УДК 004.89:004.41.

Г.А. САМИГУЛИНА, д.т.н., зав. лаб. Института проблем информатики и управления Министерства образования и науки Республики Казахстан, Алматы,

С.В. ЧЕБЕЙКО, к.х.н., с.н.с. Института проблем информатики и управления Министерства образования и науки Республики Казахстан, Алматы

РАЗРАБОТКА ТЕХНОЛОГИИ ИММУННОСЕТЕВОГО МОДЕЛИРОВАНИЯ ДЛЯ КОМПЬЮТЕРНОГО МОЛЕКУЛЯРНОГО ДИЗАЙНА ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ

Разработан иммунносетевой подход к моделированию зависимостей "структура-свойство" лекарственных препаратов. Предложенная интеллектуальная технология на основе искусственных иммунных систем позволяет уменьшить погрешности энергетических оценок и повысить достоверность прогноза зависимости "структура-свойство" химических соединений. Библиограф.: 12 назв.

Ключевые слова: технология иммунносетевого моделирования, интеллектуальная технология, погрешности энергетических оценок.

Постановка проблемы и анализ литературы. Создание методов прогнозирования свойств новых химических соединений и направленный компьютерный молекулярный дизайн соединений с заданным набором свойств являются важнейшими и актуальными задачами биоинформатики. Применение последних достижений вычислительной техники и новейших информационных технологий открывает широкие возможности для решения одной из главных проблем современной науки – целенаправленного поиска новых веществ и материалов с заранее заданными свойствами, в том числе проектирование новых лекарственных средств.

История дизайна с помощью компьютеров началась более 25 лет назад, когда стало возможным изображение и вращение молекул на экране компьютера [2]. Компьютерный молекулярный дизайн основан на концепции взаимосвязи молекулярной структуры и биологической активности химических соединений. Данное направление предполагает создание принципиально новых компьютерных алгоритмов и программ поиска и отбора активных веществ целевого назначения.

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. Построение адекватной компьютерной молекулярной модели позволяет в дальнейшем прогнозировать различные терапевтические и физико-химические свойства синтезируемых молекул, что определяет актуальность и перспективность развития данного научного направления.

Среди методов прогнозирования зависимости "структура – свойство" следует отметить рост исследований по искусственным нейронным сетям [4]. В рамках поиска зависимостей между структурами органических соединений и их биологической активностью наиболее популярна многослойная нейронная сеть прямого распространения, обучающаяся по методу обратного распространения ошибки.

Моделирование биологической активности органических соединений также возможно с помощью нового биологического направления искусственного интеллекта – искусственных иммунных систем (ИИС).

Процессы, происходящие при обработке информации естественными системами и принципы их функционирования, поражают своей эффективностью, экономичностью и быстродействием [5, 6]. Прежде всего, вызывает интерес способность данных систем решать многомерные задачи огромной вычислительной сложности в реальном времени. ИИС – это адаптивные системы для обработки и анализа данных, которые представляют собой математическую структуру, имитирующую некоторые функции иммунной системы человека и обладающие способностью к обучению, к прогнозированию на основе уже имеющихся временных рядов и принятию решения в незнакомой ситуации. ИИС в принципе не нуждаются в заранее известной модели, а строят ее сами на основе полученной информации в виде временных рядов. Данные системы применяются при решении плохо алгоритмизируемых задач, таких как прогнозирование, классификация и управление.

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. В работах [7, 8] предложена математическая модель формального пептида.

При реализации интеллектуальных систем, основанных на выше приведенных принципах, существует ряд проблем [9, 10]. Основная трудность заключается в создании алгоритмов безошибочного

распознавания образов, так как ошибки энергетических оценок не позволяют добиться сто процентного распознавания. Особенно эта проблема актуальна для схожих по структуре формальных пептидов, которые находятся на границах различных классов и разделение между классами нелинейно. Как и в искусственных нейронных сетях, существует проблема создания эффективных и простых методик обучения иммунной сети за минимально короткое время. Необходимо из множества факторов выделить главные, которые оказывают наибольшее влияние на процесс обработки информации, построить оптимальную структуру иммунной сети на основе информативных дескрипторов, обучать ИИС и оценить процесс обучения. Проблема значительно усложняется при увеличении размерности системы.

