

Слайд 1

Титульный лист

Слайд 2

Последовательность длины L — строка D состоящая из L символов алфавита Σ . Выравнивание последовательностей — размещение двух или более последовательностей друг под другом таким образом, чтобы было легче увидеть их сходные участки. Например, даны последовательности ACEAAFAE и CEAFDCE, если расположить их друг под другом, то не будет ни одного совпадения соответствующих символов, но если вставить пропуск восьмого символа в первой последовательности и пропуски первого и пятого символов во второй последовательности, то мы получим 5 совпадений.

Значимость выравнивания — действительное число s , отражающее сходство последовательностей. Способом вычисления значимости выравнивания s может быть, например, увеличение значимости на 1 при совпадении символов, стоящих друг под другом, и уменьшение на $\frac{1}{2}$ при несовпадении. Тогда значимость s приведенного выше выравнивания будет равна 3. Способ вычисления значимости выравнивания выбирается исходя из целей и вида выравнивания.

Сходство последовательностей может отражать функциональные, структурные или эволюционные взаимосвязи объектов, которые описывают эти последовательности. Таким образом вычисление значимости выравнивания последовательностей может быть полезно в задаче определения степени родства биологических организмов путем сравнения их ДНК или РНК, нуклеотидных последовательностей, задаче анализа свойств белков, аминокислотных последовательностей, задаче распознавания речи человека или письменного языка и многих других приложениях.

На слайде приведен пример попарного выравнивания двух строк, но если сходство последовательностей слабое, то через такое выравнивание может не выйти идентифицировать взаимосвязь описываемых последовательностями объектов. Однако сравнение сразу трех и более последовательностей может позволить выявить эту взаимосвязь, такое выравнивание называется множественным. Проводить множественное выравнивание стандартными методами динамического программирования для попарного выравнивания вычислительно неэффективно, но оказывается, что аппарат скрытых марковских моделей (СММ) позволяет эффективно решать эту задачу.

Слайд 3

СММ будут описаны далее, пока что зададимся следующим вопросом. Если есть множество последовательностей, описывающих взаимосвязанные объекты, имеется еще одна последовательность и была посчитана значимость выравнивания этой последовательности ко всему множеству каким-либо способом, то

- достаточно ли высокая эта значимость, чтобы считать объект, описываемый последовательностью, родственным к объектам, описываемым множеством, или шум, т.е. случайная последовательность, мог добиться такой значимости.

- достаточно ли низкая эта значимость, чтобы считать объект описываемый последовательностью, не родственным к объектам, описываемым множеством, или сигнал, т.е. последовательность, описывающая взаимосвязанный с множеством объект, мог получить такую значимость.

Ложноположительная вероятность значимости s — это вероятность того, что шум получит значимость равную или выше s .

Далее будет описаны метод, который позволяет эффективно вычислять введенный термин.

Слайд 4

Сначала опишем модели, затем алгоритмы, которые используются для манипуляции ими.

Метод предполагает, что даны профильная СММ, с помощью которой будут оцениваться последовательности, и фоновая модель B , которая будет описывать шум.

Определение 1. Пусть X_n и Y_n дискретные стохастические процессы, $n \geq 1$. Пара (X_n, Y_n) называется скрытой марковской моделью, если

- X_n — марковский процесс, поведение которого напрямую не наблюдается ("скрытый");
- $P(Y_n = y_n | X_1 = x_1, \dots, X_n = x_n) = P(Y_n | X_n = x_n)$ для любого $n \geq 1$, где x_1, \dots, x_n — значения, принимаемые процессом X_n (**состояния модели**), y_n — значение, принимаемое процессом Y_n (**наблюдаемый символ модели**).

Слайд 5

Примером простой СММ может быть модель, изображенная на рисунке 1 и описывающая подбрасывание двух монет. Пусть между наблюдателем и человеком с монетами стоит ширма, которая позволяет наблюдателю видеть только пол, куда падают монеты. Пусть есть две монеты: одна — честная монета, вторая — нечестная монета с перевесом в одну из сторон. Пусть человек с монетами с некоторой вероятностью либо подбрасывает монету, которую он бросил в прошлый раз, либо меняет монеты и бросает новую. При этом наблюдатель не знает, какая монета используется в конкретный момент времени, так как он не видит рук бросающего монеты и не может отличить одну монету от другой по их внешнему виду, он видит только последовательность результатов бросков.