

Das APG II Projekt

Entwicklung einer skalierbaren Datenbank zur Erfassung genetischer Verwandtschaft bei Bedecktsamern (Angiospermae)

Von Fabian Golle

Kurzbeschreibung

Das vorliegende IT-Projekt beschäftigt sich mit der Entwicklung einer Datenbank zur Erfassung genetischer Verwandtschaften bei Angiospermen (Bedecktsamer).

Seit Mitte 1998 wurde die Systematik der Bedecktsamer durch die Angiosperm Phylogeny Group (kurz „APG“) neu eingeteilt. Das Hauptziel dieser Gruppe war es, die Pflanzen nicht an morphologischen, also anhand der Ähnlichkeit, sondern nach *phylogenetischen* Merkmalen zu unterscheiden und zu klassifizieren.

Ziel ist es, in einem fortlaufenden Prozess alle Pflanzenarten in den entsprechenden Klassen, Ordnungen und Familien – das sind derzeit mehr als 350.000 Arten – möglichst kompakt und flexibel in eine Datenbank zu packen.

Die Anwendung muss möglichst skalierbar, also anpassbar sein. Denn täglich werden neue Pflanzenarten molekularbiologisch untersucht und beinahe wöchentlich stoßen Forscher auf neue genetische Verwandtschaften unter diesen Pflanzen. Regelmäßig finden damit Veränderungen an der Systematik der Bedecktsamer statt.

Diese Änderungen wurden bisher nie öffentlich kompakt und übersichtlich zusammengefasst oder über eine längere Zeit - wenn nicht sogar auf Dauer - gepflegt.

Bisherige Ansätze, wie z.B. das „Tree of Life“ Projekt scheiterten letztendlich an fehlenden Finanzmitteln oder an der technischen Realisierung.

Unser Projektziel ist es, die Datenbank möglichst skalierbar aufzubauen, und somit umfangreiche Änderungen von unserem Team mit nur wenig Aufwand zu ermöglichen.

Dazu wird die APG II Systematik komfortabel über eine grafische Oberfläche (sog. GUI) eingefügt. Das Programm schreibt diese Eingaben dann in eine riesige relational aufgebaute Tabelle. Aufgrund der Tatsache, dass alle Einträge – egal ob es sich um Klassen, Ordnungen, Familien oder Arten handelt – in einer einzigen Tabelle liegen, lassen sich Einträge technisch gesehen sehr einfach verschieben oder bearbeiten.

Ein weitere Teil des Projektes ist es, die Oberfläche für den Endnutzer möglichst übersichtlich und komfortabel zu halten. So weist z.B. ein automatisch erzeugtes Kladogramm (eine grafische Baumansicht) immer auf den Stammbaum der Pflanzenarten hin. Mit einem „intelligenten“ Spider ausgestattet werden automatisch Literaturangaben, Bilder und Links zum aktuellen Eintrag der Datenbank aus entsprechenden Online-Suchmaschinen gesucht und dem Nutzer geboten.

Auch eine Art API (Programmierschnittstelle) für andere Webseiten oder Forschungsinstitute wäre durchaus vorstellbar – Dem Projekt sind also fast keine Grenzen gesetzt.

Inhaltsverzeichnis

Kurzbeschreibung	1
1 First of all: Was ist APG II?	2
2 Einleitung.....	2
3 Problematik	3
4 Potenziale des Projektes	3
5 Die GUI	4
6 Technische Umsetzung.....	4
7 Literaturverzeichnis.....	6
8 Anlagen.....	7
8.1 Stammbaum der Angiospermen	7

