НЕЙРОСЕТЕВЫЕ МЕТОДЫ ПОИСКА И АНАЛИЗА НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

С.И. Барцев, И.А. Белоус, Ю.П. Ланкин, В.В. Межевикин
Институт биофизики СО РАН, г. Красноярск

В настоящее время накоплено огромное количество информации о геноме человека и других организмов. В частности, завершена программа "Геном человека", приведшая к почти полной расшифровке генетического аппарата homo sapiens. Большие надежды, при этом, связывались с ростом возможностей медицины при лечении наследственных болезней, понимания эволюционных процессов и др. Однако до сих пор не существует достаточно быстрых и эффективных методов анализа и поиска соответствий в нуклеотидных последовательностях, образующих гены и другие части генома. Существующие на данный момент методы основываются, в основном, на "жестких" критериях сравнения схожести фрагментов нуклеотидных цепочек, опирающихся на их совпадение. В частности, для подобных исследований активно используются статистические алгоритмы [1]. Однако в реальных генетических последовательностях встречается множество отклонений от "идеальных" нуклеотидных цепочек (мутаций) благодаря заменам и вставкам отдельных или нескольких нуклеотидов в искомые и анализируемые фрагменты генетических текстов, что сильно затрудняет поиск и анализ.

Из сказанного вполне очевидно, что большое значение имеет разработка "гибких" и быстродействующих методов поиска и анализа генетического кода. Одним из основных претендентов на место таких методов являются *искусственные нейронные сети*, имитирующие фрагменты нейронных сетей мозга. Это направление бурно развивается и прекрасно зарекомендовало себя при решении трудно формализуемых задач [2]. Одним из примеров успешного динамического поиска известных, но сильно варьирующих закономерностей (фонем) в нестационарном квазишумовом сигнале, является успешное решение задачи распознавания речи [3]. Хотя фонемы и являются устойчивыми компонентами речевого потока, они могут сильно отличаться даже в речи одного человека при изменении интонации, настроения и т.д. Следовательно, нейросетевой подход должен обладать необходимой способностью динамического выявления неточно заданных закономерностей в нуклеотидных последовательностях.

Цель настоящей работы – разработать достаточно быстрый и эффективный метод поиска нуклеотидных последовательностей, приближенно сходных с заданным образцом и пригодный для полуавтоматического анализа больших информационных массивов, содержащихся в базах данных.

Алгоритмы нейронных сетей

В данной работе использован вариант алгоритма двойственного функционирования [4] со скользящей оценкой [5]. Алгоритм [4] аналогичен известному на Западе алгоритму обратного распространения ошибки back-propagation [6]. Оба алгоритма получены независимо и опубликованы одновременно.

Следует отметить, что эксперименты проводились на двух классах нейронных сетей – слоистых (статических) и рекуррентных (динамических).

Слоистые сети представляют из себя несколько (обычно 3) слоев формальных нейронов, соединенных связями с весами (синапсами). Результат работы такой сети появляется сразу вслед за подачей сигнала на ее вход.

Рекуррентные сети предназначены для функционирования во времени. При обработке генетических текстов такая нейросеть последовательно "пробегает" по всем нуклеотидам ДНК, реагируя на каждый из них прогнозам последующего нуклеотида. На приведенных ниже рисунках 1 и 2 видно, что обнаружение искомого фрагмента последовательности отражается в виде резкого падения ошибки прогноза в сторону нулевых значений.

Использованный в программе нейросетевой алгоритм [4] прекрасно зарекомендовал себя при решении широкого спектра различных задач. Однако имеются определенные трудности при использовании его для работы на больших временных интервалах времени. В частности, возникают трудности, связанные как с неудобством и затратностью организации процесса, так и затуханием ошибки при ее обратном распространении на интервалах более 10 тактов. Для преодоле-

ния этих проблем в данной работе использована версия этого алгоритма со "скользящей оценкой" [5] с внесением в нее дополнительных модификаций.

С учетом полученного опыта принято решение дальнейшие исследования вести на базе нейронных сетей с самостоятельной адаптацией [7], свободных от выявленных ограничений back-propagation.

Программное обеспечение

Для выполнения описываемых исследований была создана специализированная программа, обеспечивающая весь необходимый уровень сервиса от этапа подготовки данных (подготовка нуклеотидных последовательностей, преобразование их из символьной формы в цифровой код и др.) до вывода на экран результатов вычислений, включая компоненты графического отображения выходных данных численных экспериментов. Элементы графического интерфейса программы приведены на рисунках 1 и 2.

Программа исследований включает два этапа. Первый из них подразумевает работу с тестовыми последовательностями и позволяет оценить потенциал используемых нейронных сетей для решения поставленных задач. Второй – включает исследование реальных нуклеотидных последовательностей.

Для анализа способностей нейросетевой программы к обнаружению заданных фрагментов генетического кода на начальном этапе были взяты последовательности, содержащие случайный набор нуклеотидов. В ходе работы с ними выяснилось, что программа эффективно запоминает тестовые последовательности и справляется с задачей поиска фрагментов, подобных тестовому в любых последовательностях.

Дальнейшие численные эксперименты были выполнены для оценки влияния мутаций на качество распознавания нейронной сетью требуемых фрагментов символьных последовательностей. Как и в первом случае, качество распознавания оказалось высоким.

