Full of beans: a study on the alignment of two flowering plants classification systems

Yi-Yun Cheng and Bertram Ludaescher

{yiyunyc2,ludaesch}@illinois.edu

https://github.com/yiyunyc2/NKOS18

NKOS2018, Sept. 13, Porto, Portugal

ILLINOIS | School of Information Sciences

The iSchool at Illinois

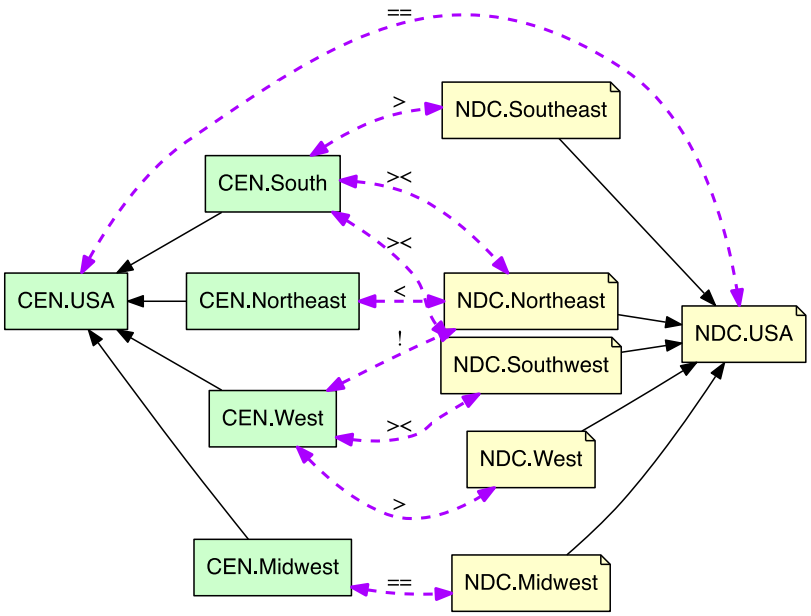
Reasons to align different classifications

- Difficulty to organize information using a stable unitary classification scheme over time
- In biodiversity domain, It is common for taxonomists to contradict each other's or even their own previous taxonomies
- KOS are dynamic, time-specific, and responsive to both empirical signals and human classification interests

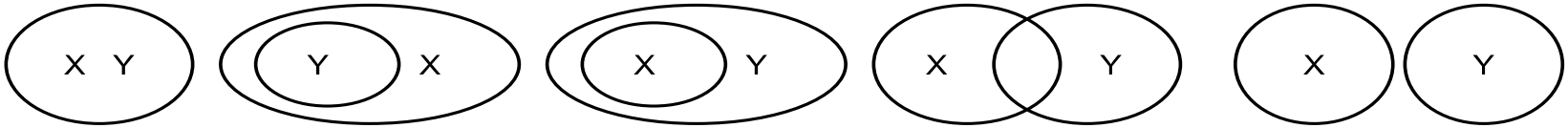
Purpose of this study

- Demonstrate the feasibility of integrating two classifications that result in numerous possible solutions
  - the computational power that can aid us in aligning KOS which could not have been possible when working with alignments manually
- With the hope that this work will further shed lights on:
  - the possible alignments of the classifications in the information science community
  - bring a novel approach for aligning KOS in the future

5 ways to relate concepts (regions)



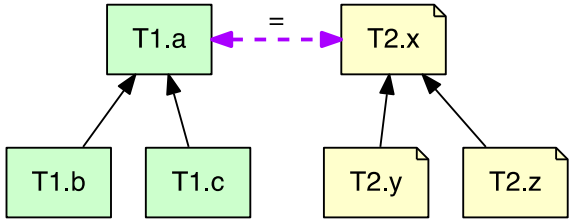
- **Idea**: relate concepts X and Y with ***articulations***
- Articulation Language: ***Region Connection Calculus*** (**RCC5**): *congruence*, *inclusion*, *inverse inclusion*, *overlap*, *disjointness*



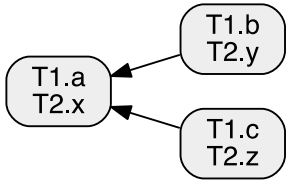
Congruence X == Y | Inclusion X > Y | Inverse Inclusion X < Y | Overlap X>< Y | Disjointness X ! Y

School of Information Sciences
The iSchool at Illinois

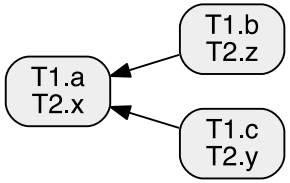
Reasoning about taxonomies



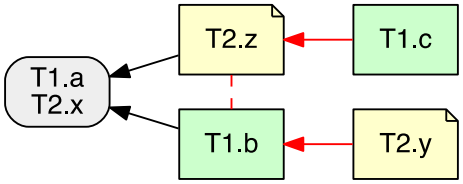
Input Taxonomies



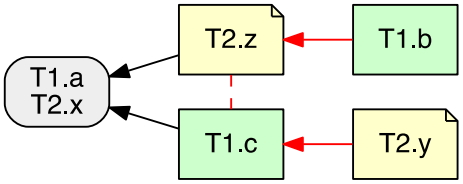
PW1



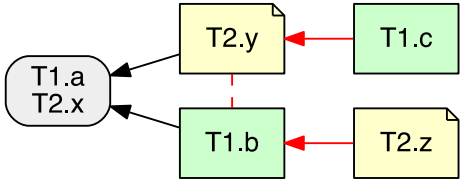
PW2



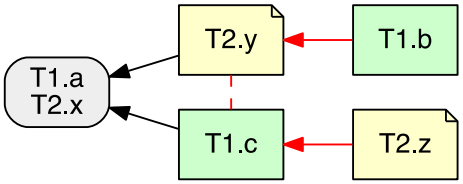
PW3



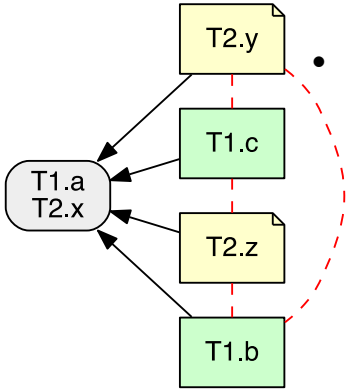
PW5



PW4



PW6



PW7

- **Given**:
  - taxonomies T1, T2
  - and relations T1 ~ T2 (*articulations*, *alignment*)
- **Find**:
  - merged taxonomy T3
- **Such that**:
  - T1, T2 are **preserved**
  - all pairwise relations are **explicit**

