BayesHammer - Hammer + Echo = Bayecho
Скрещивая подходы

Дмитрий Грошев

24.03.2012

Все делают ошибки:

- Бог делает ошибки
- Человек делает ошибки
- Секвенатор делает ошибки

Однако, последние можно попытаться исправить

Используя некоторые наблюдения, можно попытаться исправить ошибки в ридах

- Геном достаточно неравномерен, чтобы риды в большинстве случаев сильно отличались друг от друга
- Покрытие каждого нуклеотида значительно больше 1
- Секвенаторы Illumina редко делают инделы, чаще замены
- Секвенаторы Illumina делают больше ошибок ближе к концу рида
- Ошибки обычно независимы друг от друга и зависят от исходного нуклеотида

- k-меры кластеризуются по дистанции Хемминга ($k \approx 55$), кластер = группа связанности в графе Хемминга
- В полученных кластерах ищется консенсус
- Полученные k-меры считаются верными

- Так же находим кластеры k-меров
- Кластеризуем сами кластеры ещё раз, анализируя вероятности замены из q-values
- Находим центры кластеров — верные k-меры
- Расширяем множество верных k-меров — рид, целиком покрытый верными k-мерами, целиком верен
- Исправляем риды, внося изменения из исправленных k-меров

- Анализируются риды, а не k-меры
- Ищутся достаточно сильно перекрывающиеся риды
- Вводится понятие confusion matrix

$$\Phi^{(m)}_{b,b'} = \mathbb{P}(r_m = b' | H_m = b)$$

r_m — нуклеотид в риде в позиции m
H_m — нуклеотид в истинном сиквенсе в позиции m

- Ищутся наиболее вероятные нуклеотиды в перекрывающихся областях

Почему бы не взять лучшее от обоих?

- Анализировать риды, а не k-меры
- Кластеризовать перекрывающиеся области ридов, используя confusion matrix
- Возможно, кластеризация перекрытий позволит уменьшить влияние ошибок на стадии отбора пересечений

WUT?