Результаты исследований

Рисунок 1 – Реакция нейросети на "знакомый" фрагмент

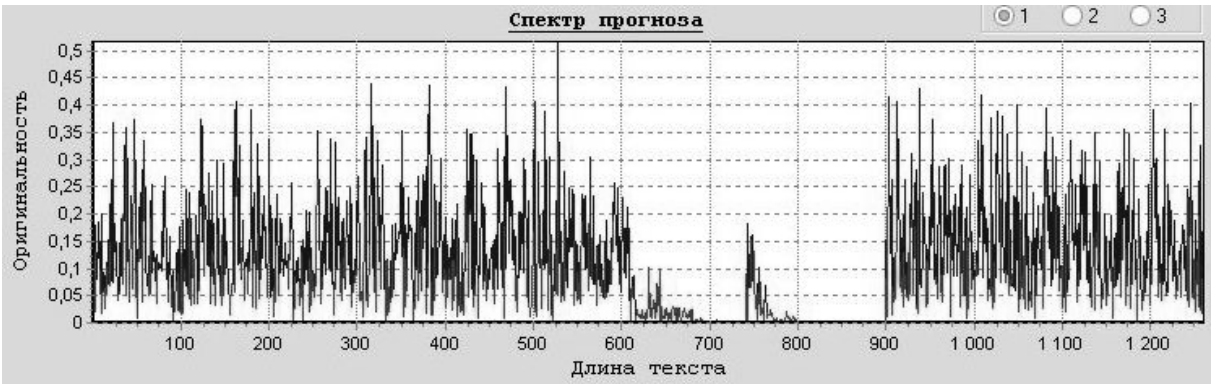

Рисунок 2 – Реакция нейросети на "знакомый" фрагмент с мутацией

Полученные результаты продемонстрировали эффективность использования нейронных сетей не только в рамках проводимых исследований, но и для работы с любыми

другими символьными последовательностями. Так, например, нейронные сети могут быть использованы для поиска слов или предложений в электронных книгах или на сайтах сети Интернет, отличающихся приставками, окончаниями и др.

На втором этапе выполнено тестирование программы на реальных природных последовательностях.

На приводимых рисунках 1 и 2 в качестве тестового образца взята ДНК вируса Vibrio phage fs2. Обучение нейросети проводилось на фрагментах ДНК длиной в 300 нуклеотидов. Реакция обученной нейросети на нуклеотидную последовательность, содержащую искомый фрагмент, приведена на рисунке 1. Время, потребовавшееся на обработку и выдачу результата, равно 0,157 с для компьютера IBM PC с тактовой частотой процессора – 2,1 ГГц.

Во второй серии экспериментов в тестовых последовательностях был заменен ряд нуклеотидов, что соответствует возникновению мутации. Из рисунка 2 видно, что нейросеть продолжает уверенно определять исходный фрагмент, а также указывает на возникшие в нем изменения всплеском на графике спектра прогноза. Эта особенность позволяет локализовать положения мутировавших участков ДНК, что представляет самостоятельную ценность при проведении подобных исследований.

Как указано выше, эксперименты проводились на двух классах нейронных сетей – слоистых (статических) и рекуррентных (динамических). Исследования показали, что для рассматриваемого класса задач рекуррентные сети более эффективны как с точки зрения качества распознавания, так и в плане вычислительных затрат при подготовке данных.

Выводы

Проведенные эксперименты показали эффективность нейросетевых методов при анализе генетического кода и необходимость дальнейших исследований, с целью развития предложенного подхода и апробации его на более широком классе объектов, с целью внедрения в практику и научные исследования.

Выполненные исследования демонстрируют тот факт, что нейросетевые алгоритмы представляют собой гибкий инструмент для поиска требуемых фрагментов в нуклеотидных последовательностях, так как слабо чувствительны к отдельным мутациям в генетическом коде в отличие от большинства традиционно используемых для этой цели алгоритмов. Нейронные сети требуют меньше времени на обработку, так как поиск осуществляется всего за один проход по последовательности.

Важным фактором является также адаптивность нейросетевых алгоритмов, позволяющая приспосабливать их к решению широкого круга задач путем обучения или доучивания.

Данный подход обеспечивает тем большее преимущество во времени и вычислительных ресурсах перед алгоритмическими методами (методы поиска на основе деревьев, хеширование и др.), чем большее количество последовательностей нужно сравнить с исходной. Таким образом, нейросетевые технологии обещают стать высокоэффективным инструментом при работе с базами данных, содержащими большие объемы генетических текстов.

Дополнительным ценным фактором является возможность использования этого аппарата для работы с последовательностями самой различной природы.

Список литературы

1. Садовский М.Г. Информационно–статистический анализ нуклеотидных последовательностей // Диссертация на соискан. уч. степ. д.ф.-м.н. - Красноярск, 2004. - 394 с.
2. Хайкин С. Нейронные сети: полный курс: Пер с англ. - М.: Издат. дом "Вильямс", 2006. - 1104 с.
3. Kohonen T. The "Neural" Phonetic Typewriter // IEEE Computer, March 1988. - P.11-22.
4. Bartsev S.I., Okhonin V.A. Variation principle and algorithm of dual functioning: examples and applications // In proc. of International Workshop "Neurocomputers and attention II", Manchester Univ. Press, 1991. - P. 445-452.
5. Bartsev S.I., Okhonin V.A., Self-learning neural networks playing "Two coins" // In proc. of International Workshop "Neurocomputers and attention II", Manchester Univ. Press, 199. - P. 453-458.
6. Rumelhart D.E., Hinton G.E. & Williams R.J. Learning representations by back-propagating errors // Nature.- 1986.- 323. - P.533-536.
7. Lankin J.P., Baskanova T.F. Algorithms of self-adaptation for atmospheric model designing // SPIE, 2004. - Vol. 5397. - P. 260-270.