ИММУННОСЕТЕВОЕ МОДЕЛИРОВАНИЕ ЗАВИСИМОСТИ «СТРУКТУРА – СВОЙСТВО» ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ

Г.А.Самигулина, С.В.Чебейко

Разработан иммунносетевой подход к моделированию зависимостей «структура-свойство» лекарственных препаратов. Предложенная интеллектуальная технология позволяет уменьшить погрешности энергетических оценок и повысить достоверность прогноза зависимости «структура-свойство» химических соединений.

1. Введение

Актуальнейшим вопросом биоинформатики в настоящее время является целенаправленный поиск новых веществ и материалов с заранее заданными свойствами, в том числе разработка новых лекарственных средств на основе компьютерного молекулярного дизайна. Применение последних достижений вычислительной техники, искусственного интеллекта, инновационных суперкомпьютерных технологий открывает широкие возможности для успешного решения данной проблемы. Вполне реальным становится создание систем по типу САПР (напомним, что в технике это означает «системы автоматического проектирования») для проектирования и конструирования молекулярных устройств с заданными функциями [1].

Разработка новейших интеллектуальных технологий позволит существенно сократить временные и финансовые расходы при производстве новых лекарств. Даже простая предварительная информация о возможных способах посадки предполагаемой молекулы – лекарства на соответствующий рецепторный участок позволяет резко сократить расходы на создание новых фармакологических препаратов.

Время на создание нового лекарства составляет 12-15 лет, а общие затраты денег более 800 млн. долларов [2]. Из 100 тысяч потенциальных лекарственных соединений только 1000 могут стать основой для нового препарата, из которых только 3 смогут выйти на фармакологический рынок. С помощью компьютерного молекулярного дизайна, интеллектуальных технологий возможно создание принципиально новых компьютерных алгоритмов, программ поиска и отбора активных веществ целевого назначения на основе концепции взаимосвязи молекулярной структуры и биологической активности химических соединений. Процесс создания нового лекарства включает в себя поиск мишени действия нового препарата; поиск биологически активного вещества, обладающего нужным фармакологическим действием; изучение этого соединения и т.д.

В том случае, когда неизвестна структура мишени, а есть только информация, что у каких-то веществ есть нужная активность, обычно используют метод QSAR (Quantitative Structure-Activity Relationship) - количественное соотношение структура-свойство. Это на-

правление возникло на стыке органической химии, математического моделирования и компьютерной химии.

Количественное описание молекулярной структуры химических соединений в компьютерном молекулярном дизайне осуществляется с помощью дескрипторов [3]. Дескриптор - это математический параметр, который характеризует структуру органического соединения, отмечая наиболее важные черты этой структуры. Существует проблема создания дескрипторов наиболее полно характеризующих рассматриваемое соединение и позволяющих в удобной форме использовать их в вычислительном процессе.

В современном компьютерном дизайне выделяется три основных этапа исследований:

- формирование обучающей выборки соединений с заданным свойством (активностью);
- описание молекулярной структуры исследуемых соединений (дескрипторов);
- установление взаимосвязи «структура – свойство (биологическая активность)» с последующим созданием устойчивых прогностических математических моделей.

Среди методов прогнозирования зависимости «структура – свойство» следует отметить рост исследований по искусственным нейронным сетям [4]. Наиболее популярна многослойная нейронная сеть прямого распространения, обучающаяся по методу обратного распространения ошибки.

2. Подход искусственных иммунных систем

Моделирование биологической активности органических соединений также возможно с помощью нового биологического направления искусственного интеллекта – искусственных иммунных систем [5] (ИИС). Под ИИС понимаются информационные методологии, использующие понятия теоретической иммунологии для решения различных прикладных задач. Интерес к таким задачам объясняется их сложной структурой, неполнотой и зашумленностью данных, невозможностью успешного использования классических математических подходов. ИИС обладают такими свойствами как способность к обучению, к прогнозированию и принятию решения в незнакомой ситуации. Биологическим прототипом данного подхода является иммунная система человека и принципы обработки информации молекулами белков на основе результатов самосборки. Рассматриваются взаимодействия между белками иммунной системы человека и чужеродными антигенами, процедура молекулярного узнавания посредством определения минимальной энергии связи между формальными пептидами.

Наиболее важными при обработке информации молекулами белка являются: пространственная конфигурация белков, принципы самосборки белков, их комплексов и сетей. Принципы действия механизмов репарации, то есть исправления ошибок в процессе функционирования ИИС позволяют строить совершенно новые алгоритмы распознавания

образов на основе искусственных иммунных сетей. Достоинствами ИИС являются: распределенность, самоорганизация, небольшая требовательность к вычислительным ресурсам, отсутствие централизованного контроля, самообучаемость, индивидуальный подход к уникальным событиям.