Кроме того, очень важным является способность иммунной сети обобщать результат на новые данные, которые не были использованы в обучающем множестве. Таким образом, решение задачи минимизации ошибки обобщения позволяет повысить прогностическую способность модели и является наиболее трудной при построении данных систем.

Цель статьи. Разработать эффективную интеллектуальную информационную технологию для компьютерного молекулярного дизайна (моделирования и предсказания свойств новых лекарственных препаратов с заданными параметрами) на основе биологического подхода искусственных иммунных систем.

Технология иммунносетевого моделирования. Разработана интеллектуальная информационная технология, позволяющая моделировать зависимость "структура – свойство" на основе искусственных иммунных сетей [11].

Используется следующий алгоритм:

– описываются структуры исследуемых соединений числовыми параметрами (дескрипторами), создаются базы данных (БД);

– осуществляется предварительная обработка данных: нормирование, центрирование, заполнение пропущенных данных;

– выбирается оптимальный набор дескрипторов, строится оптимальная структура иммунной сети;

– весь массив данных разбивается на обучающую и контролирующую выборки;

– экспертами осуществляется классификация решений;

– производится обучение иммунной сети с учителем;

– решается задача распознавания образов и нахождения минимальной энергии связывания между формальными пептидами (антителами и антигенами);

– осуществляется оценка решения задачи распознавания образов на основе гомологов и расчет коэффициентов риска прогнозирования на основе ИИС;

– осуществляется прогноз свойств неизвестных соединений.

Рассмотрим подробнее реализацию данного алгоритма. Разработанная интеллектуальная технология состоит из трех основных этапов:

Этап1. Предварительная обработка данных

Пусть исходная совокупность данных записана в виде матрицы $A = (a_{ij})$ $(i = 1, \ldots, m,\ j = 1, \ldots, n)$. 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: $x'_{ij} = \dfrac{x_{ij} - m_j}{s_j}$ где m_j – среднее значение исходных элементов *j*-го вектора; s_j – стандартное отклонение исходных элементов j -го вектора, которое вычисляется по формуле:

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

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 [12].

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 виде:

$$A = CV^T,$$

где V – матрица, столбцы которой ортогональны оси; C – матрица, строками которой являются координаты проекций каждого вектора данных в базисном пространстве R. Новую матрицу B получим следующим образом:

$$B = R^T A.$$

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

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

Рассчитывается корреляционная матрица:

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

где $N-$ число столбцов в матрице X.

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

Необходимо найти матрицу преобразования R^{T} такую, чтобы, применив ее к матрице X, получить новую матрицу 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:

$$(C - \lambda I)R = 0, \qquad (1)$$

где λ – вектор диагональных элементов в матрице Λ.

Задача будет иметь решение при выполнении условия:

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

После нахождения вектора λ подставим его 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 (для построения оптимальной структуры иммунной сети) необходимо отбросить те, которые лежат ближе к началу координат и являются наименее информативными.

Этап 2. Распознавание образов

Ключевым моментом в разработанной интеллектуальной технологии на основе ИИС является решение задачи распознавания образов [8]. Для каждого класса, выделенного экспертами, формируются матрицы эталонов $A_1, A_2, A_3, \ldots, A_n$ (n – количество классов). Выполнив

сингулярное разложение данных матриц, получаем правые и левые сингулярные векторы $\{x_1, y_1\}$, $\{x_2, y_2\}$ и т. д. эталонных матриц. Затем формируется множество матриц, рассматриваемых в качестве образов: B_1, B_2, B_3, …, B_m (m – количество образов).

Согласно подходу ИИС энергию связи между формальными пептидами можно представить в виде:

$$W_1 = -x_1^{\mathrm{T}} B y_1,\ W_2 = -x_2^{\mathrm{T}} B y_2,\ W_3 = -x_3^{\mathrm{T}} B y_3,\ \ldots,\ W_n = -x_n^{\mathrm{T}} B y_n,$$

где т – символ транспонирования.

Нативная (функциональная) укладка белковой цепи соответствует минимуму энергии связи, поэтому минимальное значение энергии связи определяет класс n, которому принадлежит данный образ: $W_n = \min\{W_1, W_2, W_3, \ldots, W_n\}$.

Этап 3. Оценка энергетических погрешностей

Обработка многомерной совокупности данных на основе технологии ИИС неизбежно приводит к увеличению энергетических погрешностей, зависящих от ряда факторов, и существенно влияет на достоверность прогноза. Разработана процедура оценки энергетических погрешностей на основе гомологичных белков [10].

