УДК 519.168

Кластеризация белков с учётом их доменной структуры

Зверков О.А., Горбунов К.Ю., Селиверстов А.В., Любецкий В.А.
Институт проблем передачи информации им. А.А. Харкевича РАН

На примере белков рассматриваются *три подхода к кластеризации данных*, каждый из которых показал хороший результат на соответствующем типе данных.

Важные белки имеют несколько субъединиц, каждая из которых имеет несколько доменов. Число субъединиц, доменов и взаимное расположение доменов могут значительно различаться даже у близких белков. Информацию о субъединицах и их доменах можно получить в базах данных. Субъединицы и домены имеют разные степени консервативности, которые определяются стандартными методами, или они просто указаны в базе данных Pfam. Например, в составе β-субъединицы РНК-полимеразы бактериального типа имеются семь доменов, хотя в отдельных видах некоторые из них могут отсутствовать [1].

В докладе будут рассмотрены две тесно связанные между собой задачи о кластеризации данного набора белков и о качестве некоторой уже полученной кластеризации. Для их решения предлагаются три подхода, по существу, применимые к любым данным, не обязательно к белкам. Будем считать, что белки изображаются точками конечномерного банахового пространства, а для простоты изложения качество кластеризации оценивается только двумя значениями: 1 («да») и 0 («нет»). Каждый домен соответствует определённой координате этих точек, поэтому размерность пространства равна максимальному числу различных доменов у рассматриваемых белков.

Первый подход. Назовём «хорошей» кластеризацию, для которой выполняется следующее: для любых двух её кластеров эллипсоиды минимального объёма, описанные около них, не пересекаются. Такой эллипсоид (называемый эллипсоидом Левнера) определяется однозначно в соответствующем подпространстве [2]. Эллипсоиды – это аффинные многообразия, заданные многочленами второй степени с положительно определённой квадратичной формой. Объём эллипсоида зависит только от нормы на координатных осях.

Второй подход. Близость двух белков как точек пространства

определяется нормой – суммой модулей координат («метрикой L_1»). «Сжатый ортоплекс» – шар в этой метрике, сжатый вдоль части координатных осей. Назовём кластеризацию «хорошей», если для любых двух кластеров сжатые ортоплексы минимального объёма, описанные около них, не пересекаются. Мы предложили простые алгоритмы проверки, является ли данная кластеризация хорошей для подходов 1-2.

Третий подход. Здесь кластеры строятся следующим образом. Фиксируем норму в исходном пространстве. Для заданного r определим r-граф: его вершины – все n точек пространства, которым заранее сопоставлены белки; две вершины соединены ребром, если расстояние между ними не больше r. С небольшим шагом увеличиваем r от нуля и определяем функцию $f(r)$ в точках сетки, равную числу связных компонент в r-графе. Эта функция невозрастающая, меняется от n до 1. Ищем максимально длинный отрезок, на котором $f(r)$ остаётся постоянной или медленно меняется. На этом отрезке можно выбрать точку, для которой связные компоненты r-графа соответствуют кластерам. Такая процедура зависит от параметров, которые можно успешно подбирать.

Вместо связных компонент нами рассматривались двусвязные компоненты (что позволяет выделить два кластера даже в случае, когда они соединены тонкой «перемычкой»). Вместо числа компонент нами рассматривалось число: пусть x – одна из точек, обозначим $|x|$ число элементов в компоненте, содержащей x; положим $f(r)$ равным среднему значению $|x|$ по всем точкам x. Эта функция неубывающая, меняется от 1 до n (что позволяет с ростом r присоединять маленькие кластеры к подходящим большим кластерам). Мы предложили эффективный алгоритм такой кластеризации.

Литература

1. *Cramer P., Bushnell D.A., Kornberg R.D.* Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution // Science. – 2001. – V. 292. P. 1863–1876.
2. *Загускин В.Л.* Об описанных и вписанных эллипсоидах экстремального объёма // Успехи математических наук. – 1958. – Т. 13, № 6. – С. 89–93.