УДК 615.015.11

ДИСКРИМИНАНТНЫЙ АНАЛИЗ БИОЛОГИЧЕСКОЙ АКТИВНОСТИ ПРОИЗВОДНЫХ ЭСТРАДИОЛА

В.В.Иванов, Л.А.Слета, А.А.Толстая*

Харьковский национальный университет им. В.Н.Каразина,
61077, г. Харьков, пл. Свободы, 4. E-mail: anov@univer.kharkov.ua
* Национальный фармацевтический университет

Ключевые слова: дискриминантная функция; эстрадиолы; биологическая активность

Проанализирована связь биологической активности ряда эстрадиоловых производных с набором молекулярных параметров. Вычислена соответствующая дискриминантная функция, позволяющая находить соединения с высоким относительным сродством эстрадиолов к рецептору. Качество прогноза активных соединений составляет 87%.

DISCRIMINANT ANALYSiS OF ESTRADIOL DERIVATIVES BIOLOGICAL ACTIVITY
V.V.Ivanov, L.A.Sleta, A.A.Tolstaya
The biological activity of series estradiol derivatives has been analysed in connection with set of molecular parameters. The corresponding discriminant function which gives possibility to find compounds with high relative binding affinity of the estradioles to receptor has been calculated. The quality of prediction for compounds activity is 87%.

ДИСКРИМІНАНТНИЙ АНАЛІЗ БІОЛОГІЧНОЇ АКТИВНОСТІ ЕСТРАДІОЛОВИХ ПОХІДНИХ
В.В.Іванов, Л.А.Слєта, А.А.Толстая
Проаналізовано зв'язок біологічної активності ряду естрадіолових похідних з набором молекулярних параметрів. Розрахована відповідна дискримінантна функція, яка дозволяє знаходити сполуки з високою відносною спорідненістю до рецептора. Якість прогнозу активних сполук складає 87%.

Биологическая активность эстрогенов (в частности эстрадиолов), их синтетических модификаций, а также различных нестероидных аналогов многие годы находятся в центре внимания биохимиков и фармацевтов. Этот интерес стимулирует многочисленные попытки осмысления связи молекулярной структуры соединений с их биологической активностью. Методологической основой этой проблематики (Quantity Structure-Activity Relationship, QSAR[1]) является статистический анализ экспериментальных данных о биоактивности и установление связи этих данных с набором дескрипторов — разнообразных характеристик молекулярной структуры [1, 2]. Среди множества возможных подходов QSAR к проблеме активности эстрогенов можно выделить хорошо известные многопараметрические методы Фри-Вильсона и Хэнча, опирающиеся на регрессионный анализ данных по биоактивности. Развитие соответствующего математического аппарата этих моделей продолжается до сих пор [3]. Применение подобных подходов к анализу относительного сродства к рецептору (Relative Binding Affinity, RBA) эстрогенов позволило получить ряд корреляционных уравнений, связывающих численное выражение активности некоторых соединений с молекулярными параметрами (рефракция, константы Гаммета и т.д.). Обзор данных см. в [4]. Для исследования этих структур применяются также современные довольно дорогостоящие подходы типа метода сравнения молекулярных полей (CoMFA) [5, 6] из группы так называемого трехмерного QSAR (3D-QSAR).

Следует отметить однако, что экспериментальное исследование сродства лигандов к соответствующим рецепторам часто проводится с использованием тканей различных животных (мыши, крысы...) при различных температурах (0°C, 25°C и выше), с разным периодом экспозиции. Это приводит к тому, что методы регрессионного анализа (основанные на уравнениях Фри-Вильсона и Хэнча) так же как и подходы типа CoMFA, опирающиеся на методологию неполного метода наименьших квадратов (Partial Least Squares, PLS), не могут быть корректно применимы к большим массивам разнородных экспериментальных данных, поскольку такие методы предполагают единую шкалу биоактивности для всех препаратов. Указанная проблема в значительной степени "снимается" при использовании подходов в духе теории распознавания образов (pattern recognition, PR) [7], получивших в последнее время широкое распространение. Методы PR могут оказаться удобными также и в случаях, когда биоактивность

[1] В данной работе мы используем стандартные англоязычные аббревиатуры.

препаратов не имеет явного численного выражения, но имеются лишь указания на ее наличие или отсутствие. В связи с этим представляется целесообразным использование подобных методик для классификации соединений по степени их активности[2]. Наиболее подходящей методологией в рамках PR, с нашей точки зрения, является метод дискриминантных функций (DF), широко использующийся в социологии, психологии, технике и т.д. Имеются также данные (не столь многочисленные) о его применении в проблеме QSAR [8].

