

University of Groningen

What fruits can we get from this tree?

Laudanno, Giovanni

DOI:
[10.33612/diss.155031292](https://doi.org/10.33612/diss.155031292)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Laudanno, G. (2021). *What fruits can we get from this tree? A journey in phylogenetic inference through likelihood modeling*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. <https://doi.org/10.33612/diss.155031292>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Chapter 8

Samenvatting

Deze thesis geeft nieuwe methoden om informatie over diversificatie uit een fylogenetische boom te winnen. De gangbare aanpak bij het schatten van de beste parameters voor diversificatie modellen is het gebruik van *likelihood*-functies en *likelihood*-maximalisatie. De parameters binnen deze modellen vertegenwoordigen de snelheden waarmee verschillende evolutionaire gebeurtenissen (e.g. speciaties en extincties) plaatsvinden gedurende het diversificatieproces. Zoals aangegeven in hoofdstuk 1 bestaan er al veel diversificatie modellen. Deze modellen kunnen verschillende vragen beantwoorden en om die reden weerspiegelen de geschatte parameters dan ook verschillende biologische aspecten. Veel van deze methoden maken gebruik van het oplossen van de zogenoemde P-vergelijking zoals weergegeven in 1.3.2, die wordt gebruikt in gangbare *birth-death* modellen (BD), en worden toegepast binnen het Nee et al. framework (waaraan soms wordt gerefereerd als het P-framework in deze thesis) via het gebruik van vergelijking 1.3.3. Later hetzelfde hoofdstuk presenteren we ook het zogenoemde Q-framework (1.3.6.) Dit framework, oorspronkelijk ontwikkeld om met diversiteitsafhankelijke diversificatie om te gaan, stelt ons in staat om het aantal soorten te volgen tijdens het diversificatie-proces. Een gevolg hiervan is dat deze aanpak, wanneer uitgebreid, toegepast kan worden bij een breder scala aan problemen, zoals in Valente, Phillimore, and Etienne (2015).

In hoofdstuk 2 hebben we het Q-framework bestudeerd in alle gevallen waarin het ook mogelijk is om het P-framework toe te passen (n - en ρ -sampling schema's,

constante snelheden en tijdsafhankelijke snelheden). In deze gevallen bewijzen we analytisch dat het Q-framework oplossingen biedt die equivalent zijn aan oplossingen die worden gegeven door in de literatuur beschikbare modellen. Dergelijk bewijs was nodig omdat het artikel waarin het Q-framework oorspronkelijk werd geïntroduceerd enkel een heuristische justificatie gaf, verkregen door numerieke methoden. Dit is verre van optimaal en kan alleen als waar worden beschouwd voor een eindig aantal parameter instellingen. Aangezien het framework in potentie naar meer gevallen kon worden uitgebreid, hadden we een sterkere basis nodig. Nadat we extra analytisch bewijs hadden verzameld, hebben we het in twee andere hoofdstukken van deze thesis gebruikt.

In hoofdstuk 3 hebben we de kern vergelijkingen opgesteld voor het *multiple-births and death* (MBD) model door gebruik te maken van het Q-framework. Dit model staat explosieve spurts van simultane soortvormingsgebeurtenissen toe waarvan de intensiteit afhankelijk is van het huidige aantal soorten. Dit is geen algemeen model, maar een op maat gemaakt model specifiek bedoeld om met ingewikkelde, drukke fylogenieën om te gaan. De focus ligt hier op het bestuderen hoe het effect van (cyclische) veranderingen in de omgeving invloed kan hebben op de fylogenetische geschiedenis van een clade. We hebben laten zien dat we op een betrouwbare manier de parameters met *maximum likelihood* kunnen schatten van een breed scala aan gesimuleerde fylogenieën (met bekende parameters). We hebben ook onderzocht of de *BD likelihood* adequaat is voor het verkrijgen van de eigenschappen van een MBD proces. We vonden dat dit niet triviaal is en ontwikkelden een nieuwe metriek, de DNBT metriek, die in staat is om BD- en MBD- bomen van elkaar te onderscheiden.

In hoofdstuk 4 was het doel niet om een geheel nieuw model te bouwen. In plaats daarvan hebben we gefocust op het verbeteren van sommige van de modellen uit de literatuur die de invloed van *single lineage shifts* op fylogenetische *likelihoods* in acht nemen. Dit is bijvoorbeeld het geval wanneer één enkele *lineage* een cruciale innovatie ontwikkelt. Wanneer dit gebeurt kan zo'n *lineage* ontsnappen aan competitie van andere soorten uit dezelfde clade (zoals in Etienne en Haegeman, 2012), waardoor het sneller kan diversifiëren ten opzichte van de fylogenetische achtergrond. In dit hoofdstuk hebben we eerst cruciale aspecten van bestaande modellen geïdentificeerd, daarna presenteerden we de correcte analytische expressies voor de *likelihood* in de gevallen waarbij zich in een fylogenie: (1) één observeerbare *lineage shift* voordoet bij constante snelheden; (2) één observeerbare snelheidsverandering voordoet met diversiteitsafhankelijke snelheden; (3) één niet-observeerbare *lineage shift* voordoet bij constante snelheden; (4) meerdere observeerbare *lineage shifts* voordoen bij constante snel-

heden; (5) meerdere observeerbare snelheidsveranderingen voordoen met diversiteitsafhankelijke snelheden. We bewijzen ook dat, wanneer één snelheidsverandering aanwezig is, het mogelijk is om, door het combineren van *likelihoods* van gevallen met observeerbare met gevallen met niet-observeerbare snelheidsveranderingen, de originele Nee et al. formule te verkrijgen. Dit demonstreert de consistentie van onze aanpak.

In hoofdstuk 5 hebben we een methode geconstrueerd om te beoordelen of de ontwikkeling van een nieuw *likelihood* model noodzakelijk is en daarmee of op dit moment beschikbare inferentie modellen goed genoeg zijn. Hiervoor hebben wij een *R package* ontwikkeld genaamd *pirouette*. *pirouette* gebruikt als input een verdeling van gesimuleerde fylogenieën die worden gegenereerd onder het te testen diversificatie model, waaraan we hier zullen refereren als het “generatieve” model. Vervolgens genereert *pirouette*, uit elke fylogenie in de verdeling, een posterior-verdeling. Deze posterior-verdeling wordt verkregen door gebruik te maken van de standaard inferentie modellen uit BEAST2. Gebruikers kunnen nu een fout statistiek selecteren om de fout te schatten die BEAST2 maakt tijdens de reconstructie van een fylogenie met de nu beschikbare gereedschappen. Echter, de verdeling van fouten die zo verkregen wordt heeft meerdere bronnen, waarvan de gigantische, natuurlijke stochasticiteit die bij het proces betrokken is, een voorname is. Om hier rekening mee te kunnen houden, doorloopt *pirouette* ook een tweede, parallelle, *pipeline* die in alle bijna aspecten identiek is aan het origineel. Het enige verschil is dat deze *pipeline* wordt doorlopen met als input een fylogenie die gemaakt is onder een standaard diversificatie model. We beschouwen de output van deze parallelle *pipeline* als de baseline error. Wanneer de twee foutverdelingen gelijkend zijn, is de ontwikkeling van een nieuw *likelihood*-model (en diens eventuele implementatie als boom prior in BEAST2) onnodig. Als de foutverdelingen sterk verschillen moet men een nieuwe module ontwikkelen voor de betreffende boom.

8. SAMENVATTING
