Математические модели в биологии и медицине

На основе аппарата байесовских сетей построен алгоритм распознавания вторичной структуры белков. Проведено сравнение полученного результата с другими известными байесовскими процедурами.

УДК 519.68

А.М. ГУПАЛ, С.С. РЖЕПЕЦКИЙ

ЭМПИРИЧЕСКИЕ БАЙЕСОВСКИЕ СЕТИ ДЛЯ РАСПОЗНАВАНИЯ СТРУКТУРЫ БЕЛКОВ

Введение. В данной работе рассмотрена задача распознавания вторичной структуры белков. Цель исследования – построение эффективных байесовских процедур, позволяющих предсказывать вторичную структуру белка по его аминокислотной последовательности.

Рассматриваемая задача принадлежит к области биоинформатики. Биоинформатика находится на стыке сразу нескольких различных научных областей и преследует в первую очередь сугубо прикладные и практические цели. Главная задача специалиста в данной области заключается в создании алгоритмов и прикладных инструментов способных решать прикладные задачи с заданной степенью точности, потребляя при этом ограниченные ресурсы. Поэтому основной интерес представляют следующие вопросы:

• нахождение эффективной байесовской процедуры для решения задачи;

• сравнение точности результата и сложности метода с уже известными байесовскими процедурами, применявшимися для решения задачи;

• способность процедуры распознавать не гомологические белки (не имеющие длинных общих подпоследовательностей аминокислот).

Первый раздел данной работы посвящен краткому введению в задачу распознавания вторичной структуры белков. Дано определение основных понятий и указана прикладная важность рассматриваемой задачи.

Во втором разделе дано краткое введение в теорию байесовских сетей. Байесовские сети рассматриваются в контексте других байесовских процедур. В конце раздела сделан краткий обзор литературы и сравниваются результаты других подходов с применением байесовских процедур для решения данной задачи.

Раздел три содержит описание исследуемой выборки белков, методики обучения, тестирования и валидации, а также метрики оценки точности распознавания, которые применялись в этом исследовании. Приводятся результаты, полученные с использованием наиболее известных и широко используемых методик построения байесовских сетей.

В четвертом разделе предложено эмпирическое улучшение байесовской сети найденной в разделе три на основе анализа других байесовских процедур, описанных во втором разделе. Показано, что эмпирическая сеть превосходит предыдущие процедуры по качеству распознавания.

В последнем разделе подведен краткий итог статьи с основными результатами и их анализом. Указаны направления для дальнейших исследований.

Раздел 1. Распознавание пространственной структуры белков – это задача построения пространственной модели белка исходя из записи его аминокислотной последовательности. Решение этой задачи является критически важным для таких областей как медицина (например, фармакология и разработка новых лекарств) и биотехнология (например, для получения новых ферментов).

Пространственная структура также постепенно вытесняет аминокислотную последовательность и как основа для методов классификации белков. Если на раннем этапе развития биоинформатики неизвестные белки пытались классифицировать по степени подобия их аминокислотных последовательностей, то в настоящее время все чаще применяется классификация на основе анализа пространственной структуры. Отдельно следует отметить, что именно пространственная структура определяет функцию и свойства белка.

Белок представляет собой цепь аминокислот соединенных пептидными связями. Всего есть 20 аминокислот, которые могут входить в состав белка. Аминокислоты имеют различные заряды и делятся на гидрофильные и гидрофобные. В зависимости от заряда и гидрофильности отдельных аминокислот, в процессе сворачивания аминокислотной цепи в белок во время его синтеза, он принимает конкретную, для данной аминокислотной последовательности, пространственную структуру.

В пространственной структуре белков выделяют характерные вторичные структуры, состоящие из нескольких аминокислот и формирующие узнаваемый паттерн. Распознавание этих вторичных структур по аминокислотной записи белка является важным промежуточным шагом к полному распознаванию пространственной структуры белка.

