Порядок и беспорядок в мире геномов хлоропластов: как структура генома связана с таксономией носителя

[1]Садовский Михаил Георгиевич

[1]Институт вычислительного моделирования СО РАН

3 июня 2015

Краткое содержание

Что такое «функция»?

Следует признать, что понятие функции (нуклеотидной последовательности) совсем не просто: можно говорить о химических функциях (ДНК может быть катализатором), о генах, о некодирующих областях, об эпигенах и т. д.

Впрочем, временно мы будем полагать, что каждый имеет хотя бы интуитивное представление о функциях НП.

Что такое таксономия?

Несмотря на достаточно частые изменения в таксономии организмов, с ней вопросов существенно меньше. Таксономическое положение («больших животных») определяется морфологически. То есть, по соматическому геному в конце концов.

Наша задача: связь структуры и таксономии

Итак, мы проделаем следующий путь:

- Сперва определим, что такое структура;
- Затем постараемся понять, насколько разные геномы оказываются близкими по структуре и формируют ли они кластеры;
- Выделим такие кластеры (если получится!);
- Проверим, какие именно геномы попали в один кластер: случайно распределение геномов (с точки зрения таксономии) по кластерам или нет?

Частотный словарь

Нуклеотидная последовательность = символьная последовательность из $\aleph = \{A, C, G, T\}$. Число символов в ней N — её длина.

Слово — любая связная подпоследовательность $\omega = \nu_1\nu_2\nu_3, \ldots, \nu_{q-1}\nu_q$ длины q символов; в частности, при $q = 3$ будем иметь триплеты (знакомые, часто aka кодоны).

Конечный словарь — список всех слов (длины q) с указанием числа копий каждого слова в последовательности.

Заменим число копий n_ω на частоту

$$f_\omega = \frac{n_\omega}{N}$$

и получим частотный словарь (толщины q).

Всюду впредь будем работать с частотными словарями толщины $q = 3$.

Частотный словарь триплетов

Каждый частотный словарь триплетов — точка в 63-мерном пространстве.

Два словаря совпадают, если частоты соответствующих триплетов равны.

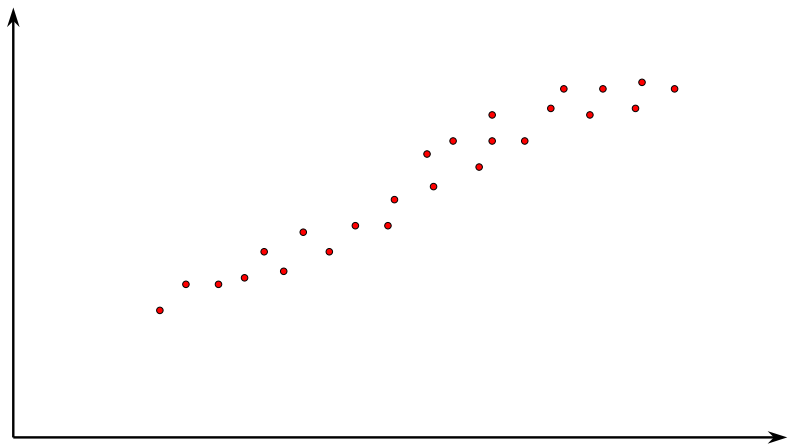
Между двумя словарями можно легко задать расстояние. Мы будем пользоваться старым добрым расстоянием Евклида:

$$\rho\left\{W_3^{(1)}, W_3^{(2)}\right\} = \sqrt{\sum_{i=\mathrm{AAA}}^{\mathrm{TTT}} \left(f_i^{(1)} - f_i^{(2)}\right)^2}.$$

Многомерные данные: как их увидеть?

С распределениями различных (биологически осмысленных) величин знакомы все. Часто они бывают многомерными (как, например, 63-мерный словарь триплетов).

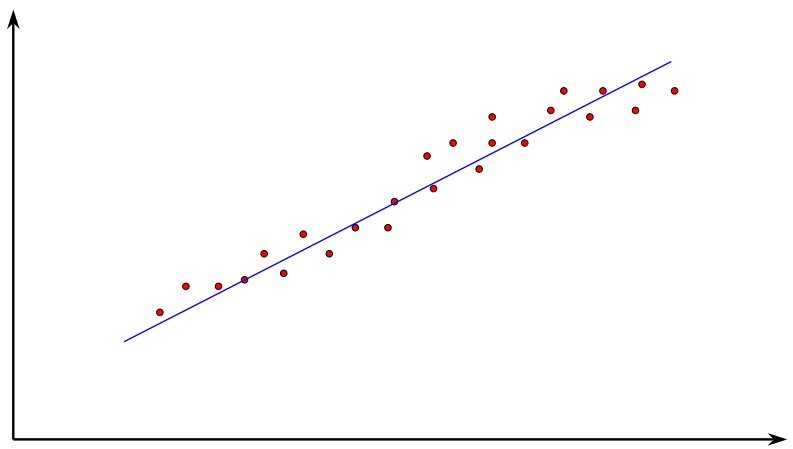
Приблизить многомерные данные многообразиями малой размерности.



Многомерные данные: как их увидеть?

С распределениями различных (биологически осмысленных) величин знакомы все. Часто они бывают многомерными (как, например, 63-мерный словарь триплетов).

Приблизить многомерные данные многообразиями малой размерности.



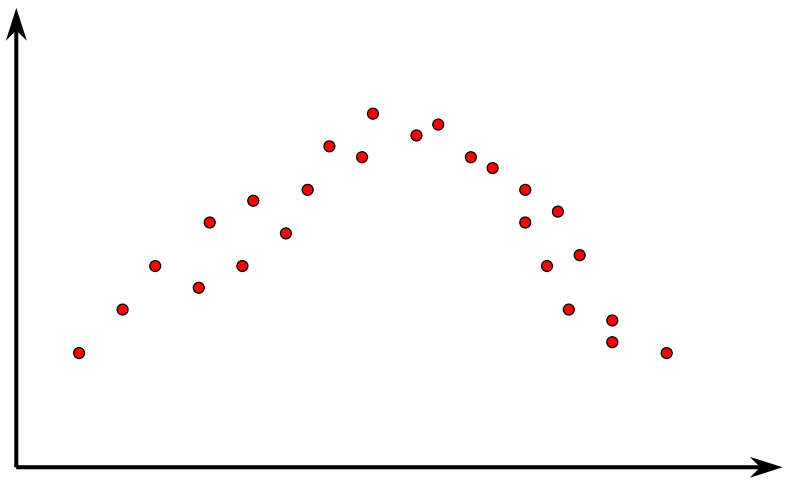
Какие бывают многообразия малой размерности?

Многообразие нулевой размерности: среднее значение.

Многообразие размерности единица: дисперсия (стандартное отклонение).