1 First of all: Was ist APG II?

Der Stammbaum der Angiospermen ist der erfolgreichste Pflanzenstamm, der sich auf dem Festland etabliert hat. Es gibt etwa 250.000 Arten. All das ist nichts Neues. Mit der Erschließung und Dokumentierung beschäftigen sich bereits seit Jahrhunderten unzählige Biologen und Phylogenetiker. Allerdings basieren viele der alten Studien noch auf den äußerlichen Merkmalen der Pflanzen, sprich die Pflanzen wurden nach rein optischen Analysemethoden in die Pflanzensystematik eingeordnet. Allerdings ist diese Methode im Zeitalter der DNA-Analyse mehr als veraltet. Inzwischen haben Forscher praktisch aufgrund der neusten Erkenntnisse die alte Systematik jedoch schon einige male abgeändert. Allerdings hat auch hier wieder jede Forschergruppe ihr eigenes Pflanzensystem erstellt und veröffentlicht, sodass es wieder zu keinerlei Einheitlichkeit kam.

Die Angiosperm Phylogeny Group (kurz APG II) hat bisher die am technisch sowie biologisch am besten durchdachte Methode angewandt: Sie untersuchten die Pflanzen jeweils in 81 unterschiedlichen Plastidgenomen (also Pflanzenzellen, die Erbinformationen beinhalten) auf 64 sog. Genomen, also Erbinformationen. Dabei bezogen sie auch 13 bisher unbekannte Genome in ihre Analyse mit ein.

Anhand der Ergebnisse war man in der Lage, den Kompletten Pflanzenstamm der Angiospermen neu einzuteilen. So fand man zum Beispiel heraus, dass die Aborellas die älteste genetische Linie von Pflanzen ist. Auch viele andere bisher offene Fragen zur Evolutionsgeschichte konnten so beantwortet werden.

2 Einleitung

Ein Traum eines jeden Botanikers soll wahr werden: Eine einheitliche Datenbank der Systematik der Blütenpflanzen, erreichbar von überall auf der Welt über das Internet. Das Ganze auch noch aktuell und top-gepflegt. Genau dass ist unsere Ziel! Wir wollen eine global zugängliche Datenbank nach dem APG II Modell der Angiospermen erschaffen. Diese Datenbank soll für Jedermann auf der Welt über einen einfachen Webbrowser zugänglich sein und im Gegensatz zu Projekten wie „The Tree of live“ oder den Detailseiten in Wikipedia immer aktuell sein.

3 Problematik

Forscherteams auf der ganzen Welt stoßen beinahe täglich auf neue genetische Zusammenhänge unter einzelnen Pflanzenarten oder entdecken komplett neue Arten. – Und genau hier liegt das große Problem im Projekt: Wie lässt sich die Datenbank der Pflanzen weiterhin aktuell halten, auch wenn Forscher XY in Neuguinea neue Pflanzenarten entdecken?

Die Lösung ist verhältnismäßig einfach: Es muss für sämtliche Forschergruppen möglich sein, entsprechend in die Datenbank einzugreifen und die betreffenden Einträge abzuändern – natürlich kontrolliert, damit nicht jeder nach seiner eigenen Lust und Laune Einträge abändert. Jede (angemeldete und verifizierte) Forschergruppe kann Einträge verändern oder innerhalb der Systematik verschieben. Diese Änderung kann aber von anderen Gruppen oder Moderatoren, sogenannten „Advisors“ rückgängig gemacht werden, falls Unstimmigkeiten auftreten.

Je nach Erfahrung mit dem Prinzip wäre auch eine Verifikation der Einträge denkbar, das also jede Änderung erst von mindestens einer(zwei, drei) Gruppen bestätigt werden muss, um öffentlich sichtbar zu sein.