Для классификации типов вторичных структур существует общепринятый стандарт DSSP [1]. Согласно этому стандарту выделяют следующие вторичные структуры:

- G = спираль из трех витков;
- H = спираль из четырех витков;
- I = спираль из пяти витков;
- T = поворот состоящий из водородных связей;
- E = продолжение нити в складке;
- B = остаток в изолированном β-мосте;
- S = изгиб;
- C = нерегулярный участок.

На практике вместо стандарта DSSP, определяющего 8 типов структур, подавляющее число исследований использует всего три структуры: спираль (H, G и I по стандарту DSSP), лист (E и B по стандарту DSSP) и нерегулярность (I, S и C согласно DSSP).

На рис. 1 показано пространственную структуру белка и схематическое изображение той же структуры с четким выделением трех исследуемых типов вторичной структуры (спирали, листы и нерегулярности).

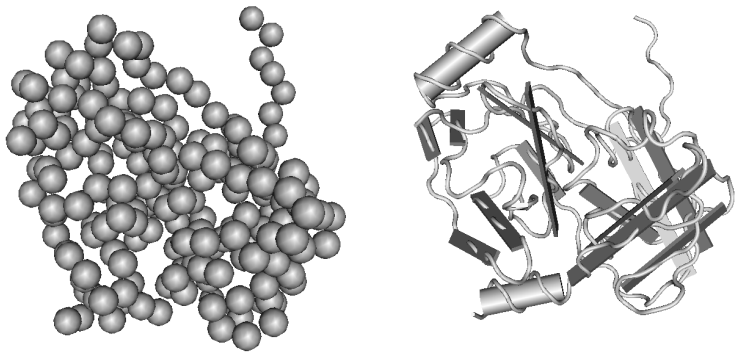

РИС. 1. Пример пространственной структуры белка и его схематическое изображение в виде вторичных структур

Отдельно следует отметить, что в статье рассматривается задача распознавания негомологических белков. То есть, белков не имеющих достаточно длинных общих подпоследовательностей. Задача распознавания вторичных структур гомологических белков в настоящее время решена с точностью превышающей 90 % и большого интереса не представляет.

Таким образом, задача распознавания вторичной структуры белка сводится к задаче перевода 20-ти буквенной аминокислотной последовательности в 3-ех буквенную последовательность вторичной структуры, в которой находятся соответствующие аминокислоты.

Раздел 2. Данная работа является продолжением предыдущей серии работ авторов, посвященной исследованию эффективности байесовского подхода и байесовских процедур распознавания в биоинформатике. Отличительными чертами байесовских процедур является их:

- индуктивность;
- эффективность (показано, что байесовская процедура распознавания на независимых признаках является субоптимальной);
- простота вывода результата и скорость расчета для уже обученной модели [2].

С другой стороны, основной сложностью при построении байесовских процедур распознавания является выбор модели структурных зависимостей между наблюдаемыми и исследуемыми переменными. Простейшим случаем является игнорирование таких зависимостей вообще, и использование модели независимых признаков. Примерами более сложных моделей могут служить: методы на цепях Маркова, деревья, марковские модели со скрытыми переменными (HMM) и т. п. В этом контексте, байесовские сети предоставляют как наиболее общую модель структуры зависимостей между переменными, так и методы нахождения этих моделей из данных.

Байесовская сеть – это вероятностная модель на ациклических ориентированных графах, представляющая набор случайных переменных и их условных зависимостей в графовой форме. Каждая вершина на графе представляет конкретную случайную переменную, а ребро – наличие условной зависимости между двумя переменными. Каждой вершине соответствует своя вероятностная функция, которая в качестве переменных использует набор значений вершин родителей на графе, а в качестве результата выдает вероятности значений для исследуемой вершины.

Байесовские сети применяются для целого ряда задач:

- установление и анализ зависимостей между исследуемыми переменными модели (например, анализ влияния отдельных стран на рынок валют [3]);
- предсказание и распознавание;
- классификация.

Существует отдельный класс байесовских сетей для случая, когда модель может быть представлена, как последовательность переменных (например, в теории сигналов, распознавании голоса и т. п.). Такой класс моделей называется динамическими байесовскими сетями. Динамическая байесовская сеть по своей графовой структуре представляет собой несколько обычных байесовских сетей соединенных в последовательность. Такие модели гораздо более требовательны к вычислительным ресурсам, чем байесовские сети, и обладают гораздо большей сложностью.