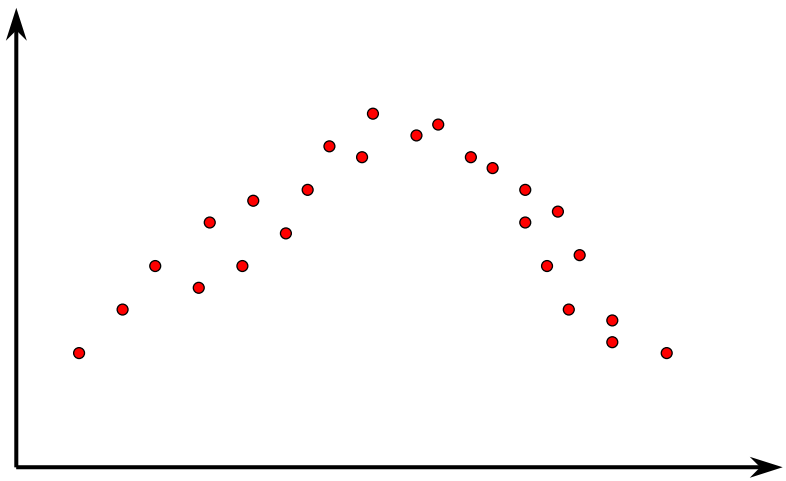
Какие бывают многообразия малой размерности?

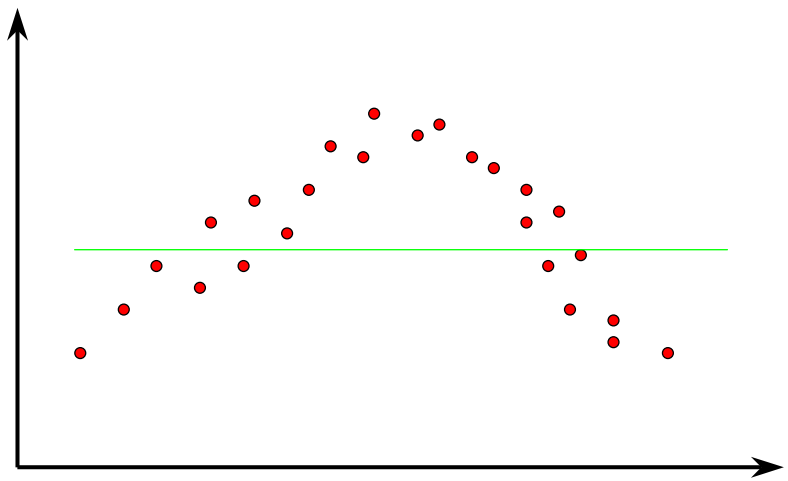
Но что делать в таком случае?



Какие бывают многообразия малой размерности?

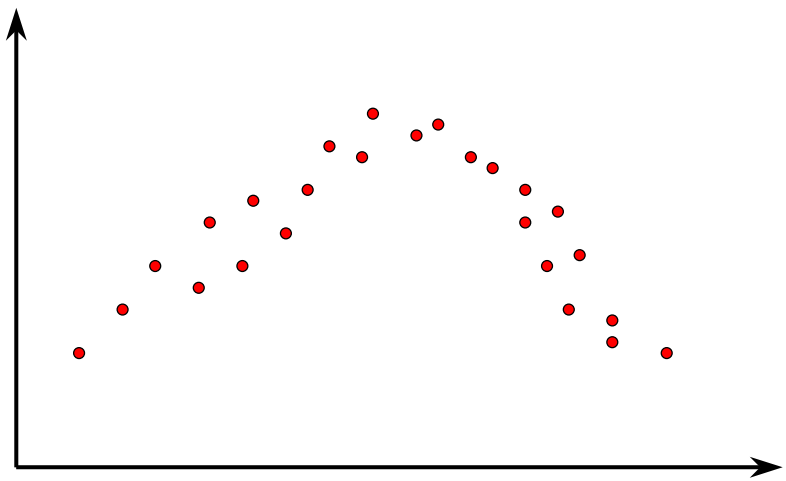
Но что делать в таком случае?

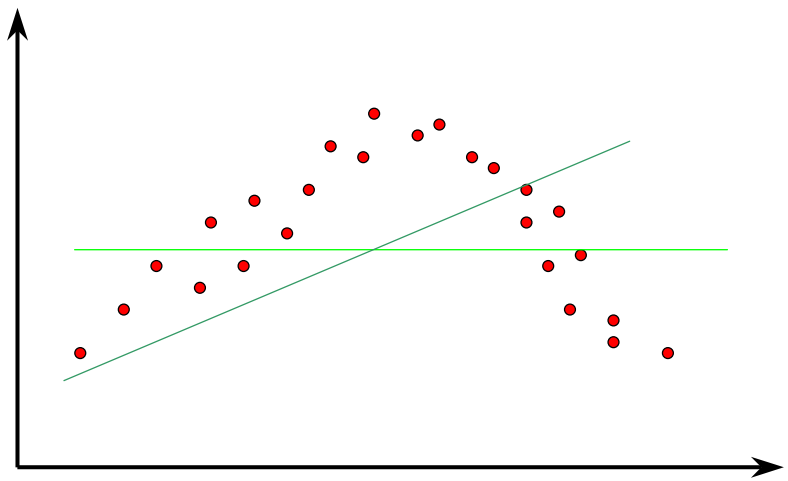




Какие бывают многообразия малой размерности?

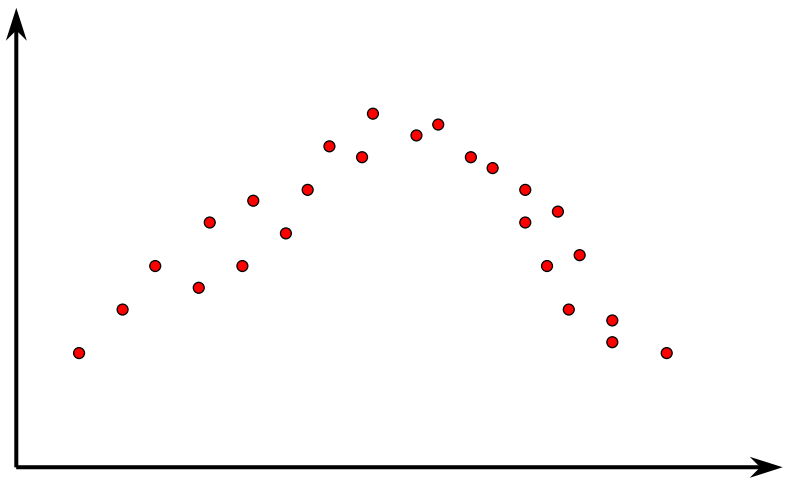
Но что делать в таком случае?

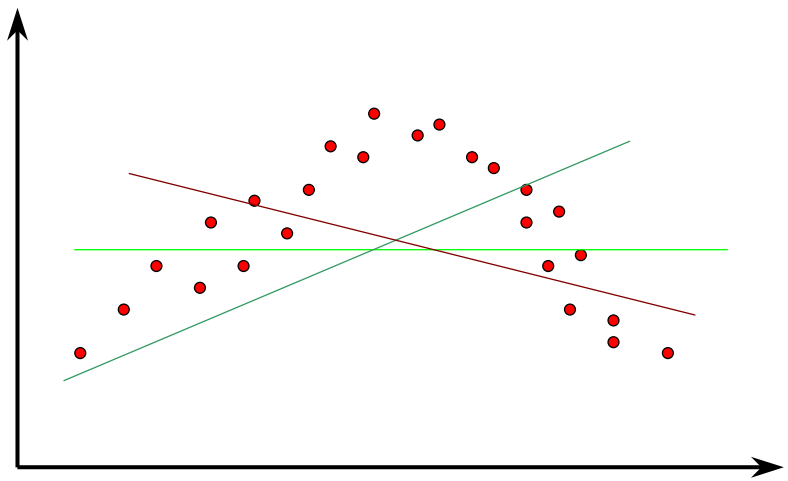




Какие бывают многообразия малой размерности?