4 Potenziale des Projektes

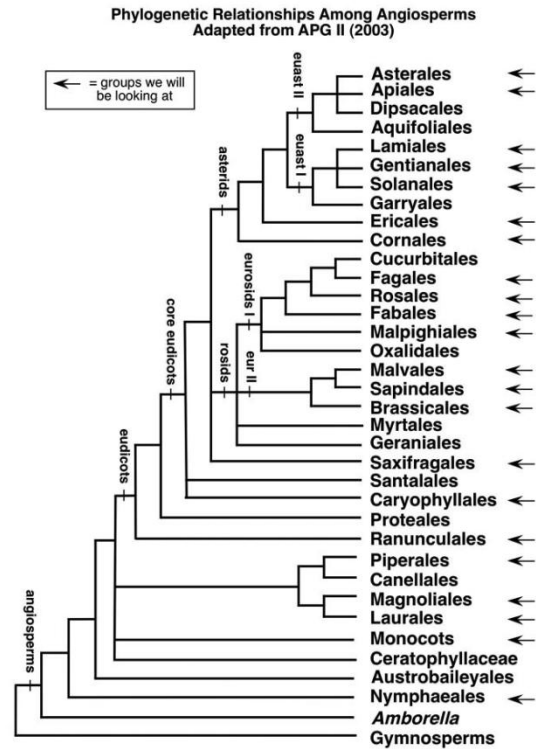
Um es kurz und bündig in einer Aufzählung darzustellen. Aus unserem Projekt ergeben sich folgende Hauptgedanken und Potenziale:

- Die aktuelle Systematik der Bedecktsamer (Angiospermen) wäre für jedermann über das Internet zugänglich
- Jegliche Änderungen oder Änderungsvorschläge an der Systematik könnten von den unterschiedlichen Forschungsgruppen online diskutiert werden und dann nach entsprechender Verifikation in die Datenbank übernommen werden.
- Auch ein Revisionsystem á la Wikipedia wäre denkbar, sodass Änderungen in der Datenbank auch noch nachträglich mit verfolgt und ggf. rückgängig gemacht werden können.
- Die übersichtlich gestaltete Oberfläche bietet viele komfortable Unterstützungsmöglichkeiten, wie z.B. eine automatische Google-Recherche, für die Nutzer. Auch eine Kooperation mit Springerlink wäre denkbar.
- Andersrum könnte man selbst aber auch eine Schnittstelle anbieten, über die andere Forschungsinstitute ihre Applikationen über unsere Datenbank aktuell halten können.
- Aufgrund der Schul- bzw. Schülerbasierten Entwicklung und Pflegeung des Systems ist die Finanzsituation weniger kritisch. Die entsprechende Hardware könnte z.B. durch Sponsoren oder Werbung auf der Plattform finanziert werden.

5 Die GUI

Die grafische Oberfläche, nachfolgend auch „GUI“ (engl. Graphical User Interface), ist schlicht und einfach gehalten. Die Navigation innerhalb der Systematik erfolgt entweder Textbasiert oder über ein sog. Kladogramm.

Wichtig ist, dass es sich bei diesem Projekt nicht um ein Projekt mit vorher festgelegten Zeitraum handelt. Vielmehr handelt es sich um einen fortlaufenden Entwicklungsprozess. Genauso werden auch die Einträge gehandhabt. Das Ziel oder vielmehr die Hoffnung ist es, zu jeder Pflanzenart – und das sind immerhin über 350.000 - nicht nur den Namen und die Einteilung im System, sondern auch eine kurze Beschreibung, Vorkommen der Pflanzenart und auch



ein kleines Bild sowie Literatur zu haben.

Abbildung 1: Beispiel Kladogramm

Teilweise könnten Suchroboter (sog. Spider) uns diese Arbeit abnehmen, indem sie sich automatisch z.B. über Suchmaschinen wie Google auf die Suche begeben, die Ergebnisse entsprechend filtern und mit in die Datenbank schreiben. Allerdings muss dann theoretisch jeder Eintrag von einem Menschen auf explizites Material, sowie auf bestehende Urheberrechte sowie explizite Inhalte überprüft werden, was diesen Vorgang nicht gerade einfacher macht.

Möglich wäre es auch, dass bestimmte Komponenten der GUI direkt beim Aufrufen der Website über eine entsprechende AJAX-API¹ aus dem Internet abzurufen.