Для практического использования байесовской сети в задачах распознавания необходимо три различных алгоритма, для решения следующих задач:

- нахождение графа байесовской сети;
- нахождение функций условных вероятностей для каждой вершины полученного графа;
- распознавание неизвестного образца с помощью обученной модели.

В работе для последних двух задач применялись стандартные известные алгоритмы, и интерес представляет только алгоритм нахождения графа байесовской сети. По сложности эта задача принадлежит к классу NP-полных. Показано, что в общем случае количество возможных ациклических ориентированных графов (АОГ) – суперэкспонента от количества вершин графа. В настоящее время все алгоритмы нахождения графа байесовской сети принято делить на две группы. К первой группе алгоритмов относятся поисковые алгоритмы, использующие различные поисковые процедуры и целевые функции, позволяющие оценивать качество модели. Ко второй группе относятся сепарационные алгоритмы, использующие различные статистические критерии для определения условных независимостей между переменными в модели и строящими граф сети на основе этой информации. Обе группы моделей имеют свои положительные и отрицательные стороны. Так, поисковые алгоритмы иногда не способны найти решение задачи, из-за попадания в локальный экстремум оценочной функции, при этом как сепарационные алгоритмы для определенного класса задач требуют чрезмерных вычислительных ресурсов.

Приведем обзор результатов других байесовских подходов, применявшихся для решения задачи распознавания вторичной структуры.

Исторически одним из первых известных методов была процедура GOR (названа именами ее авторов: Garnier, Osguthorpe, Robson), созданная в 1970-х годах [4]. GOR, как и большинство методов того времени, основывался на вероятностных параметрах, полученных эмпирическим путем, в результате изучения пространственных структур известных белков с помощью рентгеновской кристаллографии. Однако, в отличие от существовавших тогда методов, GOR учитывал не только вероятность попадания определенной аминокислоты в ту или иную вторичную структуру, но и условную вероятность того, что соседние аминокислоты тоже находятся в этой структуре. Таким образом, GOR по своей сути использовал байесовский подход.

В настоящее время для решения рассматриваемой задачи среди байесовских процедур чаще всего применяются HMM. Точность распознавания таких моделей достигает 73 % [5]. Отметим также, что в работе [6] удалось добиться точности распознавания в 78 %, используя обычную модель марковских цепей второго порядка и некоторую эвристику.

Что касается методов распознавания с помощью именно байесовских сетей, то в существующих публикациях используются в основном динамические байесовские сети, способные распознавать вторичные структуры с точностью до 78 % [7]. Обычные, не динамические сети, до настоящего времени не давали результата сравнимого даже с HMM, и нам не известно о серьезных исследованиях в данном направлении.

Одна из целей этой работы – построение метода распознавания вторичной структуры белков именно с помощью не динамических байесовских сетей, и достижения сравнимого с динамическими сетями уровня точности предсказаний. Таким образом, мы хотим с одной стороны построить модель более

сложную, чем модели Маркова и убедится, что результат улучшится. А с другой – мы стремимся построить модель более простую, чем динамические байесовские сети, без потери в точности распознавания.

Также следует упомянуть, что обычные байесовские сети в настоящее время часто используются как мета-классификаторы. В качестве переменных в этих сетях выступают предсказания вторичной структуры, полученные несколькими другими методами. Подобное использование байесовских сетей в некоторых случаях позволяет улучшить средний результат других методов на 3 – 5 %.

Раздел 3. В работе использовалась обучающая выборка из проекта EVA [8]. EVA создан для определения точности распознавания и сравнения существующих методов предсказания вторичной структуры белков. Выборка EVA обладает следующими характеристиками:

- содержит 3000 белков;
- белки не содержат одинаковых подпоследовательностей длиннее 5 аминокислот;
- представлены белки имеющие различное происхождение.