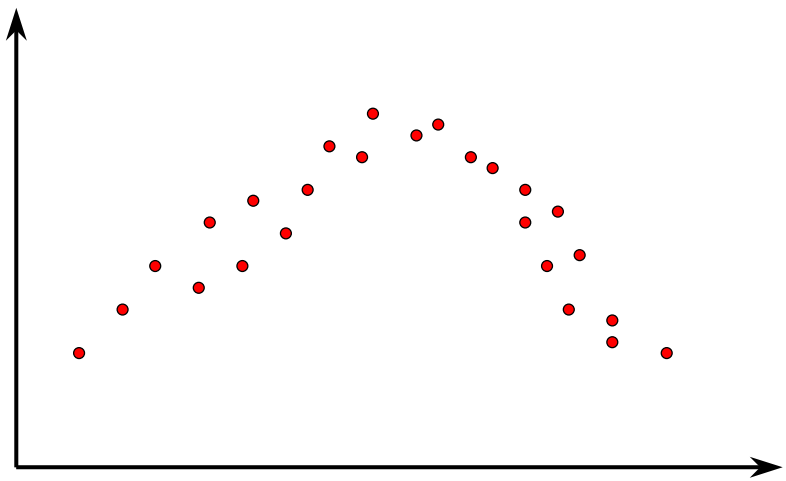
Но что делать в таком случае?

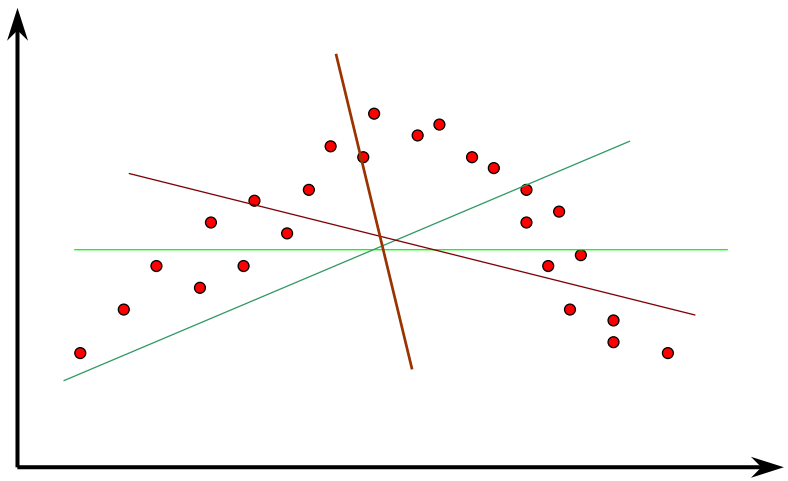




Какие бывают многообразия малой размерности?

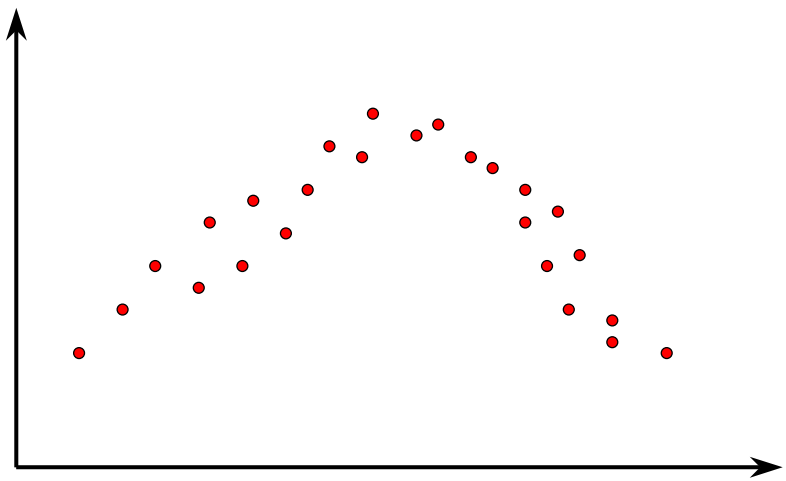
Но что делать в таком случае?



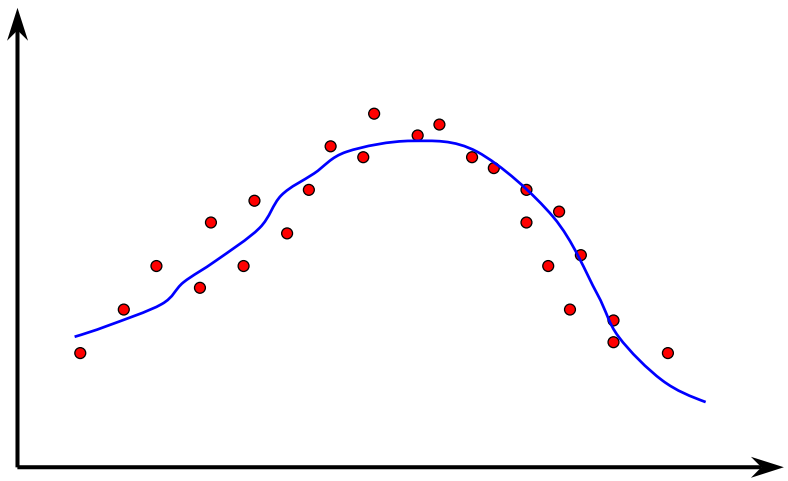


Какие бывают многообразия малой размерности?

Но что делать в таком случае?



Провести одномерную кривую...



Как оно работает

1. Разбиваем точки (геномы) на L классов случайным образом.
2. Для каждого класса вычисляем центр — среднее арифметическое

$$c_j^k = \frac{1}{l_k} \sum_{i=1}^{l_k} f_j^i \,, \qquad j = 1, 2, \ldots 63 \,.$$

3. Затем для каждой точки (генома) и для каждого центра вычисляем расстояние:

$$d_i^k = \rho\left(C^k,\ F^i\right) \,, \qquad i = 1, 2, 3, \ldots, M \,.$$

4. Переопределяем точки (геномы): геном уходит в тот класс, к центру которого он ближе всего.
5. Всё повторяем до тех пор, пока все точки не перестанут переходить из класса в класс.

Как оно работает: в чём проблемы

1. Метод динамических ядер не увеличивает число классов.
2. Надо бы проверять разделимость классов; мы не делали этого.
3. Метод чувствителен к начальному разбиению (оно случайное): проблема волатильных геномов (точек).
4. Проблема выбора начального числа классов.

Упругие карты — двумерные (нелинейные) многообразия, приближающие точки в (многомерном) пространстве.

