Нейросетевые подходы к анализу данных: оценка параметров сложных молекулярных систем по кривым затухания флуоресценции

П.В. Назаров, М.В. Репич, Н.Н. Яцков, В.В. Апанасович

Белорусский Государственный Университет, пр. Скорины 4, Минск 220028, Беларусь

1. Введение

Флуоресцентная спектроскопия с разрешением по времени является важнейшим инструментом при изучении сложных молекулярных объектов и систем, включая биологические мембраны, протеины, ДНК и др. Этот экспериментальный метод позволяет получать детальную информацию о структуре и динамических свойствах молекулярных систем. Ключевым моментом успешного применения флуоресцентной спектроскопии является адекватный анализ кривых затухания флуоресценции. Однако, существует ряд факторов, которые превращают анализ этих данных в достаточно сложную задачу: из набора экспериментальных данных требуется определить не один, а несколько неизвестных параметров системы; практически все зависимости между параметрами и получаемыми данными носят нелинейный характер; экспериментальные данные подвержены искажениям, обусловленным неидеальностью регистрирующей системы флуориметра. Поэтому стандартные подходы к анализу (регрессионные методы, преобразования Фурье и Лапласа и т.д.) зачастую не дают приемлемого результата. Одним из наиболее перспективных методов интерпретации экспериментальных данных спектроскопии является метод фитинга экспериментальных данных результатами имитационного моделирования (simulation-based fitting) [1, 2]. Однако применение этого подхода затруднено на практике из-за значительных вычислительных затрат. Как правило, идентификация системы осуществляется стандартными методами многопараметрической оптимизации, в которых имитационная модель выступает в качестве стохастической функции, аппроксимирующей результаты эксперимента. Очевидно, что в этом случае число пусков моделирования совпадает или превосходит число вычислений функции невязок. Зачастую высокие временные затраты не позволяют применять такой подход для идентификации систем. Поэтому актуальной является разработка методов и алгоритмов, позволяющих снизить вычислительные и временные затраты этого подхода. Также открытым остается вопрос о поиске хороших начальных приближений искомых параметров.

В настоящее время для анализа искаженных данных всё чаще используются искусственные нейронные сети (ИНС), характеризующиеся высокой устойчивостью к шумам, способностью к обучению и обобщению [4, 5]. Нами были предложены два нейросетевых метода анализа кривых затухания. Первый метод предназначен для получения приближенных оценок параметров экспериментальной системы. Второй – является расширением подхода к идентификации систем через имитационное моделирование, значительно снижающим его вычислительные и временные затраты.

2. Принципы флуоресцентной спектроскопии с разрешением по времени.

Флуоресцентная спектроскопия основывается на испускании фотонов из молекулы находящейся в возбужденном состоянии [3]. Наиболее чувствительный метод флуоресцентной спектроскопии – спектроскопия с разрешением по времени [3]. Молекулы возбуждаются очень коротким импульсом света длительностью от пико- до наносекунд. Это позволяет исследовать изменение числа возбужденных молекул во времени. В ансамбле невзаимодействующих молекул статистика деактивации четко определена: если система содержит N типов флуоресцирующих невзаимодействующих молекул, то экспериментально наблюдается многоэкспоненциальное затухание, представленное на рис. 1*a*

$$F(t) = \sum_{I=1}^{N} a_i \exp(-t/\tau_i), \quad \sum_{i=1}^{N} a_i = 1, \tag{1}$$

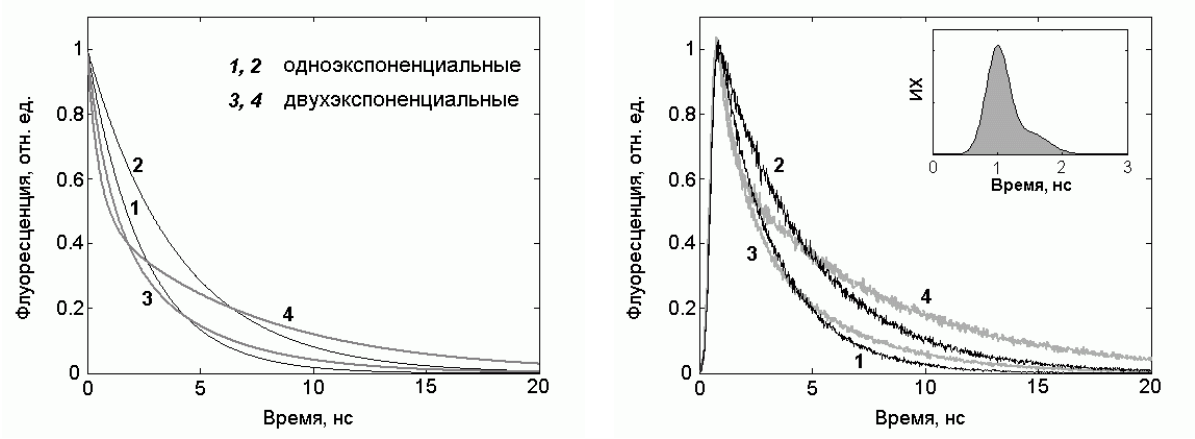
где τ_i – время жизни молекулы *i*-того типа в возбужденном состоянии.