Выборка EVA применялась так же и в [5 – 7], что позволяет производить сравнение полученных результатов.

Валидация результатов проводилась по методике «без одного». Белки по очереди по одному исключались из обучения, а затем для них проводилась оценка точности распознавания, после чего результат был усреднен для всех белков в выборке. Для оценки точности распознавания вторичной структуры в белке использовалась следующая формула для случая определения точности распознавания конкретной структуры:

$$C_\alpha = \frac{p_\alpha n_\alpha - u_\alpha o_\alpha}{\sqrt{[n_\alpha + u_\alpha][n_\alpha + o_\alpha][p_\alpha + u_\alpha][p_\alpha + o_\alpha]}},$$

где p_α – количество верно определенных оснований, попадающих в структуру типа α; n_α – количество верно определенных оснований, не попадающих в структуру типа α; u_α – количество неверно определенных оснований, попадающих в структуру типа α; o_α – количество неверно определенных оснований, не попадающих в структуру типа α.

Также использовалась метрика точности распознавания, обозначаемая в литературе, как C_3 – отношение правильно предсказанных вторичных структур к всей длине белка. Далее, если не указано иначе, используется метрика C_3.

При выборе переменных и построении байесовской сети применялся рамочный метод. Рамка длины L при обучении пошагово смещалась вдоль аминокислотной последовательности белка. В качестве наблюдаемых переменных выступали L аминокислот попавших в рамку, а вторичная структура центральной аминокислоты в рамке выступала в роли исследуемой переменной. Таким образом, обучающая выборка состояла из всех возможных рамок. На основе этой обучающей выборки строился граф байесовской сети из $L+1$ вершины.

Для построения графа байесовской сети применялись алгоритмы обеих типов, упомянутых во втором разделе. В качестве основы для реализации алгоритма нахождения графа байесовской сети с помощью поискового подхода испытывались все алгоритмы из пакетов BNT [9] и GeNIe [10]. Также был испробован и сепарационный подход, а именно метод PC и его улучшенная модификация [11]. Всего было испробовано 9 известных алгоритмов.

Поисковые методы превзошли по точности распознавания сепарационные. Наилучшая точность распознавания вторичной структуры белков оказалась равной 56 %. Этот результат был получен сразу несколькими поисковыми методами (GES, K2, MSWT), построившими идентичный граф сети (рис. 2).

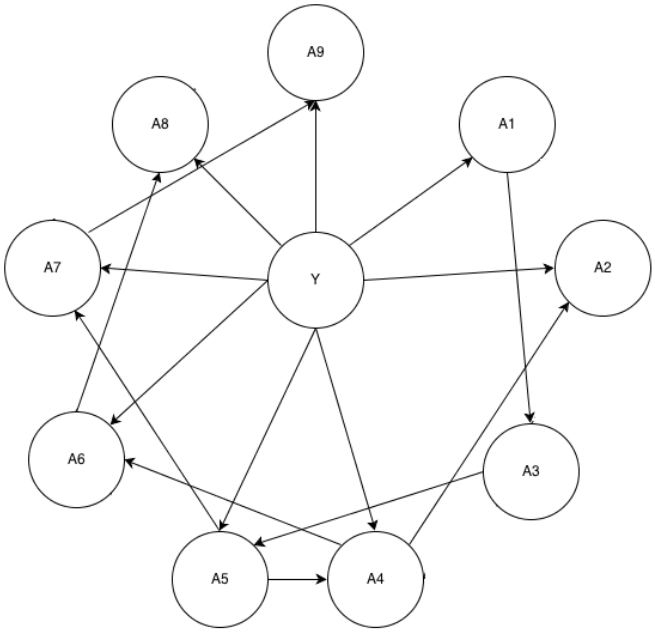

РИС. 2. Граф байесовской сети для рамки размером L = 9 аминокислот:
A1 – A9 – аминокислоты входящие в рамку; Y – вторичная структура
для центральной аминокислоты (A5)

Таким образом, известные методы построения графов байесовских сетей не позволяют решить задачу распознавания вторичной структуры с необходимым уровнем точности распознавания.