- вычисляется корреляционная матрица между значениями координат точек;
- вычисляются первые две главные компоненты этой матрицы (= направления самых больших различий в данных);
- на двух этих векторах строится (обычная знакомая) плоскость. Каждая точка данных проектируется на неё;
- каждая точка связывается пружинкой с проекцией;
- плоскости разрешается деформироваться упруго, вся система отпускается и стремится в минимум потенциальной энергии;

Упругие карты — продолжение

- затем положения точек на карте переопределяются: каждая экспериментальная точка отображается на карте в ту, которая к ближе всего к экспериментальной;
- карта готова к употреблению (почти);
- карта подвергается нелинейному преобразованию, которое её «разглаживает»;
- вот теперь всё готово для дальнейшего анализа!

О структуре базы геномов

Геномы депонированы в EMBL–банке, использовался релиз от марта 2011 года.

Релиз содержал $3{,}5 \times 10^3$ геномов, использовалось только 1132 генома.

Были исключены «единичные» геномы, представляющие высокие таксоны: база содержит лишь те геномы, для которых в каждом роде содержится не менее 5 видов.

Таксон	*M*	Таксон	*M*	Таксон	*M*
Batrachia	51	*Chondrostei*	5	*Crocodylidae*	7
Cryptodira	25	*Dinosauria*	94	*Eutheria*	193
Gymnophiona	16	*Metatheria*	18	*Neopterygii*	500
Squamata	78				

Об исключении триплета

Всего триплетов 64, а пространство 63-мерное. Один триплет исключён: почему? Потому, что сумма частот всех триплетов равна 1.

Какой исключать?

Об исключении триплета

Всего триплетов 64, а пространство 63-мерное. Один триплет исключён: почему? Потому, что сумма частот всех триплетов равна 1.

Какой исключать? Такой, для которого дисперсия (по базе) самая маленькая: он даёт наименьший вклад в различение геномов.

Об исключении триплета

Всего триплетов 64, а пространство 63-мерное. Один триплет исключён: почему? Потому, что сумма частот всех триплетов равна 1.

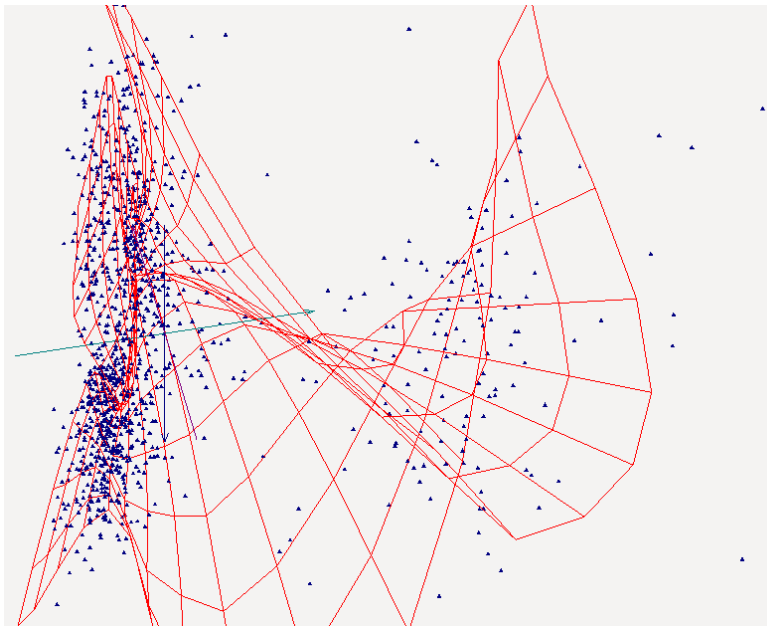
Какой исключать? Такой, для которого дисперсия (по базе) самая маленькая: он даёт наименьший вклад в различение геномов.

$\min\{\sigma\}$	
GCG	0,001299
TGA	0,001533
ATG	0,001560
CGA	0,001602
AGT	0,001607
GAT	0,001674

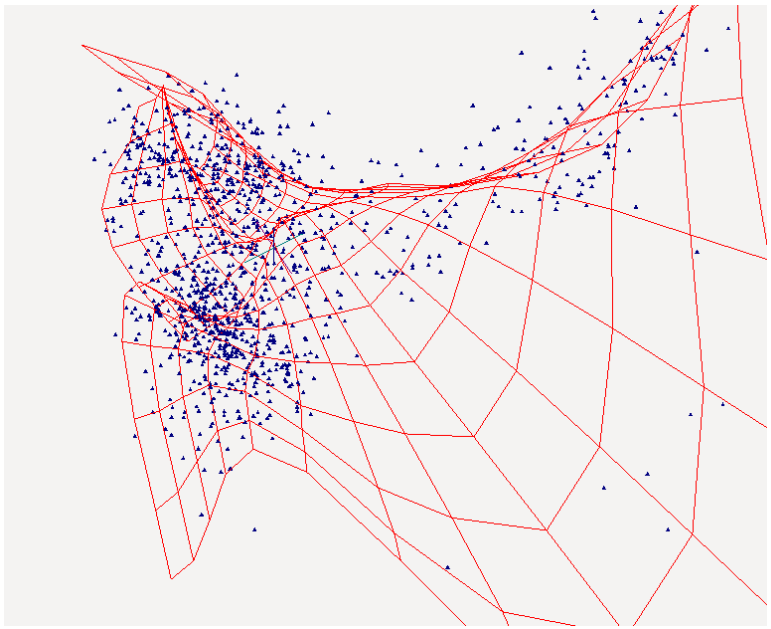
$\max\{\sigma\}$	
AAA	0,016346
TTT	0,015903
AAT	0,015026
ATT	0,014700
TTA	0,013165
TAA	0,013074

Линейная классификация: два класса

Разбиение оказалось очень устойчивым.