Математическая основа подхода ИИС заключается во введении понятия формального пептида как математической абстракции свободной энергии белковой молекулы от ее пространственной формы, описанной в алгебре кватернионов. Формальным пептидом называют упорядоченную пятерку [5]:

$$P = < n, U, Q, V, v >,$$

которая включает следующие компоненты:

- количество звеньев $n>0$;

- множество торсионных углов: $U = \{ \varphi_k, \psi_k \}, \ k=1,...,n,$

где $-\pi \le \varphi_k \le \pi, \ -\pi \le \psi_k \le \pi$;

- множество единичных кватернионов: $Q = \{Q_0, Q_k\},$

где кватернионы $Q_k = Q_k(\varphi_k, \psi_k)$ и результирующий кватернион ФП Q_0 определяется как их произведение: $Q_0 = Q_1 Q_2 ... Q_n$;

- множество коэффициентов $V = \{v_{ij}\}, \ i = 1,2,3,4, \ j \ge i$;

- функция v (без индекса), определенная на элементах результирующего кватерниона Q_0 следующей квадратичной формой: $v = -\sum_{j \ge i} v_{ij} q_i q_j$.

3. Постановка задачи

Постановка задачи формулируется следующим образом: необходимо разработать эффективную интеллектуальную информационную технологию моделирования и предсказания свойств новых лекарственных препаратов с заданными свойствами на основе биологического подхода искусственных иммунных систем [6, 7].

Специфика состоит в том, что при построении математических моделей необходимо одновременно использовать как распознающие, так и оптимизационные способности иммунной сети [8]. Задачу прогнозирования можно рассматривать как задачу оптимизации: необходимо построить оптимальную модель (структуру), которая бы наиболее точно описывала требуемую биологическую активность (свойство) органического соединения.

Реальные данные (дескрипторы) почти всегда содержат нежелательные составляющие, которые называются шумом. Существуют различные виды неопределенностей данных, например случайные погрешности (неточности при сборе и формировании дескрипторов); систематические (причинные) погрешности; ошибки при моделировании из-за использования упрощенных моделей и т.д. Шум и избыточность данных проявляются в корреляции между

переменными. За исключением простых случаев, искажения в данных не могут быть устранены полностью. Основной целью при построении систем на основе иммунносетевого моделирования является уменьшение ошибки обобщения, поскольку малая ошибка обучения гарантирует адекватность модели лишь в заранее выбранных точках. Важным является способность иммунной сети обобщать результат на новые данные, которые не были использованы в обучающемся множестве.

Возможна ситуация когда почти идентичные по структуре химические соединения имеют диаметрально противоположные свойства и следовательно биологическую активность, что совершенно недопустимо при конструировании лекарственных соединений. В качестве примера можно привести парадокс похожести. Оптически активное лекарство и его зеркальный изомер могут значительно различаться по биологической активности [1, 7]. Некоторые хиральные барбитураты в одной из форм обладают седативной активностью, тогда как другой их энантиомер вызывает судороги. В случае синтетических аналогов морфина один энантиомер может быть сильным анальгетиком, а другой обладать противокашлевым эффектом.

Таким образом, решение задачи минимизации ошибки обобщения позволяет повысить прогностическую способность модели и является наиболее трудной при построении данных систем.

4. Алгоритм иммунносетевого моделирования

Используется следующий алгоритм, который состоит из 9 шагов:

Шаг 1.Описываются структуры исследуемых соединений числовыми параметрами (дескрипторами), создаются базы данных (БД);

Шаг 2. Осуществляется предварительная обработка дескрипторов: нормирование, центрирование, заполнение пропусков;

Шаг 3. Выбирается оптимальный набор дескрипторов, строится оптимальная структура иммунной сети;

Шаг 4. Весь массив данных разбивается на обучающую и контролирующую выборки;

*Шаг 5. Э*кспертами осуществляется классификация решений;

Шаг 6. Производится обучение иммунной сети с учителем;

Шаг 7. Решается задача распознавания образов и нахождения минимальной энергии связывания между формальными пептидами (антителами и антигенами);

Шаг 8. Осуществляется оценка решения задачи распознавания образов на основе гомологов и расчет коэффициентов риска прогнозирования на основе ИИС;

*Шаг 9. О*существляется прогноз свойств неизвестных соединений.

Рассмотрим подробнее реализацию данного алгоритма.

4.1. Предварительная обработка данных

Методы предварительной обработки многомерных данных предназначены для перевода данных в наиболее информативное для исследуемой задачи представление.