Расчет дискриминантных функций

Основная идея DF заключается в выборе такого (желательно минимального) набора дескрипторов, при котором группы активных и неактивных молекул образуют своеобразные кластеры. При этом молекулы можно рассматривать как условные "точки" в многомерном пространстве дескрипторов. После такого отбора сама дискриминантная функция определяется поверхностью в пространстве многих измерений, наиболее точно разделяющей группы активных и неактивных молекул. В частном случае линейная дискриминантная функция (DF) строится как суперпозиция дескрипторов:

$$DF(d_{1,2},...) = x_0 + x_1d_1 + x_2d_2 + ..., \quad (1)$$

где: $x_0,x_1,x_2...$ — коэффициенты, обеспечивающие эффективную классификацию молекул, а $d_1,d_2,...$ — отобранные дескрипторы. Уравнение вида

$$DF(d_1,d_2,...) = 0 \quad (2)$$

определяет разделяющую гиперплоскость. Обобщение функции (1) естественным образом приводит к квадратичной зависимости вида:

$$DF(d_1,d_2,...) = x_0 + \sum_i x_i d_i + \sum_{i,j} x_{ij} d_i d_j \quad (3)$$

Уравнение для разделяющей поверхности второго порядка таким образом определяет гиперквадрику.

Знание подходящим образом откалиброванной DF дает возможность проводить классификацию молекул на активные и неактивные.
Критерий активности имеет вид:

$$DF(d_1,d_2,...)>0, \quad (4)$$

а критерий неактивности:

$$DF(d_1,2,...)<0. \quad (5)$$

Важным моментом дискриминантного анализа является понятие центроида. Это условная точка в многомерном пространстве молекулярных параметров, имеющая в качестве координат среднее значение дескрипторов по данному классу активности. Таким образом, центроид соответствует гипотетической молекуле — наиболее типичному представителю данного класса активности. Значение функции DF для центроидов может служить опорным при оценке активности исследуемых молекул.

Техника нахождения величин x_I, x_{ij} хорошо известна [9] и представляет собой обобщенную задачу на собственные значения для ковариационных матриц, описывающих внутригрупповую и межгрупповую дисперсии. Соответствующий программный комплекс был разработан нами специально для оценок биологической активности молекул. Он включает регрессионный, дискриминантный и факторный анализ.

В настоящей работе мы провели дискриминантный анализ связи RBA с дескрипторами молекулярной структуры для синтетических производных эстрадиола. Экспериментальные данные о биоактивности, а именно lg(RBA), были взяты нами из обзора [4]. Они численно выражают активность относительно незамещенной молекулы, для которой величина lg(RBA) полагалась равной 2. В наших расчетах мы предполагали, что соединения с lg(RBA)≥1,5 являются активными, а соединения с lg(RBA)<1,5 — неактивными или слабоактивными. Анализировались данные для 130 различных производных с общей формулой:

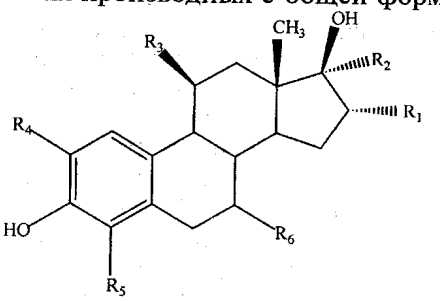

Рис. Замещенные эстрадиола

При этом заместители (R_1-R_6) являются различными химическими структурами (галогенами, алкенами, алкинами и т.д.), их полный перечень содержится в [4].

Для всех соединений проводилась оптимизация геометрии полуэмпирическим методом PM3. Кроме того, в рамках этого метода вычислялся набор электронных дескрипторов: заряды на атомах (наибольший положительный q+ и наибольший отрицательный q-), дипольный момент — μ (D), энергии верхней занятой ($\varepsilon_{ВЗМО}$) и нижней вакантной ($\varepsilon_{НВМО}$) молекулярных орбиталей (в эВ), теплота образования молекулы (ΔH) в ккал/моль. Кроме того, нами были оценены параметры, характеризующие геометрическую структуру молекул: объем $V(A^{\circ 3})$ и площадь поверхности $S(A^{\circ 2})$. Эти величины вычислялись по аддитивным схемам с использованием данных о Ван-дер-Ваальсовых радиусах атомов. Ряд параметров вычислялся с использованием программного комплекса DRAGON [1]. Среди них важнейшими являются молекулярная рефракция $MR(A^{\circ 3})$, поляризуемость α $(A^{\circ 3})$ и липофильность (логарифм кон-

[2] В настоящей работе использовался наиболеее рациональный вариант классификации молекул по двум классам: активные - неактивные.