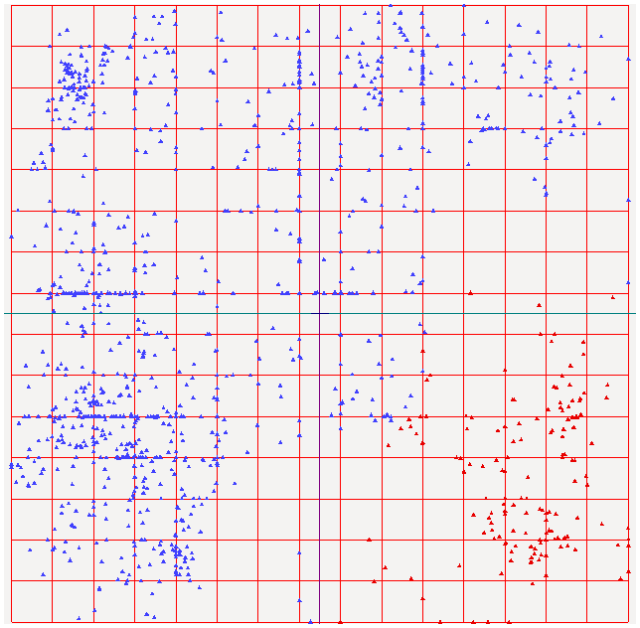


естественные координаты,

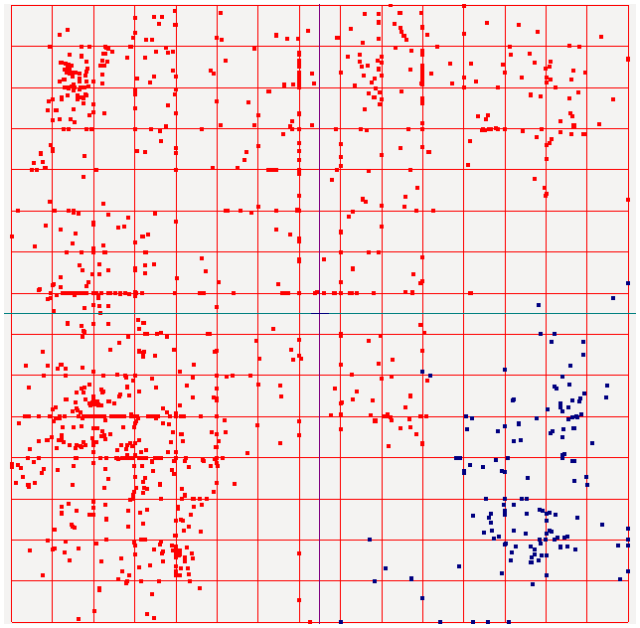


главные компоненты

Линейная классификация: два класса, метод динамических ядер



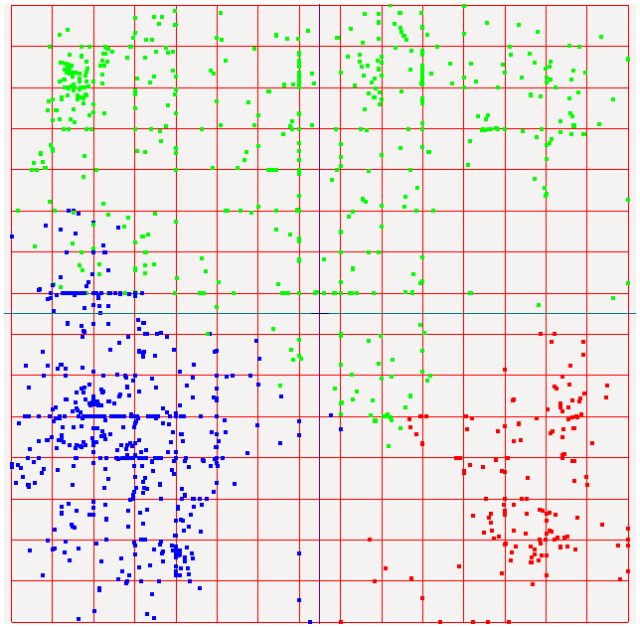
Классификация



Позвоночные/б-п

«Оппортунистами» были два генома из беспозвоночных и двенадцать — из позвоночных.

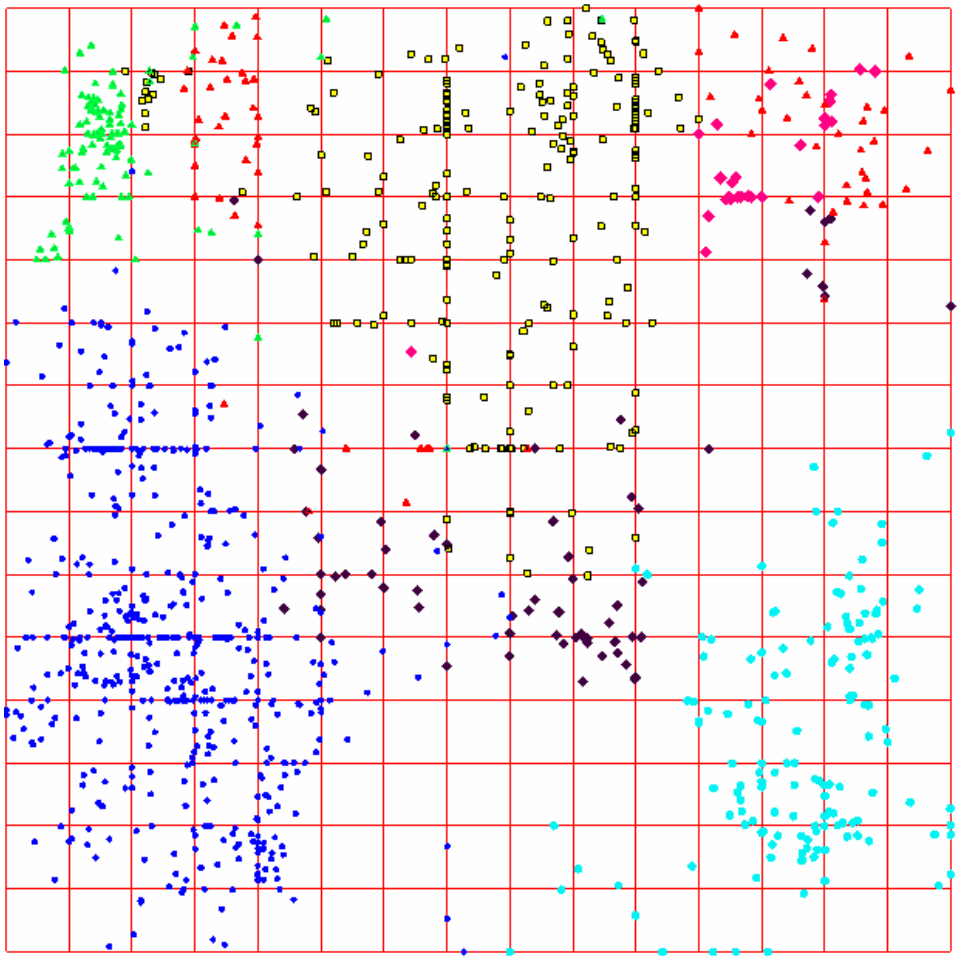
Линейная классификация: три класса, метод динамических ядер



Таксон	*N*	I	II	III
Actinopterygii	510	464	46	0
Amphibia	65	40	17	8
Archosauria и *Lepidosauria*	177	1	176	0
Mammalia	212	0	1	211
Neoptera	143	0	4	139
Testudines	25	0	25	0

Классификация на три класса менее устойчива; возможно вырождение в два класса.

Линейная классификация: три класса, распределение таксонов



- – *Actinopterigii*;
- – *Amphibia*;
- – *Archosaura*;
- – *Lepidozaura*;
- – *Mammalian*;
- – *Neoptera*;
- – *Testudines*.

Классификация сверху вниз

- Начинаем со всего массива данных и делим на два класса (если получится).
- Каждый получившийся класс делим в свою очередь на два (либо три или четыре) класса и т. д.
- Где-то останавливаемся, получив на выходе структуру типа дерева.

Данная структура полностью релевантна классической классификации на основе морфологических признаков.

Классификация снизу вверх

- По-прежнему, начинаем со всего массива данных и делим на два класса (если получится). Или на три; в общем случае — на минимальное, дающее устойчивое разбиение на классы.
- Затем весь массив данных делим на число классов, большее на единицу и так продолжаем до тех пор, пока всё более или менее устойчиво делится. Получаем серию классификаций

$$\mathfrak{C}_2, \mathfrak{C}_3, \mathfrak{C}_4, \ldots, \ \mathfrak{C}_{K-1}, \mathfrak{C}_K \, .$$

- Затем изучаем судьбу каждого класса при переходе от классификации $\mathfrak{C}_j$ к $\mathfrak{C}_{j-1}$.