Рассматривается предварительная обработка данных двух видов. Прежде всего, это работа с базами данных, так как входной информацией для ИИС служат дескрипторы, характеризующие химическое соединение, которые занесены в базу данных в виде временных рядов. Одной из обязательных задач является заполнение недостающих данных в БД. Существует множество причин, по которым образуются пропуски в данных. Например, неисправность измерительных приборов, невозможность произвести исследование и т.д.

Можно использовать различные методы: заполнения средним; заполнения главным компонентом; метод случайного заполнения или применить подход, основанный на max правдоподобия и т.д. Выбор конкретного метода зависит от постановки задачи исследования и структуры данных.

Пусть исходная совокупность данных записана в виде матрицы $A=(a_{ij})$ $(i, j=1,....n)$ размерности $(m \times n)$. Так как дескрипторы, характеризующие вещества, измеряются в разных единицах, то результат может существенно зависеть от выбора масштаба измерения. Поэтому необходим переход к безразмерным величинам с помощью нормирования и центрирования дескрипторов. Для этого элементы каждого вектора преобразуем таким образом, чтобы математическое ожидание было равно нулю, а дисперсия единице.

Основной целью нормирования данных является приведение их к сопоставимому виду. Новая матрица стандартизированных переменных X записывается из элементов:

$$x_{ij}' = \frac{x_{ij} - m_j}{s_j},$$

где m_j – среднее значение исходных элементов j-го вектора; s_j – стандартное отклонение исходных элементов j- го вектора, которое вычисляется по формуле:

$$s_j = (\frac{1}{N-1}\sum_{i=1}^{n}(x_{ij} - m_j)^2)^{\frac{1}{2}}.$$

4.2. Построение оптимальной структуры иммунной сети

Способность ИИС обрабатывать большой объем информации неизбежно приводит к увеличению времени на обучение. В связи с данной проблемой используются различные методы предварительной обработки исходных данных для оптимального отбора переменных, выделения основополагающих факторов, уменьшения анализируемого пространства признаков и снижения времени на обучение.

Задача снижения размерности анализируемого признакового пространства и отбора наиболее информативных дескрипторов решена с помощью факторного анализа и метода главных компонент на основе вращения собственного вектора [10].

Под оптимальной структурой иммунной сети будем понимать структуру сети, построенной на основе весовых коэффициентов выделенных информативных дескрипторов, наиболее полно характеризующих состояние рассматриваемого вещества в зависимости от основополагающих факторов, влияющих на его свойства. Критерием является максимальное сохранение информации при минимальном количестве дескрипторов.

Основные задачи факторного анализа:

- исследование структуры взаимосвязей переменных (дескрипторов). В этом случае каждая группировка переменных будет определяться фактором, по которому эти переменные имеют максимальные нагрузки;

- идентификация факторов как скрытых (латентных) переменных - причин взаимосвязи исходных переменных;

- сокращения количества дескрипторов с минимальными потерями исходной информации.

Ниже предложен алгоритм построения оптимальной структуры ИИС.

Алгоритм

Шаг 1. Нормирование исходных дескрипторов таким образом, чтобы математическое ожидание было равно нулю, а дисперсия единице.

Шаг 2. Выделение информативных дескрипторов с помощью метода главных компонент.

Шаг 3. Выделение и анализ основополагающих факторов на основе развития варимаксного подхода.

Шаг 4. Редукция малоинформативных дескрипторов.

Шаг 5. Визуальное представление многомерных данных на дисплее.

Шаг 6. Ранжирование информативных дескрипторов в зависимости от весовых коэффициентов.

Шаг 7. Построение оптимальной структуры иммунной сети для дальнейшего решения задачи распознавания образов на основе ИИС.

Определим базисное пространство R и проекции векторов данных на каждую из n-ортогональных осей. Тогда исходную матрицу данных A размерности $(m \times n)$ можно представить в матричной форме:

$$A = CV^{\mathrm{T}},$$

где V – матрица, столбцы которой ортогональны оси; C – матрица, строками которой являются координаты проекций каждого вектора данных в базисном пространстве R. Тогда координаты новой матрицы B будут записаны в матричной форме:

$$B = R^{\mathrm{T}} A.$$

Матрица преобразования R^{T} в двумерном пространстве имеет вид:

$$R^{T} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

Рассчитывается корреляционная матрица:

$$C = \frac{1}{N-1}(X^{T} X),$$

где $N-$ число столбцов в матрице X.