Взаимодействие между флуоресцирующими молекулами посредством тушения или безизлучательного переноса энергии приводит к существенному усложнению закона затухания. В этом случае флуоресценция может быть представлена как стретч-экспонента:

$$F(t) = F_0 e^{-ct} e^{-G(t)}, \tag{2}$$

где F_0, c – некоторые константы, а $G(t)$ – функция, зависящая от геометрии пространственного распределения флуоресцентных молекул. Аналитическое выражение этой функции существует лишь для нескольких упрощенных распределений молекул в одно-, дву- и трехмерных пространствах.

Инструментальная погрешность приводит к дополнительным сложностям в анализе данных. Наблюдаемая флуоресценция в действительности представляет собой свертку между теоретической флуоресценцией (формулы 1-2) и импульсной характеристикой (ИХ) $h(t)$ детектирующей системы, которая обусловлена конечной шириной импульса возбуждения, неидеальностью ИХ фотодетектора и электроники (см. рис 1*б*).



а) *б*)

Рис. 1. Теоретические (а) и получаемые на практике (б) кривые затухания флуоресценции. Параметры затухания: для кривой 1 τ=2.5 нс; для кривой 2 τ=4 ns; 3 – τ_1=4 ns, τ_2=1 ns; 4 – τ_1=7 ns, τ_2=0.5 ns.

3. Нейросетевые подходы к анализу данных

Общий подход к анализу и идентификации сложных экспериментальных систем. Задача определения параметров системы по экспериментальным данным является обратной задачей вычислительного эксперимента и, как следствие, может приводить к неоднозначным решениям. Анализ данных в этом случае требует построения адекватной модели процессов в системе и детального исследования всего пространства параметров. На рис. 2 представлен общий подход к анализу экспериментальных данных, в том числе, полученных методами флуоресцентной спектроскопии для сложных молекулярных систем. Пусть над некоторой системой (блок 1) был поставлен ряд экспериментов, результатом которых явился набор экспериментальных данных (кривых затухания флуоресценции) (блок 2). Исходя из априорной информации о системе и используя результаты предварительного анализа данных (блок 3), выбирается модель физических процессов приводящих к их получению. Кроме того, в блоке 3 могут быть получены оценки искомых параметров системы. С учетом сделанных оценок и выбранной модели производится точная

идентификация параметров. Для сложных систем, у которых отсутствует аналитическое описание, наилучшие результаты при определении параметров дает метод фитинга экспериментальных данных моделью (simulation-based fitting) [1, 2].