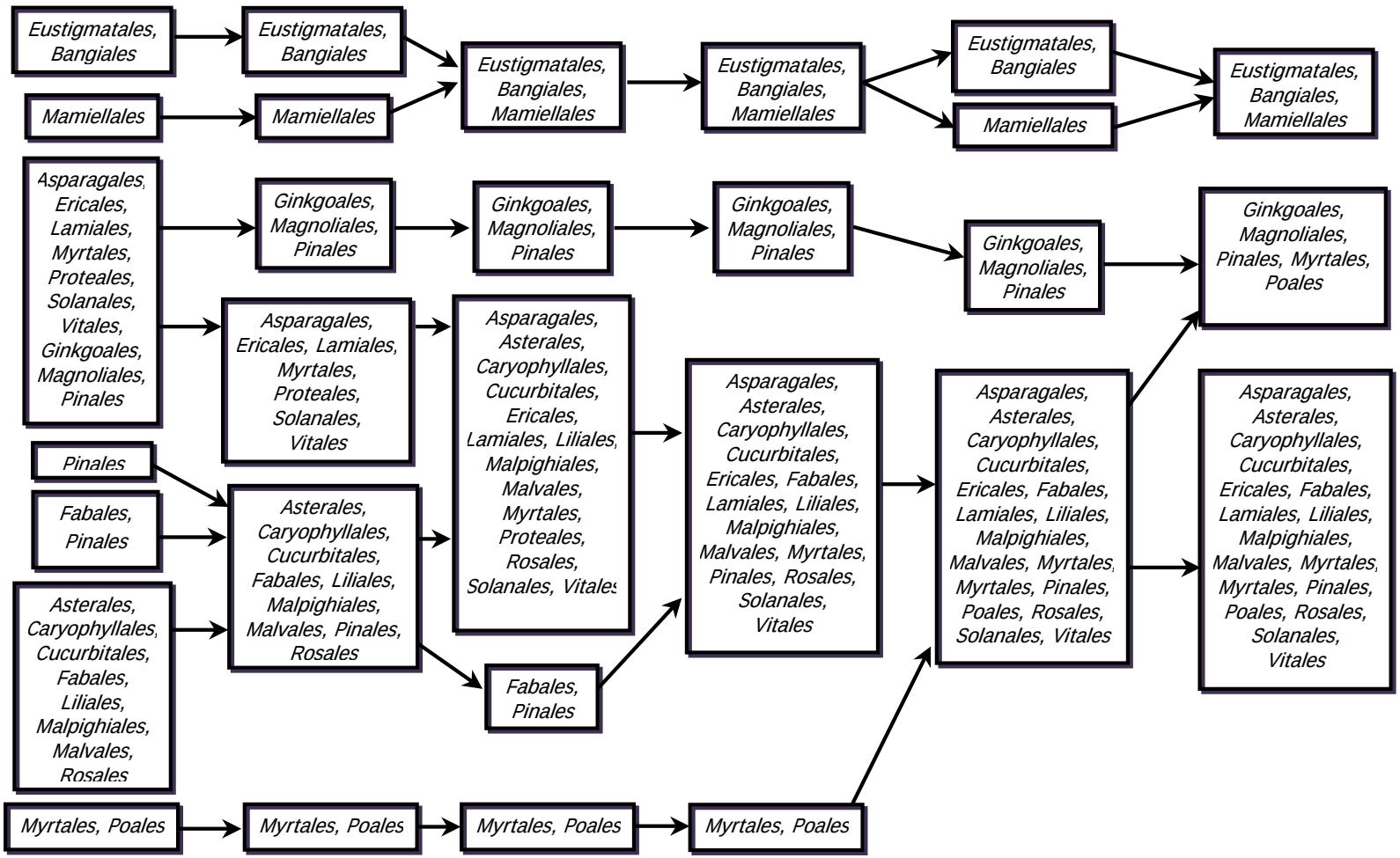
Классификация снизу вверх — продолжение

- Судьба может быть такой: некоторый класс $\mathfrak{C}_j^n$ целиком включён в класс $\mathfrak{C}_{j-1}^l$;
- некоторый класс $\mathfrak{C}_j^n$ в основном включён в класс $\mathfrak{C}_{j-1}^l$ и частично — в класс некоторый класс $\mathfrak{C}_j^n$ целиком включён в класс $\mathfrak{C}_{j-1}^m$;
- некоторый класс $\mathfrak{C}_j^n$ почти случайным образом распределился по набору классов $\mathfrak{C}_{j-1}^l$, $l = 1, 2, \ldots, l^*$.

В общем случае классификация «снизу вверх» даст структуру типа графа с циклами; чем дальше будет граф отстоять от полносвязного, тем лучше классификация.