Пусть $Y = B^{\mathrm{T}}$, $X = A^{\mathrm{T}}$, тогда получим:

$$Y = XR, \qquad Y^{\mathrm{T}} = R^{\mathrm{T}} X^{\mathrm{T}}.$$

Необходимо найти матрицу преобразования R^{T} такую, чтобы, применив ее к матрице X, получить новую систему координат Y, которая удовлетворяет выражению:

$$Y^{\mathrm{T}} Y = R^{\mathrm{T}} X^{\mathrm{T}} XR = R^{\mathrm{T}} CR = \Lambda,$$

где Λ – диагональная матрица.

Необходимо, чтобы выполнялось условие:

$$CR = \lambda R,$$

тогда получим:

$$(C - \lambda I)R = 0,$$

где λ – скалярные, диагональные элементы в матрице Λ.

Задача будет иметь решение при выполнении:

$$|C - \lambda I| = 0.$$

После нахождения решения для λ, подставим их вновь в (1) и найдем матрицу преобразования R. На основе проведенных преобразований исходные данные можно изобразить в новой системе, где координатные оси являются собственными векторами. После анализа дескрипторов необходимо отбросить те, которые лежат ближе к началу координат и являются наименее информативными.

4.3. Решение задачи распознавания образов

При реализации интеллектуальной технологии используется следующий алгоритм для решения задачи распознавания образов [5].

Алгоритм

Шаг 1. Для каждого класса, выделенного экспертами, формируются матрицы эталонов:

$$A_1,\ A_2,\ A_3,\ldots n,$$ где n - количество классов.

Шаг 2. Для реализации шага 1. информация берется из сформированных баз данных и баз знаний. Для улучшения специфичности узнавания каждый выбранный временной ряд сворачивается в квадратную матрицу.

Шаг 3. После сингулярного разложения данных матриц получаем правые и левые сингулярные вектора:

$$\{x_1, y_1\},\ \{x_2, y_2\}$$ и т.д., эталонных матриц.

Шаг 4. Формируется множество матриц, рассматриваемых как образы:

$$B_1,\ B_2,\ B_3,\ B_4,\ldots.., m,$$

где m - количество образов.

Шаг 5. Расчитываются энергии связи между формальными пептидами, которые представлены в виде:

$$W_1 = -x_1^T B y_1,$$

$$W_2 = -x_2^T B y_2,$$

$$W_3 = -x_3^T B y_3,$$

$$W_4 = -x_4^T B y_4,$$

…….

$$W_n = -x_n^T B y_n,$$

где T – символ транспонирования, n - количество классов.

Шаг 6. Определяется минимальное значение энергии связи. Нативная (функциональная) укладка белковой цепи соответствует минимуму энергии связи, поэтому минимальное значение энергии связи определяет класс n, к которому принадлежит данный образ:

$$n : W_n = \min\{W_1, W_2, W_3, W_4, ..., W_n\}$$

Шаг 7. Результаты распознавания образов заносятся в таблицу.

Основная проблема при решении задачи распознавания образов связана с данными, которые находятся на границе разделения классов (особенно при нелинейном разделении классов). Пептиды, почти похожие по структуре могут быть отнесены к неправильным классам, что существенно уменьшает достоверность прогноза.

4.4. Оценки энергетических погрешностей ИИС

Обработка многомерной совокупности данных на основе технологии ИИС неизбежно приводит к увеличению энергетических погрешностей, зависящих от ряда факторов, и существенно влияет на достоверность прогноза. Определение нативной укладки цепи, соответствующей минимуму энергии, затруднено вследствие наличия различных погрешностей. Актуально применение свойств реальных белков для решения этой проблемы на основе ИИС. Главной характеристикой аминокислотных последовательностей, определяющей основные физические свойства белка, является повышенная стабильность нативной структуры. Устранение погрешностей энергетических оценок в подходе ИИС осуществляется на основе свойств гомологичных белков. Гомологичными формальными пептидами будем называть пептиды, которые при решении задачи распознавания образов относятся к одному классу решений.

Вычисленная энергия нативной структуры E'_N представляется в виде:

$$E'_N = E_N + \Delta E_N ,$$

где E_N – истинное значение энергии нативной структуры; ΔE_N – погрешность энергетической оценки. Приведем алгоритм оценки энергетических погрешностей ИИС [11].

Алгоритм

Шаг 1. Осреднение потенциалов по гомологам:

$$< E_i^* >_G = \frac{\sum_i^G E_i^*}{G},$$

где символ <…> означает осреднение по гомологам; G – число гомологичных пептидов.