Two Flowering Plant Classifications

- Cronquist system
  - Arthur Cronquist (1981)
  - classifying plant resemblances based on evolution and *morphological* similarity (similar characteristics)
  - the most fully developed phyletic system of flowering plant classification systems by far
- Angiosperm Phylogeny Group System (APG IV)
  - uses both morphological and *molecular data* to group plants
  - The de facto standard classifications now

Biodiversity Taxonomies

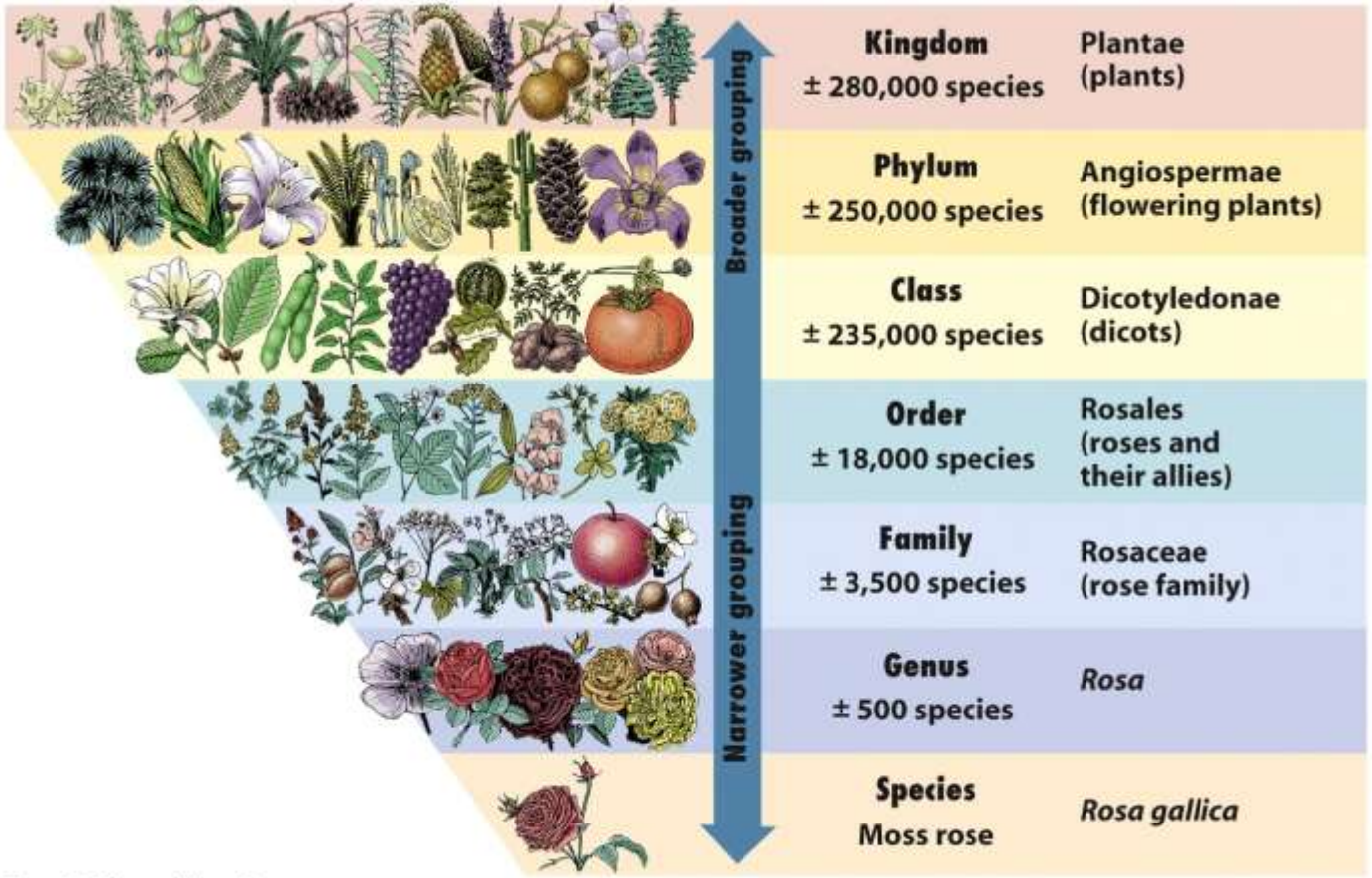


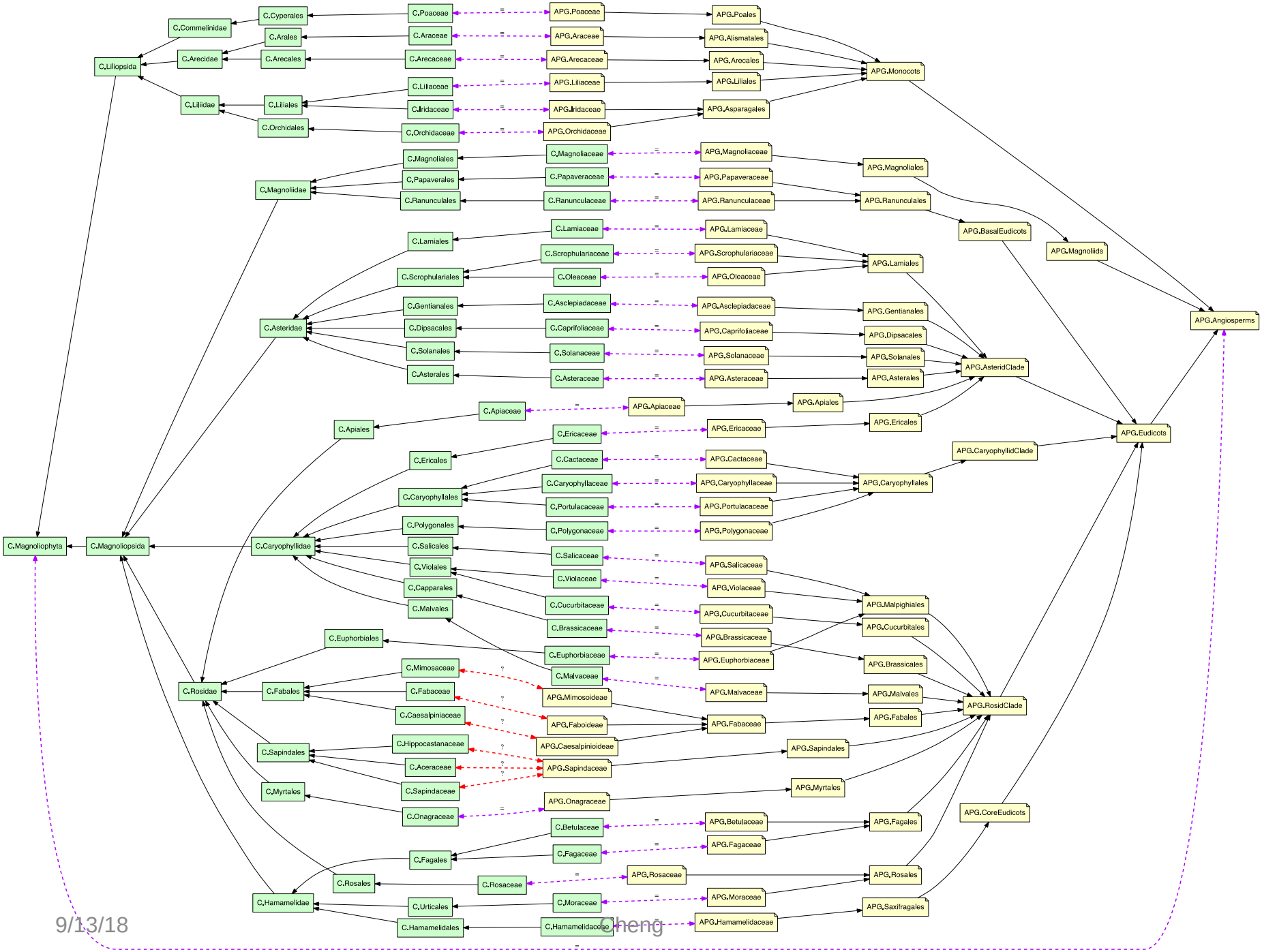
Figure 2-6 Discover Biology 3/e

What are we aligning in our use case

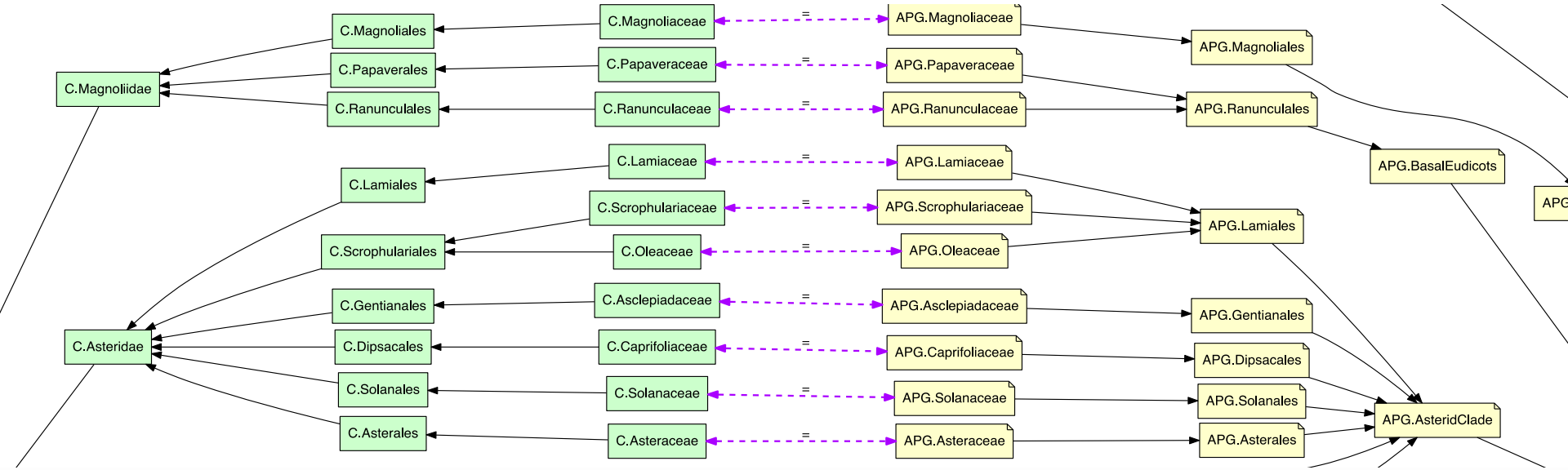
- *Family* level concepts
- Total number of flowering plant families: 416
- Number of families we align
  - 40 most common families/ subfamilies
- Examples: Magnoliaceae, Ranunculaceae, Papaveraceae, Cataceae, Betulaceae, Fabaceae, Rosaceae

