1 Введение

Дезоксирибонуклеиновая кислота (ДНК) является своеобразным кодом жизни. Эта молекула хранит и передает генетическую программу развития и функционирования живого организма. В то же время, все функции ДНК зависят от ее соединений с белками. Поэтому, изучение ДНК-белковых взаимодействий актуально и привлекательно. Chip-seq (chromatin immunoprecipitation - sequencing)[] является одним из современных методов, позволяющим выделить участки ДНК связанные с конкретным белком (одинаково применим к разным белкам). Однако, по понятным причинам (сложный биологический эксперимент), погрешность данного метода не может быть нулевой, и безрассудная вера ему лишена смысла. По этому, обычно, к результатам подобных методов накладывается вероятностная модель. Конечно, это добавляет ряд существенных ограничений. Однако, в качестве неоспоримого плюса можно привести тот факт, что хорошо подобранная модель позволяет понять природу данных и, изучить их свойства.

Итого, нашей задачей является нахождением модели по последовательности чисел выдаваемых Chip-seq.

В настоящее время, в качестве семейства искомых моделей, активное приминение находит HMM (Hidden Markov Model)[] второго порядка с Пуассоновским испусканием. Данное семейство допускает предположение о том, что каждое состояние (наличие/отсутствие белка в заданной части генома) завист только от одного предыдущего. Можно ограничиться и более лояльным допущением о том, что состояние зависит от n предыдущих состояний, однако такое допущение резко увеличивает сложность модели ($O(2^n)$ параметров). Также, сложность заключается в подборе этого n и переобучении в случае, если не все состояния имеют одинаковые длины контекстов зависимости. Последннее замечание приводит к идее использования VOHMM (Variable Order Hidden Markov Model)[]

2 Скрытые марковские модели переменного порядка