Шаг 2. Определение вычисленной энергии нативной (функциональной) структуры по усредненным гомологам:

$$< E_i^* >_G \approx E_N .$$

Шаг 3. Определение *Z*-факторов. Значение *Z*-фактора определяется средним числом стандартных отклонений между энергией нативной структуры и энергией случайно выбранной укладки цепи:

$$Z = \frac{E_N - < E >}{< (E - < E >)^2 >^{\frac{1}{2}}},$$

где $E_N \approx < E_i^* >_G$; $< E >$ – среднее число стандартных отклонений по гомологам; E – энергия случайно выбранной укладки цепи.

Шаг 4. Распознавание нативной структуры белков по гомологам и определение достоверности прогноза на основе ИИС в зависимости от значений *Z*-фактора. Рассчитываются коэффициенты риска прогнозирования:

$$K_R(G_i) = |1 - Z_i|, \; i = \overline{1,n},$$

где n – количество гомологичных пептидов.

Таким образом, нативная структура белковой цепи, соответствующая минимуму энергии связи, является для каждого класса определенной и позволяет определить принадлежность гомологов к какому-либо классу решений. Особенно это свойство ценно для образов, которые находятся на границах классов. Данная способность ИИС существенно уменьшает погрешности энергетических оценок, повышает достоверность прогноза интеллектуальной системы.

Актуально создание определенных стандартных модулей технологических последовательностей, которые могут быть использованы при разработке программного обеспечения для иммунносетевого моделирования лекарственных препаратов с заданными свойствами. Параллельное вычисление возможных технологических цепочек позволяет определить оптимальную технологию обработки данных иммунной сетью с наименьшей ошибкой обобщения и самыми высокими прогностическими свойствами.

5. Заключение

Достоинством предложенной интеллектуальной технологии на основе иммунносетевого моделирования является: способность системы глубоко анализировать скрытые (латентные) взаимодействия между дескрипторами и основополагающие факторы, влияющие на них; распознавать пептиды, находящиеся на границе нелинейно разделенных классов (имеющие схожие структуры); сокращение времени на обучение иммунной сети за счет построения оптимальной структуры и редукции дескрипторов, несущих существенные погрешности; уменьшение погрешностей энергетических оценок, так называемых ошибок обобщения; повышение достоверности прогноза зависимостей «структура – свойство» химических соединений.

На разработанное программное обеспечение получено авторское свидетельство о государственной регистрации объекта интеллектуальной собственности [12].

Литература

1. Шайтан К.В. Молекулярная динамика пептидов. http://www.bioeng.ru/doc/stat/moldyn.htm

2. Кубиньи Г. В поисках новых соединений-лидеров для создания лекарств //Российский химический журнал. 2006, № 2. С. 5-17.

3. Раевский О.А. Дескрипторы водородной связи в компьютерном молекулярном дизайне //Российский химический журнал. 2006, № 2. С. 97-108.

4. Гальберштам Н.М., Баскин И.И., Палюлин В.А., Зефиров Н.С. Нейронные сети как метод поиска зависимостей структура – свойство органических соединений //Успехи химии. 2003, № 72(7). С. 706-727.

5. Тараканов А.О. Математические модели бимолекулярной обработки информации: формальный пептид вместо формального нейрона // Проблемы информатизации. 1998. С.65-70.

6. Самигулина Г.А., Чебейко С.В. Прогнозирование зависимости структура-свойство органических соединений на основе иммунносетевого моделирования // Химический журнал Казахстана. –Алматы. 2010, № 3. С. 164-172.

7. Самигулина Г.А., Чебейко С.В. Технология иммунносетевого моделирования для компьютерного молекулярного дизайна лекарственных препаратов // Вестник Харьковского Университета. Тематический выпуск: информатика и моделирование. – Харьков. 2011, № 17. С.142-148. http://www.nbuv.gov.ua/Portal/natural/vcpi/IiM/2011_17/20pdf

8. Бидюк П.И., Литвиненко В.И., Фефелов А.А., Баклан И.В. Гибридная иммунная сеть для решения задач структурной идентификации//Искусственный интеллект. 2004, № 3. С.89-99.

9. http://www.bioinformatix.ru/bioinformatica/drag_dizayn.html.

10. Иберла К. Факторный анализ. –М.:Статистика, 1980.

11. Самигулина Г.А. Разработка интеллектуальных экспертных систем прогнозирования и управления на основе искусственных иммунных систем // Проблемы информатики. Новосибирск, 2010, № 1. С. 15-22.

12. Самигулина Г.А., Самигулина З.И. Разработка технологии иммунносетевого моделирования для компьютерного молекулярного дизайна лекарственных препаратов (программа для ЭВМ). Свидетельство о государственной регистрации прав на объект авторского права в Комитете по правам интеллектуальной собственности МЮ РК. Астана, 2011, №473.