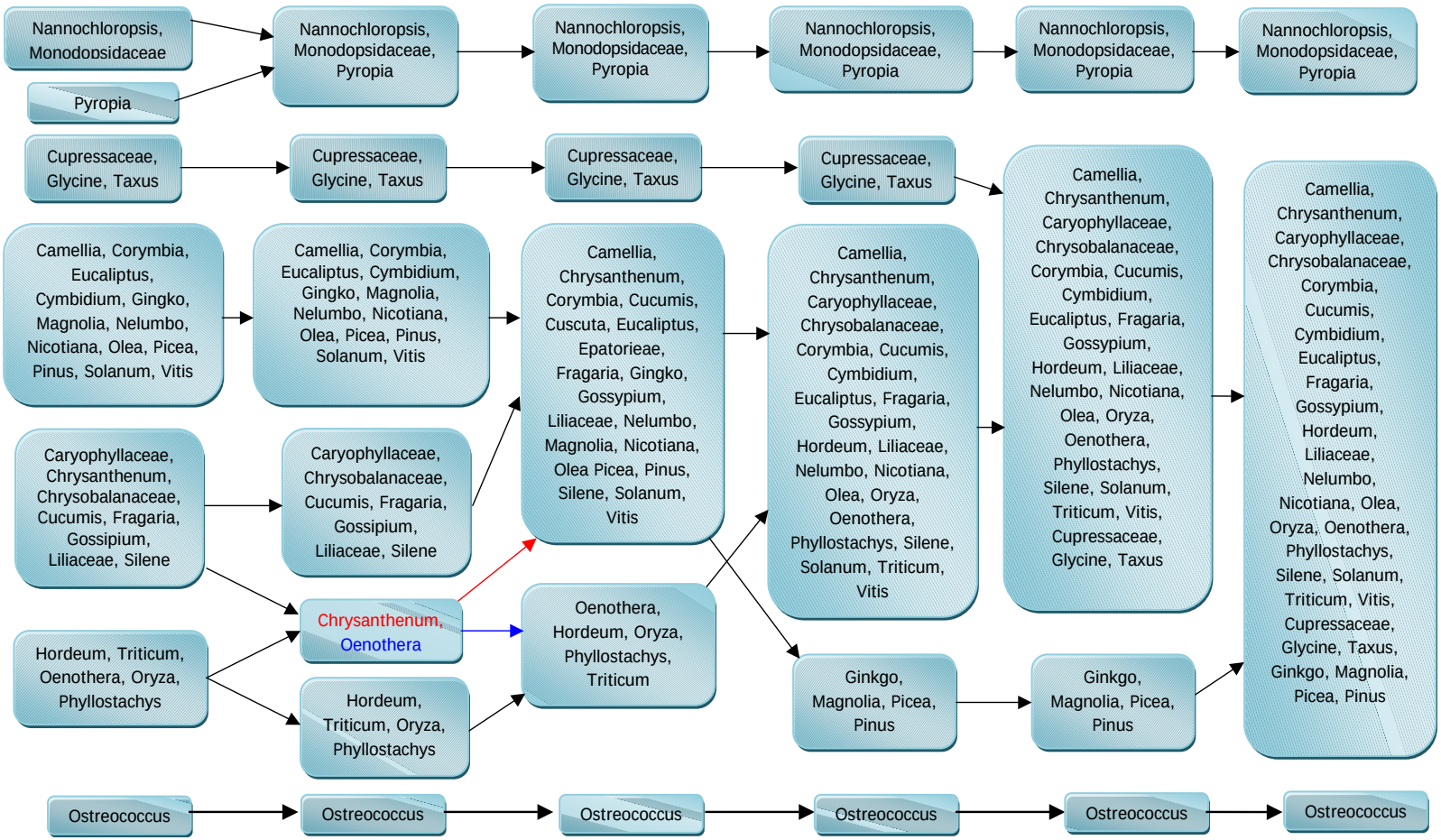
Классификация снизу вверх — пример

Примером такой классификации «снизу вверх» может быть классификация 249 геномов хлоропластов:

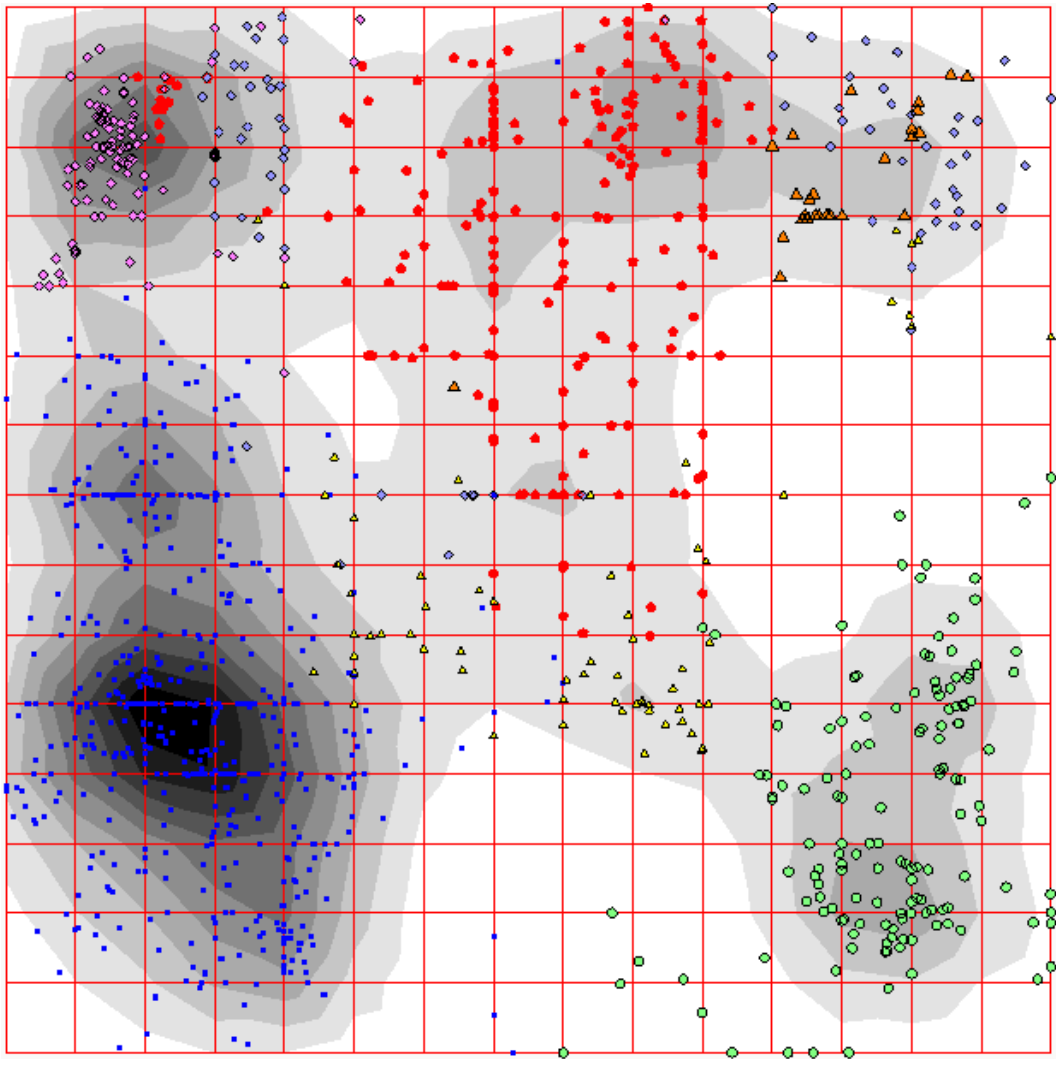


Классификация снизу вверх — пример

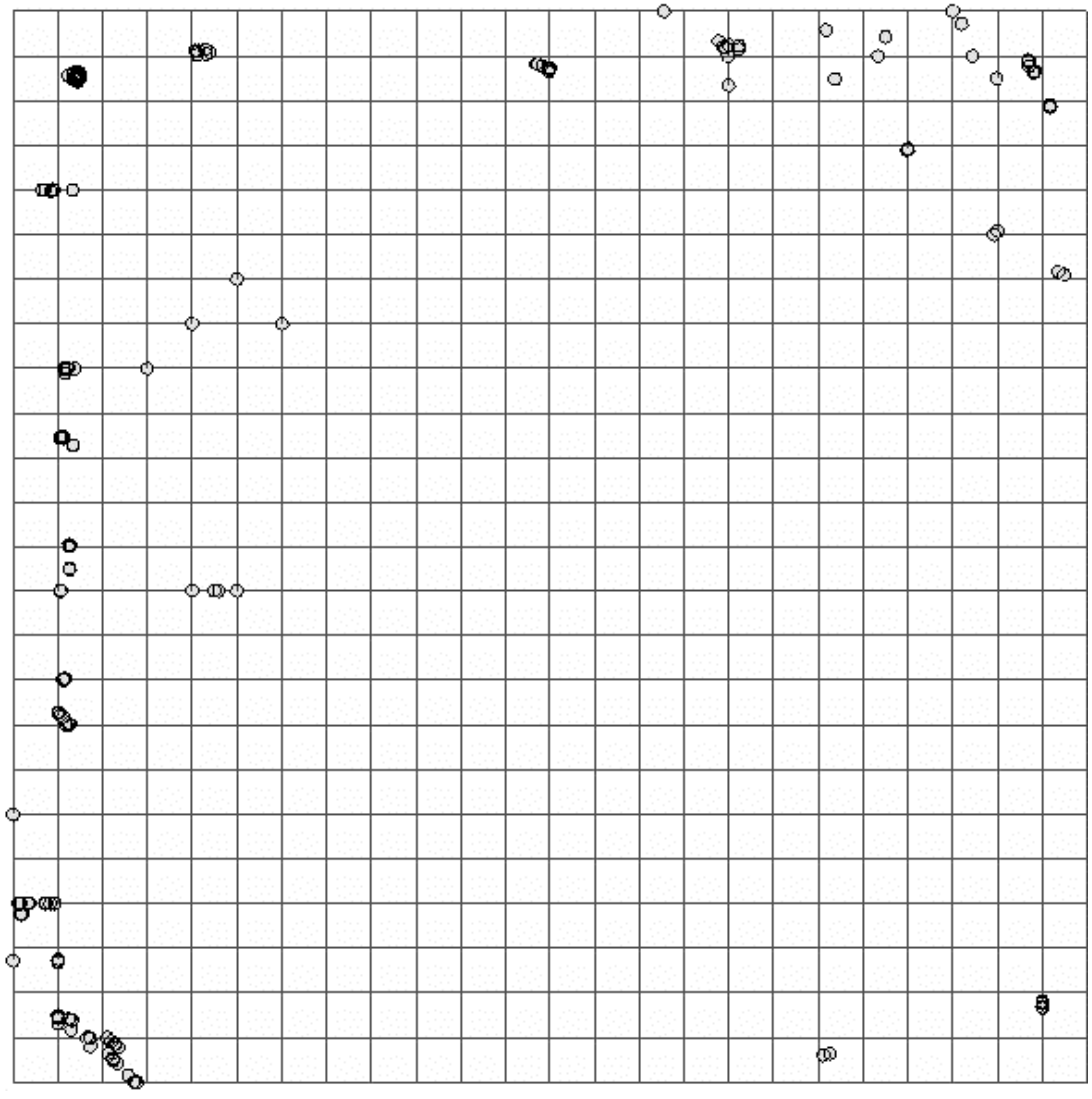
Что делать с волатильными геномами? Их тоже можно кластеризовать, понимая, что полученная кластеризация будет заметно менее устойчивой.



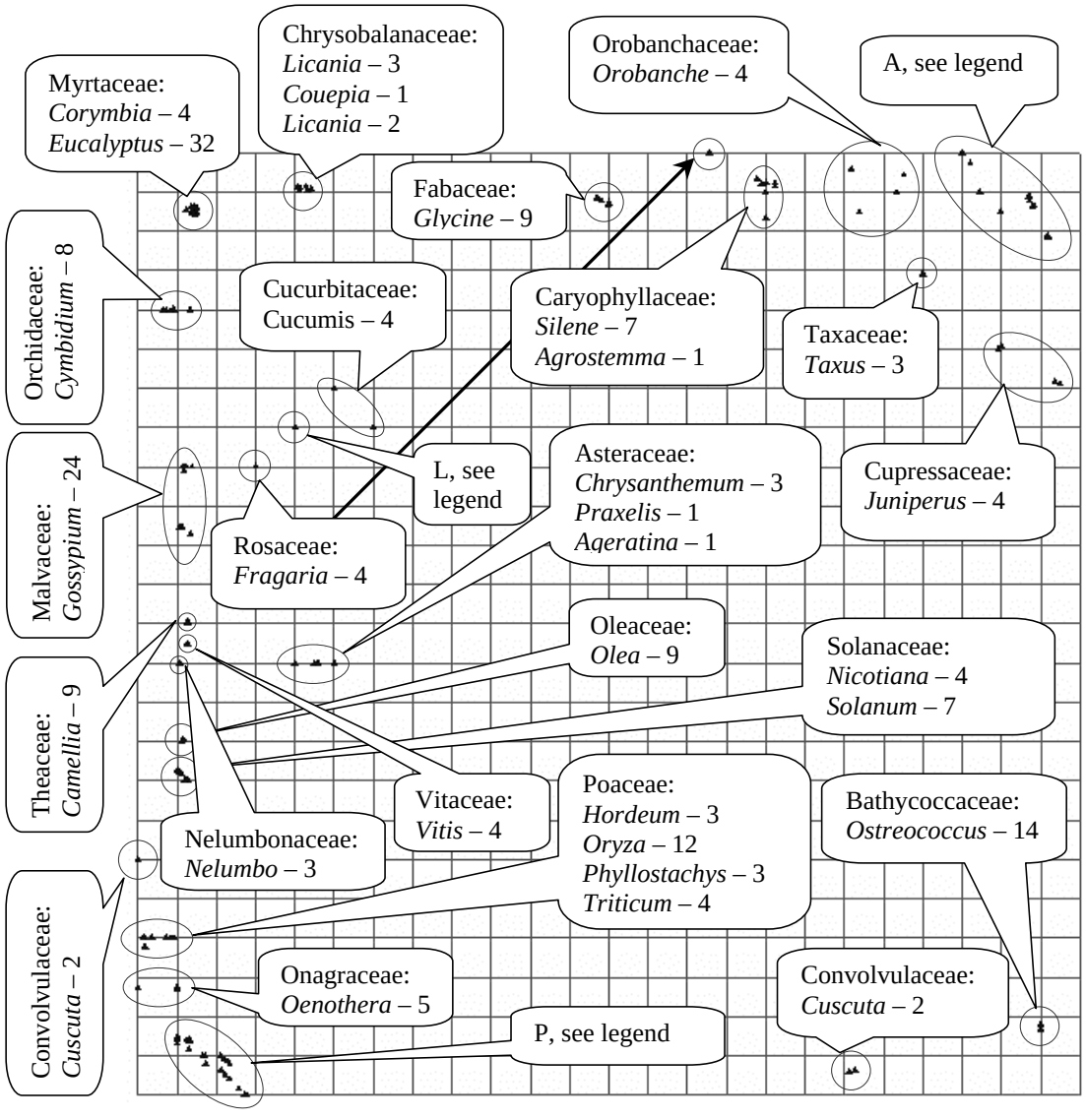
Кластеризация упругими картами. Пример с митохондриями



Кластеризация упругими картами. Случай хлоропластов



Кластеризация упругими картами. Случай хлоропластов



Заключение. Самое важное

1. Доказано существование очень высокого уровня коэволюции митохондриальных геномов и соматических. А ведь физически они друг с другом никак не связаны!
2. Указанное доказательство состоит в том, что кластеризуются организмы по геномам хлоропластов (митохондрий), а таксономия смотрится по соматическим геномам.
3. Полученные результаты следует проверить на
   - других органеллах (хлоропласты); сделано — всё прекрасно!
   - на различении структуры/функции: проверить распределение геномов митохондрий и хлоропластов одновременно.
   - на генах митохондрий и хлоропластов (в ближайших планах, желающие — три шага вперёд!).

Что осталось за бортом?

- Метод упругих карт: кластеризация требует искусства.
- Иные методы распознавания образов (в первую очередь — классификации с учителем).
- Метод топологических грамматик.

Спасибо за внимание!