First round of alignment

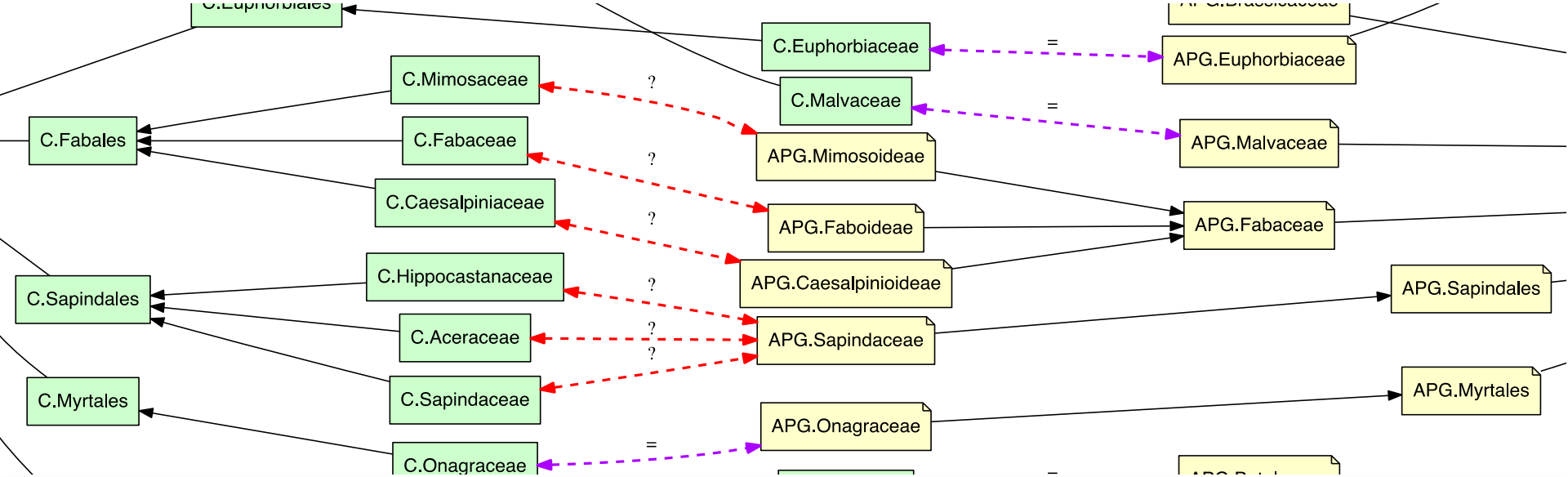
- Goal:
  - Align the 40 families (or the variations of the names) in both classifications
- Rules:
  - if the family in both systems shares the ***exact same name***, we assume (possible incorrectly) that they are ***congruent*** to each other
  - If there are ***similar but different*** names, we will leave the concepts unmapped at first


C.Cyperales
C.Poaceae
APG.Poaceae
APG.Poales
C.Commelinidae
C.Arales
C.Araceae
APG.Araceae
APG.Alismatales
C.Arecidae
C.Arecales
C.Arecaceae
APG.Arecaceae
APG.Arecales
C.Liliopsida
APG.Monocots
C.Liliaceae
APG.Liliaceae
APG.Liliales
C.Liliidae
C.Liliales
C.Iridaceae
APG.Iridaceae
APG.Asparagales
C.Orchidales
C.Orchidaceae
APG.Orchidaceae
C.Magnoliales
C.Magnoliaceae
APG.Magnoliaceae
APG.Magnoliales
C.Papaverales
C.Papaveraceae
APG.Papaveraceae
C.Magnoliidae
C.Ranunculales
C.Ranunculaceae
APG.Ranunculaceae
APG.Ranunculales
APG.BasalEudicots
C.Lamiales
C.Lamiaceae
APG.Lamiaceae
APG.Magnoliids
C.Scrophulariaceae
APG.Scrophulariaceae
APG.Lamiales
C.Scrophulariales
C.Oleaceae
APG.Oleaceae
C.Gentianales
C.Asclepiadaceae
APG.Asclepiadaceae
APG.Gentianales
APG.Angiosperms
C.Asteridae
C.Dipsacales
C.Caprifoliaceae
APG.Caprifoliaceae
APG.Dipsacales
C.Solanales
C.Solanaceae
APG.Solanaceae
APG.Solanales
APG.AsteridClade
C.Asterales
C.Asteraceae
APG.Asteraceae
APG.Asterales
C.Apiaceae
APG.Apiaceae
APG.Apiales
C.Apiales
C.Ericaceae
APG.Ericaceae
APG.Ericales
APG.Eudicots
C.Ericales
C.Cactaceae
APG.Cactaceae
APG.CaryophyllidClade
C.Caryophyllaceae
APG.Caryophyllaceae
APG.Caryophyllales
C.Caryophyllales
C.Portulacaceae
APG.Portulacaceae
C.Polygonales
C.Polygonaceae
APG.Polygonaceae
C.Magnoliophyta
C.Magnoliopsida
C.Caryophyllidae
C.Salicales
C.Salicaceae
APG.Salicaceae
C.Violales
C.Violaceae
APG.Violaceae
C.Capparales
C.Cucurbitaceae
APG.Cucurbitaceae
APG.Malpighiales
C.Malvales
C.Brassicaceae
APG.Brassicaceae
APG.Cucurbitales
C.Euphorbiales
C.Euphorbiaceae
APG.Euphorbiaceae
APG.Brassicales
C.Malvaceae
APG.Malvaceae
APG.Malvales
C.Rosidae
C.Fabales
C.Mimosaceae
APG.Mimosoideae
C.Fabaceae
APG.Faboideae
APG.Fabaceae
APG.Fabales
C.Caesalpiniaceae
APG.Caesalpinioideae
APG.RosidClade
C.Sapindales
C.Hippocastanaceae
APG.Sapindaceae
APG.Sapindales
C.Aceraceae
C.Sapindaceae
C.Myrtales
APG.Myrtales
C.Onagraceae
APG.Onagraceae
C.Betulaceae
APG.Betulaceae
APG.Fagales
APG.CoreEudicots
C.Fagales
C.Fagaceae
APG.Fagaceae
C.Rosales
C.Rosaceae
APG.Rosaceae
APG.Rosales
C.Hamamelidae
C.Urticales
C.Moraceae
APG.Moraceae
APG.Saxifragales
C.Hamamelidales
C.Hamamelidaceae
APG.Hamamelidaceae