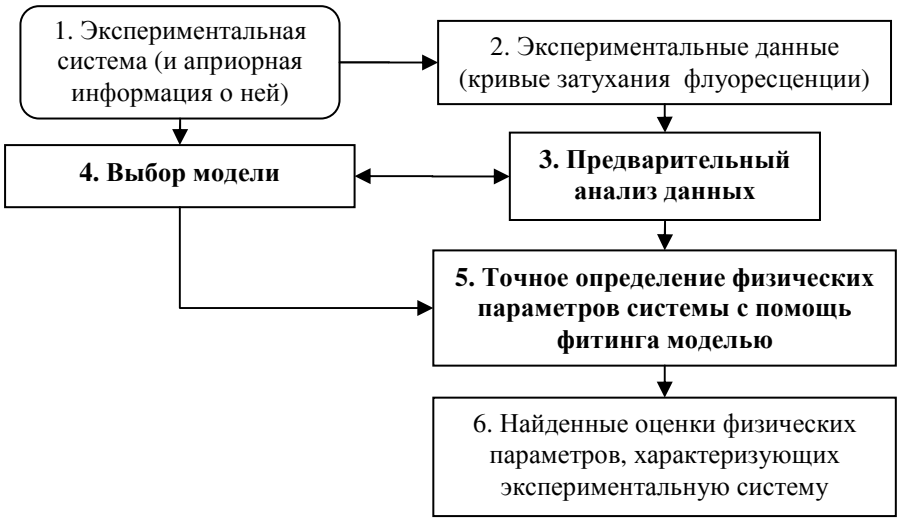
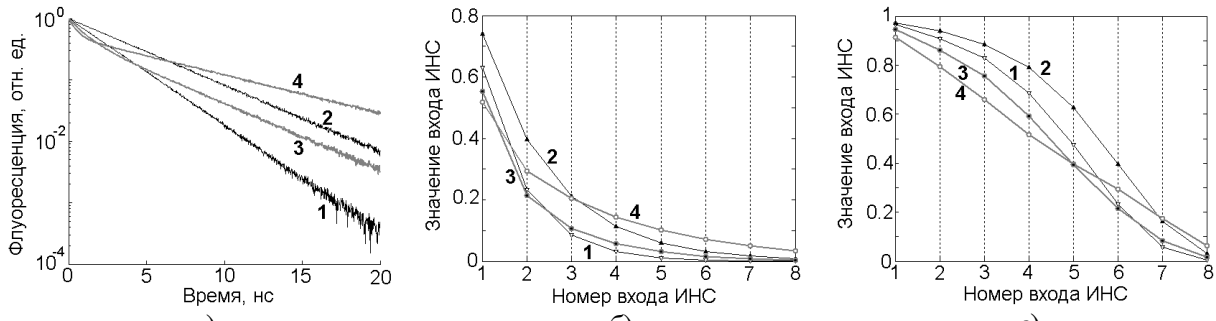


Рис. 2. Общая схема анализа данных и идентификации экспериментальной системы

Применение ИНС для предварительного анализа данных. Наиболее очевидный способ применения ИНС – непосредственное решение обратной задачи по определению параметров системы. При этом на вход нейронной сети подаются предварительно обработанные экспериментальные данные, а с выходов снимается оценка параметров экспериментальной системы.

Для успешного применения ИНС над данными должны быть произведены некоторые дополнительные операции. Во-первых, следует понизить размерность экспериментальных данных. Во-вторых, значения, подаваемые на вход сети, желательно перевести в интервал (0, 1).

Рассмотрим этап предварительной обработки в случае анализа кривых затухания флуоресценции. Исходные данные представлены на рис. 3*а*. Каждая экспериментальная кривая содержит по 1024 отсчета. Для примера пусть число входов ИНС равняется 8. Кривые затухания можно преобразовать в вектор из 8 значений, применив огрубление (усреднение внутри некоторого интервала) по временной шкале. При этом интервалы усреднения могут выбираться либо одного размера (рис. 3*б*), либо экспоненциально возрастающими. На рис. 3*в* приведен пример понижения размерности для количества отсчетов в интервале, равного: 8, 8, 16, 32, 64, 128, 256, 512.



а) *б*) *в*)

Рис. 3. Примеры одно- (кривые 1 и 2) и двухэкспоненциальных (кривые 3 и 4) кривых затухания флуоресценции (а) обработанные с путем усреднения с постоянными (б) и экспоненциально возрастающими (в) интервалами.

Еще одной областью применения ИНС при предварительном анализе данных является выбор модели. Например, при анализе многоэкспоненциальных кривых затухания флуоресценции, нетривиальной является задача определения числа экспонент. Решение этой задачи может быть предоставлено ИНС. При этом на вход сети подаются предварительно обработанные кривые затухания, с выходов снимается вероятность принадлежности кривой к одному из классов (например: одно-, двух- или трёхэкспоненциальные кривые).

Применение ИНС для моделирования системы при ее точной идентификации. Метод идентификации параметров систем с помощью имитационного моделирования был разработан для определения физических и химических параметров сложных систем, которые не могут быть полностью описаны аналитическими выражениями. Суть метода состоит в следующем (рис. 4). Пусть для некоторой экспериментальной системы (блок 1) построена ее имитационная модель (блок 4). При этом физические параметры системы фигурируют в качестве входов модели. Проводя ряд экспериментов над системой, получают набор экспериментальных данных (блок 2). Делается предположение о значениях искомых параметров (блок 3), которые подаются на вход имитационной модели (блок 4). Смоделированные данные (блок 5) сравниваются с экспериментальными. На основании сравнения алгоритм оптимизации (блок 6) корректирует входные параметры имитационной модели. Если ошибка расхождения данных не превышает допустимого значения, алгоритм оптимизации заканчивает свою работу. Входные параметры модели на последнем шаге являются оценкой искомых параметров системы (блок 7).