6 Technische Umsetzung

Das Grundprinzip der Datenbank ist relativ einfach. Es handelt sich um eine einzige, relationale Tabelle für alle Einträge. Die Tabelle basiert momentan noch auf MySQL und hat 4 wesentliche Felder:

Feldname	Beschreibung
ID	Eine fortlaufende ID (Primärschlüssel)
Parent	Die ID des jeweils übergeordneten Eintrages (Lässt die wichtigen Verschachtelungen innerhalb der Datenbank zu)
Type	Den Typ des Eintrags (z.B. Klasse, Ordnung, Familie, Art, Node usw.)
Text	Der Titel (Name) des Eintrages. Hier stehen also die Namen der Klassen, Gattungen, Familien usw.

¹ Eine Technologie, die es ermöglicht Daten ohne das Neuladen bzw. Verlassen der Website aus dem Internet abzurufen

Weitere Spalten in der Datenbank beinhalten beispielsweise eine Kurzbeschreibung, Vorkommen, ein Bild usw. der jeweiligen Art.

Die Funktionsweise der Datenbank lässt sich am besten an folgenden Beispielen erklären:

- **ROOT ID: 0**
 - **Element 1 ID: 1 Parent: 0**
 - **Sub-Element 1.1 ID: 5 Parent: 1**
 - **Sub-Element 1.2 ID: 6 Parent: 1**
 - **Sub-Element 1.3 ID: 7 Parent: 1**
 - **Sub-Element 1.3.1 ID: 11 Parent: 7**
 - **Element 2 ID: 2 Parent: 0**
 - **Sub-Element 2.1 ID: 8 Parent: 2**
 - **Sub-Element 2.2 ID: 9 Parent: 2**
 - **Sub-Element 2.2.1 ID: 12 Parent: 9**
 - **Sub-Element 2.2.1 ID: 13 Parent: 9**
 - **Element 3 ID: 3 Parent: 0**
 - **Sub-Element 3.1 ID: 10 Parent: 3**
 - **Element 4 ID: 4 Parent: 0**

ID	PARENT	TEXT
1	0	Element 1
2	0	Element 2
3	0	Element 3
4	0	Element 4
5	1	Sub-Element 1.1
6	1	Sub-Element 1.2
7	1	Sub-Element 1.3
8	2	Sub-Element 2.1
9	2	Sub-Element 2.2
10	3	Sub-Element 3.1
11	7	Sub-Element 1.3.1
12	9	Sub-Element 2.2.1
13	9	Sub-Element 2.2.1

Die Tabelle ist also intern relational aufgebaut. Jedes Element in der Datenbank, und somit im Tree hat eine eindeutige ID (Spalte 1). Das erste Element („ROOT“) hat die ID 0. Es braucht nicht in die Datenbank geschrieben werden, da es unveränderlich ist.

Bisher haben wir also eine ganz simple Tabelle, in die wir Einträge in der gleichen Ebene, nämlich direkt unter dem „Root“-Eintrag einordnen können.

Die Verschachtelung der Pflanzensystematik erfordert aber eben eine gewisse Verschachtelung der Einträge in der Datenbank. Und genau dafür sind die Einträge in der Spalte „PARENT“ zuständig. Sie definieren für jedes Element ein eindeutig übergeordnetes Element.

Beispiel: Ist der Eintrag „Sub-Element 1.3.1“ (mit der ID 11) dem Eintrag „Sub-Element 1.3“ (mit der ID 7) untergeordnet, wird dieses ganz einfach durch Ändern des Feldes „PARENT“ auf die ID des übergeordneten Eintrags, in diesem Fall also die **11**, bewirkt.

Auf diese Art lassen sich also relativ einfach komplexe verschachtelte Bäume in der Datenbank abspeichern. Sollte sich nun, wie im oben beschriebenen Szenario, die Position einer Art innerhalb der Systematik ändern, muss das Programm einfach nur die „PARENT“-ID des jeweiligen Eintrages.

7 Literaturverzeichnis

PNAS. (n.d.). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns.