Zooming in to the input alignments

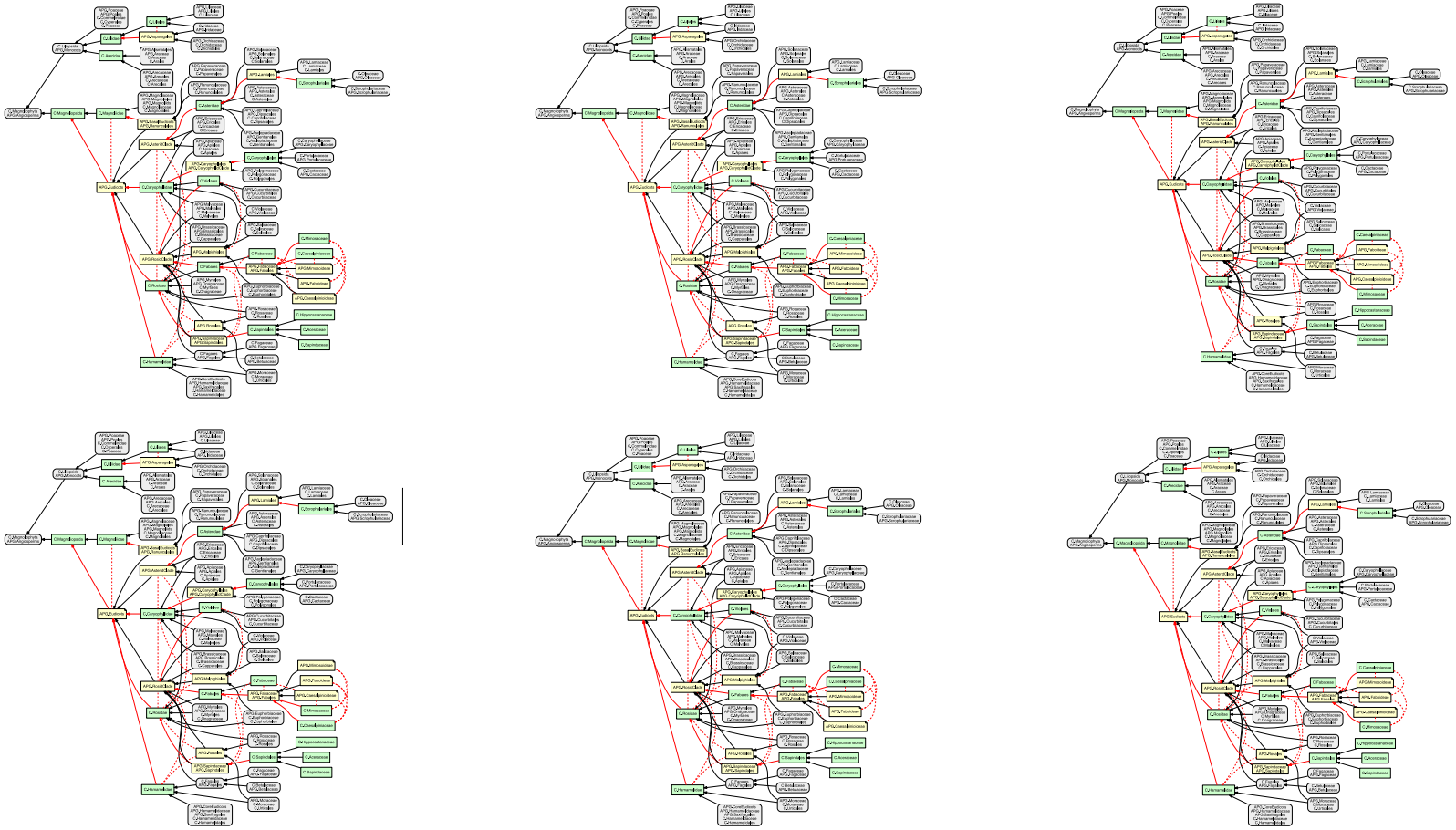


Six unknown relations

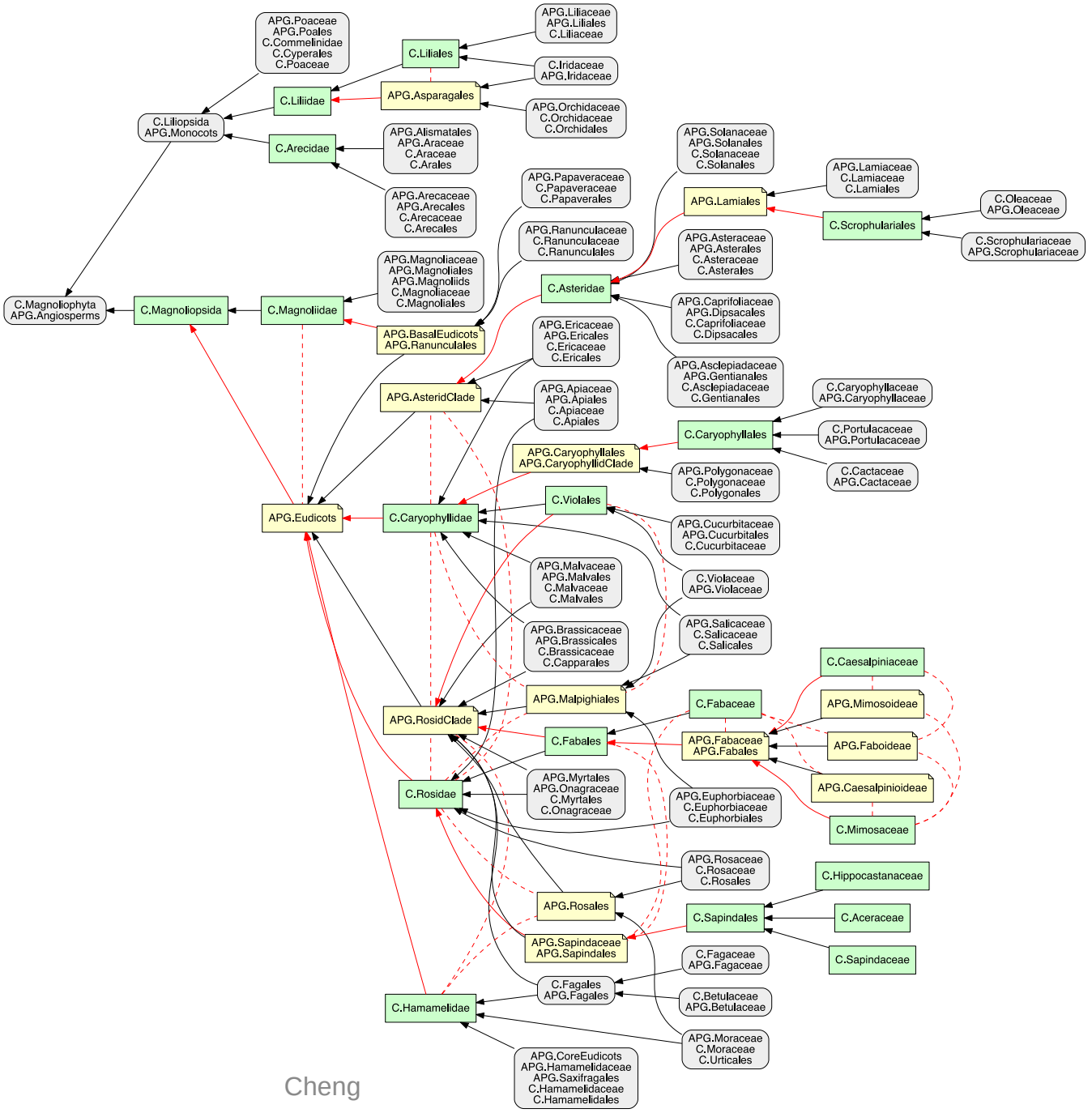


[C.Caesalpini**aceae** ? APG.Caesalpini**oideae**]
[C.Mimosa**ceae** ? APG.Mimosoi**deae**]