Кроме описанной модели сделаны попытки расширить количество переменных, добавив дополнительную информацию. Например, построены байесовские сети, которые кроме самих аминокислот входящих в рамку, содержали еще и гидрофобность, заряд и массу для каждой аминокислоты. Полученные таким образом сети, соответствовали уже известным данным о степени влияния различных факторов на вторичную структуру. Например, граф сетей четко

подтверждал то, что гидрофильность оказывает на структуру большее влияние, чем заряд аминокислоты. Тем не менее, описанный способ позволил увеличить точность распознавания не более чем на 2 %.

Раздел 4. Была сделана гипотеза, что низкий результат обычной байесовской сети, объясняется несовершенством существующих алгоритмов нахождения графа сети, а не несовершенством самой модели байесовских сетей в целом.

Одним из главных недостатков поисковых алгоритмов построения байесовских сетей, является их подверженность проблеме попадания в локальные экстремумы, после чего поиск обычно заканчивается, не дойдя до оптимальных решений. Сделана попытка применить описанные в предыдущем разделе алгоритмы, но заменить в них оценочную функцию, которая используется в поисковых алгоритмах для оценки качества текущего графа. В упомянутых алгоритмах в качестве такой функции выступал байесовский критерий информации (BIC). Одним из свойств BIC является более высокий штраф за сложность модели, в сравнении с другими аналогичными критериями. Поэтому, алгоритмы были модифицированы, и вместо BIC использован критерий Акаике (AIC), а также другие формулы, являющиеся его модификациями и накладывающие меньший штраф на сложность структуры графа, чем BIC. Описанные попытки, хотя и привели в некоторых случаях к более эффективным сетям, но не позволили получить точность сравнимую с моделями марковских цепей [6] или динамических байесовских сетей [7].

Примечательно, что в [6] получен лучший результат, чем в [5], и при этом использовалась более простая модель, а именно цепи Маркова второго порядка. Сделана эмпирическая модификация графа на рис. 2, таким образом, чтобы каждая переменная аминокислоты по возможности была связана с двумя предыдущими и двумя последующими аминокислотами в рамке. При этом граф должен был остаться ациклическим и направленным. Результат такой модификации показан на рис. 3. Достроенные ребра отмечены пунктиром.

Для предложенной модифицированной сети далее был применен стандартный алгоритм и рассчитана эффективность модели. Общая точность распознавания составила $C_3 = 83$ % при размере рамки $L = 9$. Для отдельных типов вторичных структур, точность рассчитанная по формуле из раздела 3, составила:

- для спиралей – $C_\alpha = 79$ %;
- для листов – $C_\beta = 86$ %;
- для нерегулярностей – $C_\gamma = 75$ %.

Полученный результат превосходит другие байесовские процедуры упомянутые в статье, и подтверждает нашу гипотезу о несовершенстве существующих алгоритмов построения байесовских сетей.

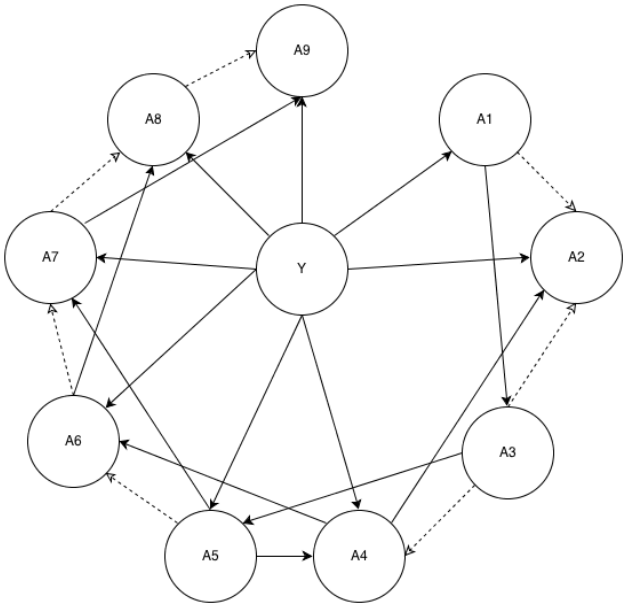

РИС. 3. Эмпирически модифицированный граф байесовской сети для распознавания вторичной структуры белков