Рис. 4. Определение параметров системы с помощью имитационной или нейросетевой модели

Метод позволяет оценивать параметры даже очень сложных систем и может работать в случае, когда обратная задача имеет несколько решений (для их нахождения алгоритм следует запускать многократно с различными начальными приближениями параметров). К минусам подхода следует отнести значительные вычислительные затраты, обусловленные моделированием данных на каждой итерации алгоритма оптимизации, а также стохастический характер функции невязки, что не позволяет использовать градиентные методы поиска. Для устранения этих недостатков ранее было предложено заменить имитационную модель (блок 4) на обученную нейронную сеть. На входы сети подаются нормированные оценочные значения искомых параметров. С выхода снимаются смоделированные данные.

Обучение ИНС. Обучение ИНС может проводиться либо непосредственно на экспериментальных данных, либо на модели, если она полностью известна и адекватна. Обученная на экспериментальных данных сеть будет учитывать погрешности эксперимента (методические ошибки в константах, характеристики прибора и т.д.), однако этот подход требует проведения значительного числа экспериментов для получения приемлемого обучающего множества и не всегда применим на практике.

При обучении на смоделированных данных снимаются проблемы связанные с получением обучающего множества. Однако следует искусственно вносить в смоделированные данные шум, соответствующий экспериментальному. Это повышает устойчивость обученной ИНС.

Использованные в работе сети обучались методом обратного распространения ошибки с модификацией Левенберга-Марквардта. Результативность обучения ИНС этим методом в значительной степени зависит от выбора начальных весов сети. Поэтому изначально рассматривались 100 вариантов нейронной сети с различными начальными весами. Каждая предварительно обучалась в течение 100 итераций; затем выбиралась сеть с наименьшей суммарной ошибкой, и дообучалась. Во избежание переобучения сети использовалась идеология контрольного обучающего множества [5].

4. Результаты

Применение ИНС для предварительного анализа данных. Рассмотрим систему флуоресцентных зондов находящихся в одном из трёх конформационных состояний характеризуемых различными временами жизни. Примером такой системы может служить протеин, содержащий аминокислоту триптофан, обладающую флуоресцентными свойствами. Кривая затухания флуоресценции такой системы представляет собой сумму трёх экспонент. Для оценки времен жизни по кривой затухания был использован трёхслойный персептрон с 16 нейронами в каждом скрытом слое и тремя выходами. Предварительно огрубленные кривые затухания подавались на вход сети. С выходов снимались значения времен жизни τ. Для того чтобы избежать многозначности обусловленной инвариантностью суммы к перестановке членов, компоненты кривой затухания были искусственно упорядочены при обучении сети: наибольшее время жизни снималось с первого выхода, наименьшее – с последнего. Сеть обучалась на 2000 смоделированных кривых затухания.

Таблица 1. Вероятности попадания оценки времен жизни в заданный интервал ошибки

Интервал ошибки	Вероятность попадания τ_1 в заданный интервал	Вероятность попадания τ_2 в заданный интервал	Вероятность попадания τ_3 в заданный интервал
< 10%	99%	73%	70%
< 20%	100%	97%	96%

Результаты вычислительного эксперимента, представленные в табл. 1, подтвердили возможность применения ИНС для предварительно анализа кривых затухания. Погрешность оценки позволяет использовать её в качестве хорошего начального приближения при точном анализе данных.

Применение ИНС для моделирования системы при ее точной идентификации. В качестве экспериментальной системы была рассмотрена биологическая мембрана с внедренными протеинами содержащими флуоресцентные метки двух типов: доноры и акцепторы (система детально описана в [1]). Энергия, поглощаемая донорами, с некоторой вероятностью передавалась затем акцепторам. Экспериментально фиксировалось излучение доноров.