[C.Faba**ceae** ? APG.Faboi**deae**]
[C.Aceraceae ? APG.**Sapindaceae**]
[C.Sapindaceae ? APG.**Sapindaceae**]
[C.Hippocastanaceae ? APG.**Sapindaceae**]

Results: 555 Possible Worlds



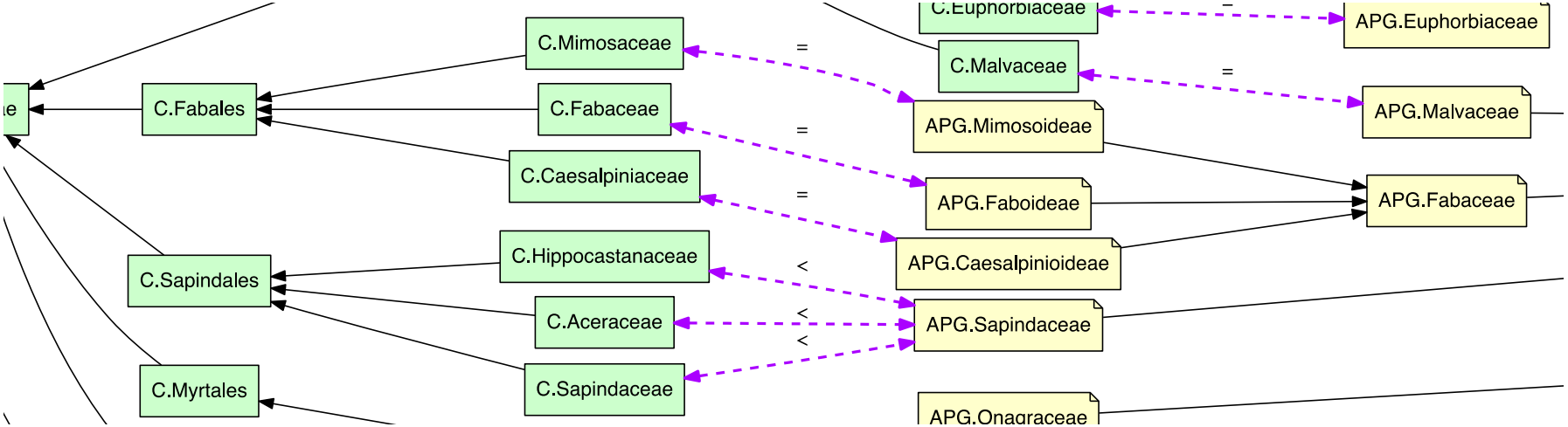
Zooming in..


