den USA.



Am besten wäre es, wenn jeder von ihnen einen anderen Stammbaum evaluieren würde. Dazu müsste der einzelne Prozessor jedoch das komplette Datenmaterial im eigenen Arbeitsspeicher verfügbar haben – wozu dieser möglicherweise nicht ausreicht. Da liegt es nahe, die Aufgabe in Teilaufgaben zu zerlegen, die jede für sich nur eine relativ kleine Teilmenge aller Daten erfordern, und diese entsprechend auf die Prozessoren zu verteilen. Allerdings darf die einzelne Teilaufgabe auch nicht zu klein sein; sonst nimmt der Austausch von Daten, der vor und nach der Erledigung jeder Teilaufgabe erforderlich ist, einen zu großen Teil der Rechenzeit in Anspruch. Die Analyse und Identifizierung solcher Teilaufgaben ist nicht einfach und bildet einen der Schwerpunkte im Teilgebiet der Informatik, das sich mit der parallelen Programmierung beschäftigt.

Auch hier gilt, dass die angewandten Parallelisierungsstrategien auf alle likelihoodbasierten Programme übertragbar sein sollten und es auch sind. Mit RAxML wurden schon bis zu 1024 Prozessoren simultan zur Berechnung der Funktion f für einen einzigen Stammbaum eingesetzt, wobei diese Zahl kein Limit darstellt. Das Programm nutzt auch die Fähigkeit zur Parallelverarbeitung bei Mehrkernprozessoren, wie sie in allen neueren Laptops und Desktops zu finden sind.

Abgesehen von unseren Bemühungen, die Effizienz der Programme zur Stammbaumberechnung zu steigern, beschäftigen wir uns aber auch mit der Analyse sehr großer biologischer Datensätze. Diese interdisziplinären Projekte verbessern unser Verständnis der Biologie und helfen uns, aktuelle rechnerische oder methodische Herausforderungen zu erkennen. Beispielhaft sei hier das »plant tree of life grand challenge project« genannt, das von der Deutschen Forschungsgemeinschaft (DFG) und der National Science Foundation in den USA gefördert wird. Sein Hauptziel besteht darin, einen umfassenden Stammbaum der Pflanzen mit etwa 500 000 Spezies zu berechnen und online zur Verfügung zu stellen, so dass Biologen ihn für weiterführende Analysen nutzen können. Das ist eine Herkulesaufgabe, zumal die benötigten Daten keineswegs komplett vorliegen. Noch nie konnte ein Stammbaum dieser Größenordnung berechnet werden. Allerdings lassen sich auf dem Hochleistungsrechner des HITS mit Daten von etwa 20 Genen schon Bäume für 120 000 Spezies berechnen. In Zusammenarbeit mit Kollegen an der Yale University und der Brown University in den USA haben wir vor Kurzem einen Stammbaum der Pflanzen mit etwa 56 000 Spezies rekonstruiert und publiziert – den größten seiner Art bisher (Kasten links).

Obwohl es noch ein weiter Weg ist, kommen wir unserem Endziel, der Berechnung des Stammbaums aller Lebewesen, allmählich näher. Die stetige Verbesserung der Sequenzierverfahren und Rechnerarchitekturen lässt uns hoffen, dass wir dieses Ziel eines Tages auch erreichen werden.

DER AUTOR

Alexandros Stamatakis leitet am Heidelberger Institut für Theoretische Studien die Scientific Computing Group. Er hat an der Technischen Universität München Informatik studiert und dort im Jahr 2004 in der Informatik promoviert. Nach Postdoc-Stationen auf Kreta und an der ETH Lausanne (Schweiz) war er von 2008 bis 2010 als Nachwuchsgruppenleiter an der Ludwig-Maximilians-Universität und später an der TU München (Emmy-Noether-Programm der DFG) tätig, bevor er im Oktober 2010 ans HITS kam.

QUELLEN

Alachiotis, N. et al.: A Reconfigurable Architecture for the Phylogenetic Likelihood Function. Konferenzbeitrag, FPL Prag 2009. Online unter: http://sco.h-its.org/exelixis/nikos/publications.html

Gritzmann, P., Brandenberg, R.: Das Geheimnis des kürzesten Weges: ein mathematisches Abenteuer. Springer, Berlin/Heidelberg 2004

Ott, M. et al.: Large-Scale Maximum Likelihood-Based Phylogenetic Analysis on the IBM BlueGene/L. In: Proceedings of IEEE/ACM Supercomputing (SC2007) Conference, Reno, Nevada, November 2007

Stamatakis, A., Izquierdo-Carrasco, F.: Result Verification, Code Verification and Computation of Support Values in Phylogenetics. In: Briefings in Bioinformatics 12, S. 270–279, 2011

Stamatakis, A., Alachiotis, N.: Time and Memory Efficient Likelihood-Based Tree Searches on Gappy Phylogenomic Alignments. In: Bioinformatics 26, S. i132–i139, 2010