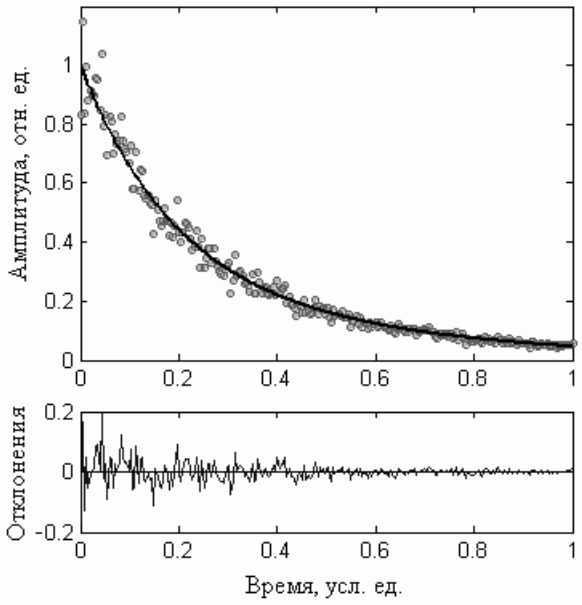


Рис. 5. Нейросетевая аппроксимация (линия) затухания флуоресценции донора (точки) в планарной системе доноров и акцепторов при наличии резонансного переноса энергии

Варьируемыми параметрами в системе являлись концентрации доноров и акцепторов, а также пространственные характеристики системы: коэффициент ассоциации протеинов, эффективный размер молекул, положение акцептора.

На рис. 5 представлено затухание флуоресценции доноров такой системы (точки), его нейросетевая аппроксимация (линия). Время (момент наблюдения) использовалось в качестве дополнительного входного параметра при нейросетевом моделировании. В табл. 2 представлены конфигурации использовавшегося трёхслойного персептрона, а так же временные затраты при обучении и моделировании экспериментальной системы. Оптимальная конфигурация сети устанавливалась эмпирически.

Таблица 2. Временные затраты при нейросетевой аппроксимации модели

Число варьируемых параметров	3	4	5
Число нейронов ИНС: входы – слой1 – слой2 – выходы	3-13-10-1	4-15-13-1	5-18-16-1
Время создания обучающей выборки	11 ч	22 ч	56 ч
Время обучения	6 мин	10 мин	14 мин
Время нейросетевого моделирования	6.0×10^{-4} с	7.0×10^{-4} с	8.0×10^{-4} с
Среднее время имитационного моделирования	40 с	40 с	40 с
Среднее ускорение при моделировании	6.7×10^{4}	5.7×10^{4}	5.0×10^{4}

5. Выводы

В результате исследований было показано, что обученная ИНС может быть использована для предварительного анализа кривых затухания флуоресценции. Метод напрямую работает с экспериментальными данными и, в отличие от классических алгоритмов оценки, не требует операции деконволюции. При анализе трёхэкспоненциальных кривых хорошие оценки параметров были получены для более чем 95% кривых. В то же время, в случае неоднозначности при сопоставлении параметров данным метод выбирает один из вариантов, теряя информацию об остальных. Это заставляет использовать фитинг моделью для точного определения параметров.

Нейросетевая аппроксимация имитационной модели позволяет значительно ускорить процесс моделирования, хотя и требует значительного по времени накопления обучающей выборки. Результатом работы ИНС является гладкая функция, значительно снижающая стохастические девиации результата имитационного моделировании. Предложенный алгоритм анализа данных был применен на практике для идентификации процессов переноса энергии в системах мембранных протеинов. В результате было получено ускорение процесса на этапе анализа более чем в 10^4 раза.

Литература

1. Nazarov, P. V. *et al.* Artificial neural network modification of simulation-based fitting: application to a protein-lipid system, *J. Chem. Inf. Comput. Sci.*, 2004, 44, p. 568-574.
2. Yatskou, M. M. *et al. N*on-isotropic excitation energy transport in organized molecular systems: Monte Carlo simulation-based analysis of time-resolved fluorescence. *J. Phys. Chem. A*, 2001, 105, 9498–9508.
3. Lakowicz, J. R. *Principles of fluorescence spectroscopy*, Kluwer Academic/Plenum Publishers: New York, 1999.
4. Уоссермен Ф. Н*ейрокомпьютерная техника: Теория и практика* Пер. с англ. Ю. А. Зуева и В. А. Точенова. М.: Мир. 1992. 184 с.
5. Tetko I. V., Livingstone D. J., Luik A. I. Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.* 1995, 35. P. 826–833.