APG.Poaceae
APG.Poales
C.Commelinidae
C.Cyperales
C.Poaceae
APG.Liliaceae
APG.Liliales
C.Liliaceae
C.Liliales
C.Iridaceae
APG.Iridaceae
C.Liliidae
APG.Asparagales
APG.Orchidaceae
C.Orchidaceae
C.Orchidales
C.Liliopsida
APG.Monocots
C.Arecidae
APG.Alismatales
APG.Araceae
C.Araceae
C.Arales
APG.Arecaceae
APG.Arecales
C.Arecaceae
C.Arecales
APG.Papaveraceae
C.Papaveraceae
C.Papaverales
APG.Ranunculaceae
C.Ranunculaceae
C.Ranunculales
APG.Magnoliaceae
APG.Magnoliales
APG.Magnoliids
C.Magnoliaceae
C.Magnoliales
C.Magnoliophyta
APG.Angiosperms
C.Magnoliopsida
C.Magnoliidae
APG.BasalEudicots
APG.Ranunculales
APG.Solanaceae
APG.Solanales
C.Solanaceae
C.Solanales
APG.Lamiaceae
C.Lamiaceae
C.Lamiales
APG.Lamiales
C.Oleaceae
APG.Oleaceae
C.Scrophulariales
C.Scrophulariaceae
APG.Scrophulariaceae
APG.Asteraceae
APG.Asterales
C.Asteraceae
C.Asterales
C.Asteridae
APG.Caprifoliaceae
APG.Dipsacales
C.Caprifoliaceae
C.Dipsacales
APG.Ericaceae
APG.Ericales
C.Ericaceae
C.Ericales
APG.Asclepiadaceae
APG.Gentianales
C.Asclepiadaceae
C.Gentianales
APG.AsteridClade
APG.Apiaceae
APG.Apiales
C.Apiaceae
C.Apiales
C.Caryophyllaceae
APG.Caryophyllaceae
C.Caryophyllales
C.Portulacaceae
APG.Portulacaceae
APG.Caryophyllales
APG.CaryophyllidClade
C.Cactaceae
APG.Cactaceae
APG.Polygonaceae
C.Polygonaceae
C.Polygonales
C.Violales
APG.Eudicots
C.Caryophyllidae
APG.Cucurbitaceae
APG.Cucurbitales
C.Cucurbitaceae
APG.Malvaceae
APG.Malvales
C.Malvaceae
C.Malvales
C.Violaceae
APG.Violaceae
APG.Salicaceae
C.Salicaceae
C.Salicales
APG.Brassicaceae
APG.Brassicales
C.Brassicaceae
C.Capparales
C.Caesalpiniaceae
APG.Malpighiales
C.Fabaceae
APG.Mimosoideae
APG.RosidClade
APG.Fabaceae
APG.Fabales
APG.Faboideae
C.Fabales
APG.Caesalpinioideae
APG.Myrtales
APG.Onagraceae
C.Myrtales
C.Onagraceae
C.Rosidae
APG.Euphorbiaceae
C.Euphorbiaceae
C.Euphorbiales
C.Mimosaceae
APG.Rosaceae
C.Rosaceae
C.Rosales
C.Hippocastanaceae
APG.Rosales
C.Sapindales
C.Aceraceae
APG.Sapindaceae
APG.Sapindales
C.Sapindaceae
C.Fagaceae
APG.Fagaceae
C.Fagales
APG.Fagales
C.Betulaceae
APG.Betulaceae
C.Hamamelidae
APG.Moraceae
C.Moraceae
C.Urticales
APG.CoreEudicots
APG.Hamamelidaceae
APG.Saxifragales
C.Hamamelidaceae
C.Hamamelidales

Second round of alignment

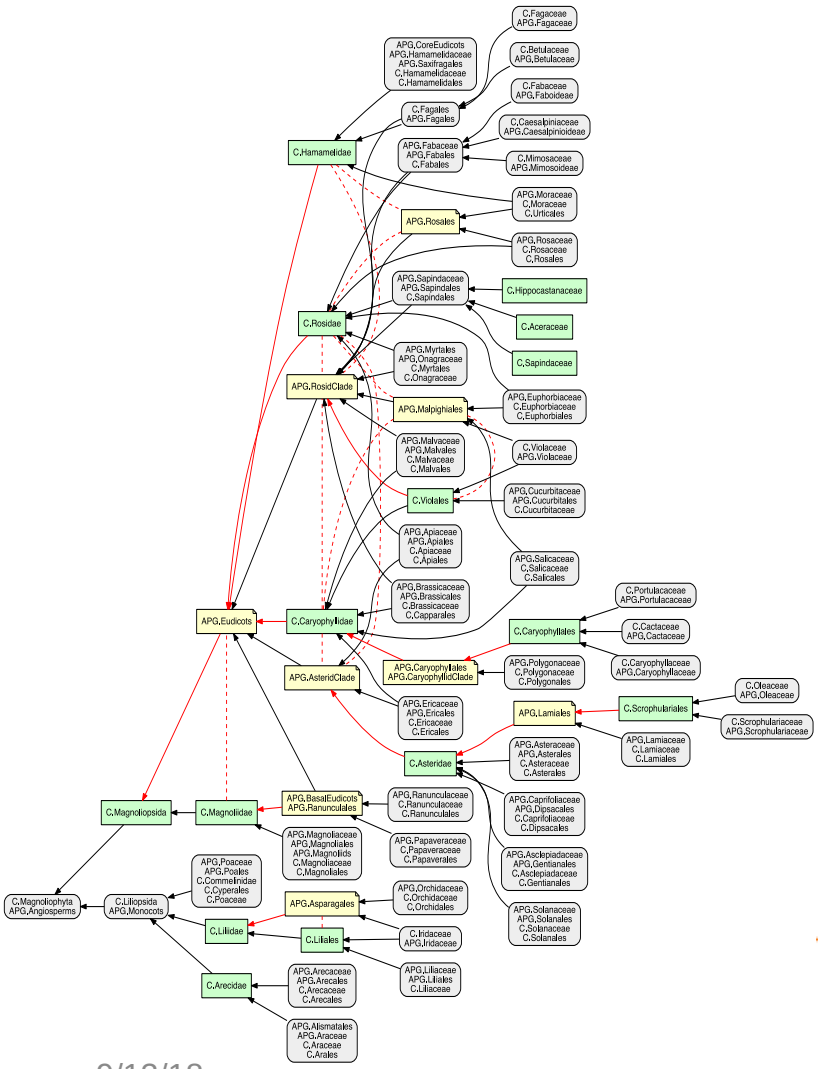
- Expert's input
  - To verify our 'congruent' alignments as well as to sort out the underspecified articulations
- Minimum viable product
  - The results from the first stage of alignment help us communicate with the expert and let him/her grasp all possible solutions for the alignment problem