Заключение. Рассмотрена задача распознавания вторичной структуры белков с помощью байесовских процедур распознавания. Сделан обзор существующих процедур и их эффективности. Построена эмпирическая байесовская сеть превосходящая другие методы по точности предсказания. Проверена гипотеза о том, что байесовская сеть, будучи более общей моделью, чем цепи Маркова, или модели Маркова со скрытыми переменными, позволяет улучшить точность предсказания вторичной структуры белков. С другой стороны показано, что обычные байесовские сети работают не хуже и даже способны превосходить результат, который показывают динамические байесовские сети.

Поднята проблема недостаточной эффективности существующих методов нахождения графов байесовских сетей. Предложено в качестве дальнейшей работы в этом направлении улучшать оценочную функцию поисковых методов построения скелета байесовской сети.

А.М. Гупал, С.С. Ржепецький

ЄМПІРИЧНІ БАЙЄСІВСЬКІ МЕРЕЖІ ДЛЯ РОЗПІЗНАВАННЯ СТРУКТУР БІЛКІВ

За допомогою апарата байєсівських мереж побудовано алгоритм розпізнавання вторинної структури білків. Отриманий результат порівнюється з іншими відомими байєсівськими процедурами.

A.M. Gupal, S.S. Rzhepetskiy

EMPIRICAL BAYESIAN NETWORKS FOR PROTEIN SECONDARY STRUCTURE PREDICTION

An algorithm for protein secondary structure prediction using bayesian networks models is constructed. The result obtained is compared with the other known bayesian methods.

1. *Kabsch W., Sander C.* Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features // Biopolymers. – 1983. – V. 22, N 6. – P. 257 – 637.
2. *Гупал А.М., Пашко С.В., Сергиенко И.В.* Эффективность байесовской процедуры распознавания // Кибернетика и системный анализ. – 1995. – № 4. – С. 76 – 89.
3. *Гупал Н.А., Ржепецкий С.С.* Применения аппарата байесовских сетей для анализа рынка валют // Компьютерная математика. – 2010. – № 1. – С. 94 – 101.
4. *Garnier J., Osguthorpe D.J., Robson B.* Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins // J. Mol Biol. – 1978. – V. 120. – P. 97 – 120.
5. *Crooks G., Brenner S.* Protein secondary structure: entropy, correlations and prediction // Bioinformatics. – 2004. – V. 20. – P. 1603 – 1611.
6. 6.*Белецкий Б.А., Васильев С.В., Гупал А.М.* Предсказание вторичной структуры белков на основе байесовских процедур распознавания // Проблемы управления и информатики. – 2007. – № 1. – С. 61 – 69.
7. *Xin-Qiu Y., Huaiqiu Z., Zhen-Su S.* A dynamic Bayesian network approach to protein secondary structure prediction // BMC Bioinformatics. – 2008. – V. 9, N 49. – P. 13.
8. *EVA* benchmark for protein structure prediction. http://cubic.bioc.columbia.edu/eva/
9. *Olivier F., Philippe L.* BNT structure learning package: Documentation and Experiments // Technical Report FRE CNRS 2645. – Laboratoire PSI, Universitè et INSA de Rouen
10. *The* GeNIe (Graphical Network Interface) software package. http://genie.sis.pitt.edu/
11. *Балабанов А.С., Гапеев А.С., Гупал А.М., Ржепецкий С.С.* Быстрый алгоритм вывода структур байесовских сетей из данных // Проблемы управления и информатики. – 2011. – № 5. – С. 73 – 80.

Получено 06.03.2014

Об авторах:

Гупал Анатолий Михайлович,
доктор физико-математических наук,
заведующий отделом Института кибернетики имени В.М. Глушкова НАН Украины,

Ржепецкий Сергей Сергеевич,
младший научный сотрудник
Института кибернетики имени В.М. Глушкова НАН Украины.