Выводы. 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 получены два авторских свидетельства о государственной регистрации объекта интеллектуальной собственности.

Список литературы: 1. *Кубиньи Г.* В поисках новых соединений – лидеров для создания лекарств / *Г. Кубиньи* // Российский химический журнал. – 2006. – № 2. – С. 5-17. **2.** *Иванов А.С.* Интегральная платформа "От гена до прототипа лекарства" in silico и in vitro / *А.С. Иванов, А.В. Веселовский, А.В. Дубанов, В.С. Скворцов, А.И. Арчаков* // Российский химический журнал, 2006. – № 2. – С. 18-35. **3.** *Раевский О.А.* Дескрипторы водородной

связи в компьютерном молекулярном дизайне / *О.А. Раевский* // Российский химический журнал, 2006. – № 2. – С. 97-108. **4.** *Гальберштам Н.М.* Нейронные сети как метод поиска зависимостей структура – свойство органических соединений / *Н.М. Гальберштам, И.И. Баскин, В.А. Палюлин, Н.С. Зефиров* // Успехи химии, 2003. – № 72 (7). – С. 706-727. **5.** *Альбертс Б.* Молекулярная биология клетки / *Б. Альбертс, Д. Брей, Дж. Льюи.с* – М.: Мир, 1994. – Т. 2. – С. 287-301. **6.** Искусственные иммунные системы и их применение / *Под ред. Д. Дасгупт.* – М.: Физматлит, 2006. – 344 с. **7.** *Тараканов А.О.* Математические модели биомолерулярной обработки информации: формальный пептид вместо формального нейрона / *А.О. Тараканов* // Проблемы информатизации. – 1998. – С. 65-70. **8.** *Tarakanov A.O.* Formal peptide as a basic of agent of immune networks: from natural prototype to mathematical theory and applications / *A.O. Tarakanov* // Proceedings of the I Int. workshop of central and Eastern Europe on Multi-Agent Systems (CEEMAS'99). – St. Petersburg, Russia, June 1-4, 1999. – P.281-292. **9.** *Самигулина Г.А.* Разработка интеллектуальных экспертных систем управления на основе искусственных иммунных систем / *Г.А. Самигулина*. – Алматы: ИПИУ МОН РК, 2010. – 252 с. **10.** *Самигулиной Г.А.* Разработка интеллектуальных экспертных систем прогнозирования и управления на основе искусственных иммунных систем / *Г.А. Самигулиной* // Проблемы информатики. – Новосибирск, 2010. – № 1. – С. 15-22. **11.** *Самигулина Г.А.* Прогнозирование зависимости структура-свойство органических соединений на основе иммунносетевого моделирования / *Г.А.. Самигулина, С.В. Чебейко* // Химический журнал Казахстана. – Алматы, 2010. – № 3. – С. 164-172. **12.** *Иберла К.* Факторный анализ / *К. Иберла*. – М.: Статистика, 1980. – 304 с.

УДК 004.89:004.41.

Розробка технології иммунносетевого моделювання для комп'ютерного молекулярного дизайну лікарських препаратів / Самігуліна Г.А., Чебейко С.В. // Вісник НТУ "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ". – 2011. – № 17. – С. 142 – 148.

Розроблений иммунносетевой підхід до моделювання залежностей "структура-властивість" лікарських препаратів. Запропонована інтелектуальна технологія на основі штучних імунних систем дозволяє зменшити погрішності енергетичних оцінок і підвищити достовірність прогнозу залежності "структура-властивість" хімічних сполук. Бібліогр.: 12 назв.

Ключові слова: технології иммунносетевого моделювання, інтелектуальна технологія, погрішності енергетичних оцінок.

УДК 004.89:004.41.

Development of immune-networks modeling technology for computers molecular design of medical products / Samigulina G.A., Chebeiko C.B. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2011. – №. 17. – P. 142 – 148.

It is developed immune nets the approach to modeling dependences "structure – property" of medical drags. The offered intellectual technology allows to reduce errors of power estimations and to raise reliability of the forecast of dependence "structure – property" of chemical compounds. Refs.: 12 titles.

Keywords: immune-networks modeling technology, intellectual technology, errors of power estimations.

Поступила в редакцию 14.02.2011