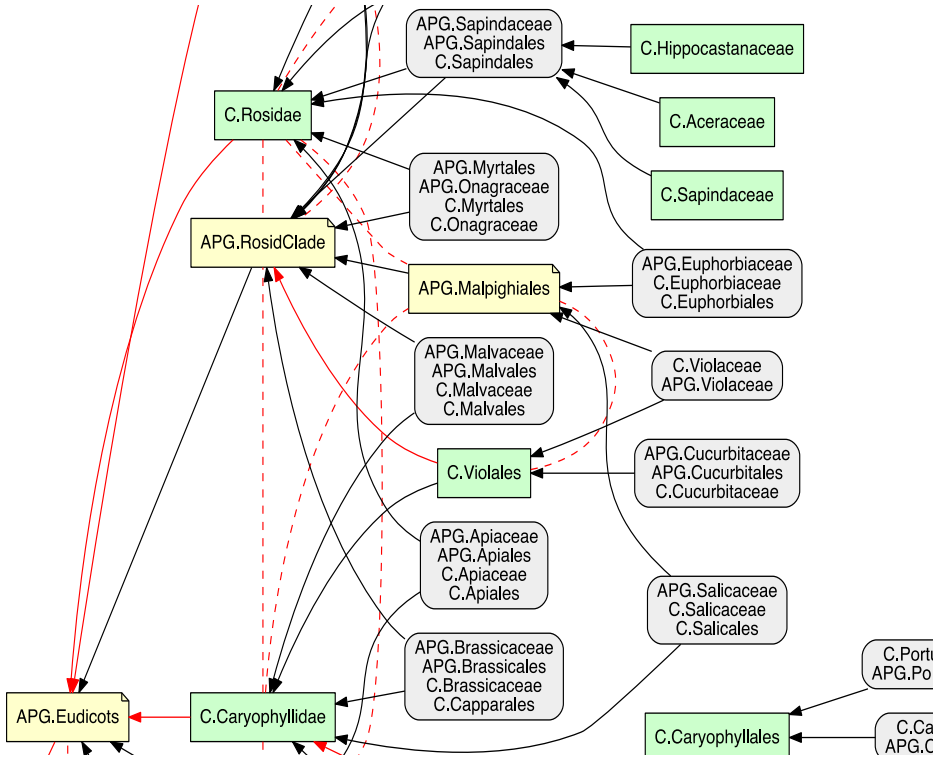
Refining the six unknown relations



[C.Caesalpiniaceae {=} APG.Caesalpinioideae]
[C.Mimosaceae {=} APG.Mimosoideae]
[C.Fabaceae {=} APG.Faboideae]
[C.Aceraceae {<} APG.Sapindaceae]
[C.Sapindaceae {<} APG.Sapindaceae]
[C.Hippocastanaceae {<} APG.Sapindaceae]

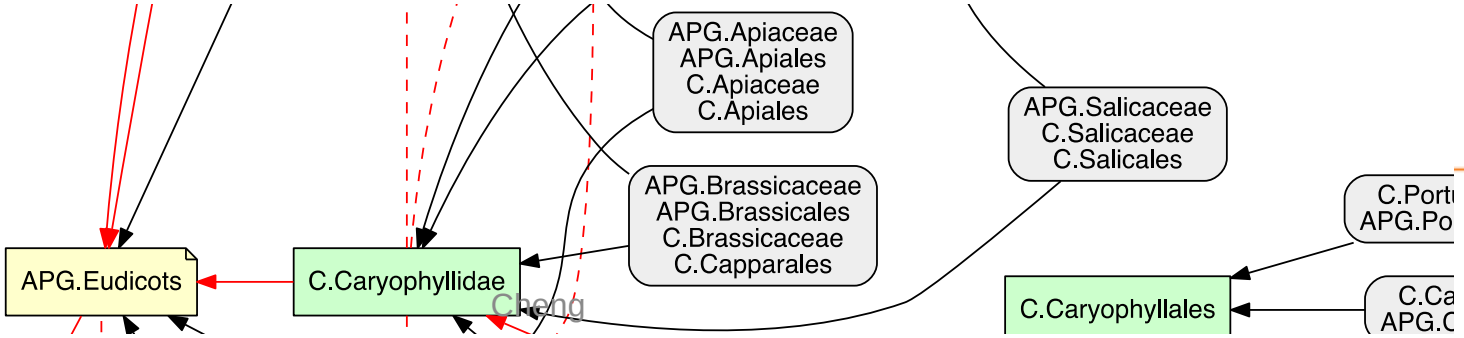
Final result: One Possible World





Limitations

- When a parent node only has one child, the RCC reasoner in Euler/X will collapse the concepts and merge them as the same node
- Our response: we have only chosen some 40 major flower families instead of all 416 families. We could add missing children or artificial children here



Conclusion and discussion

- Why using the logic-based RCC-5, Euler/X approach?
  - classifications can coexist & disambiguating the names among concepts in a merged possible world
  - solving complex alignments for cases where manual efforts would likely fail to yield all 555 different ways to merge and reconcile the two KOS
- Domain experts seems very important..?
  - KOS alignment problems are usually complex-- our logic-based alignment can serve as ***minimum viable product***

Conclusion and discussion

- Why bother aligning the ‘older’ classifications such as the Cronquist system (1981)?
  - the Cronquist system still maintains its esteemed role for its comprehensiveness and preciseness in morphologically classifying the flowering plants
- Some other implications
  - making our classification systems more “full of beans
  - to enable semantic interoperability, and enrich diversity in classification systems

Thank you!

Github Repo:
https://github.com/yiyunyc2/NKOS18

Yi-Yun (Jessica) Cheng
yiyunyc2@illinois.edu

ILLINOIS | School of Information Sciences
The iSchool at Illinois