станты распределения вещества между фазами октанол-вода, lgP). Последняя величина вычислялась по методу Моригучи, который также представляет собой разновидность аддитивной схемы (обзор методов оценки lgP см. в [10]).

Подчеркнем, что перечисленный набор дескрипторов характеризует всю молекулу в целом, а не отдельные заместители, как это принято, например, в подходе Хэнча. Это обстоятельство позволяет получить решающую функцию, являющуюся в известной степени универсальной, поскольку она не привязана к какой-либо базовой химической структуре и, таким образом, в принципе возможно ее использование для оценок lg(RBA) широкого класса соединений.

Очевидно, что многие из перечисленных выше молекулярных параметров существенно коррелируют друг с другом (например, MR, α, V). Поэтому в настоящей работе для выделения минимального набора дескрипторов, гарантирующего наиболее эффективное разделение кластеров активных и неактивных молекул, был проведен соответствующий анализ главных компонент (Principal Components Analysis, PCA), метод известен как разновидность факторного анализа. PCA позволил выделить среди наиболее важных компонент такие параметры: энергию нижней вакантной МО ($\varepsilon_{НВМО}$), поляризуемость (α) и липофильность (lgP).

Следует отметить также, что согласно уравнению Хэнча липофильность (lgP) связана с биоактивностью квадратично. Это обстоятельство важно тем, что позволяет предполагать наиболее оптимальные величины lgP для заданного типа активности. Поэтому вместо линейного варианта дискриминантной функции мы предполагаем зависимость, которая включает также квадрат липофильности $(lg\ P)^2$.

В результате проведенного дискриминантного анализа была получена следующая решающая функция для классификации молекул по указанным типам активности:

$$DF_1 = +1{,}032 - 1{,}49\varepsilon_{НВМО} - 0{,}089\alpha + 1{,}41lgP - 0{,}17(lgP)^2. \quad (6)$$

Для оценки качества прогноза нами использовался метод скользящего контроля. В англоязычной литературе по статистике он обычно известен как LOO (leave-one-out cross-validation). В этом подходе дискриминантная функция строится с использованием массива объектов (препаратов) за вычетом одного. После этого проводится прогноз активности этой (не включенной в обучающую выборку) системы. Затем процедура повторяется для следующей молекулы и т.д. После прохождения всего набора данных подсчитывается процент верных прогнозов, что и характеризует качество дискриминирования. Величину эту можно интерпретировать как вероятность обнаружения биоактивности у неисследованной активной молекулы. В наших расчетах для функции (6) качество прогноза составило 87%. Значения функции для центроидов активных $DF_1(+)$ и неактивных $DF_1(-)$ молекул соответственно равны:

$DF_1(+) = +0{,}401$;

$DF_1(-) = -0{,}502$.

Таким образом, положительное значение DF соответствует активным молекулам и, наоборот, отрицательное значение — малоактивным.

Определенный интерес может представлять также и максимально упрощенная функция разделения, использующая в качестве дискриминантного параметра лишь величину липофильности:

$$DF_2 = -4{,}386 + 2{,}307lgP - 0{,}261(lgP)^2,$$

качество прогноза для которой также относительно велико и составляет 78%.

Необходимо отметить, однако, что выбранный нами набор параметров не является единственным и уникальным, поскольку начальное пространство дескрипторов достаточно широко. По-видимому возможны иные их комбинации, которые также могут привести к достаточно эффективному разделению молекул по величине активности.

Выводы

Таким образом, на основе литературных экспериментальных данных для ряда производных эстрадиола вычислена решающая функция (DF_1) и ее упрощенный аналог (DF_2), которые позволяют классифицировать соединения по двум типам активности. Выбранные типы активности соответствуют значениям $lg(RBA) \geq 1{,}5$ и $lg(RBA) < 1{,}5$. Полученные дискриминантные функции могут быть использованы для первичной оценки биоактивности эстрогенов и их нестероидных аналогов в качестве простого и быстрого метода предварительного скрининга.

Литература

1. *Раевский О.А. // Успехи химии. — 1999. — Т. 68, №6. — С. 555-564.*
2. *Karelson M., Lobanov V.S., Katritzky A.R. // Chem. Rev. — 1996. — Vol. 96. — P. 1027-1043.*
3. *Holik M., Halamek J. // QSAR. — 2002. — Vol. 20. — P.422-428.*
4. *Gao H., Katzenellenbogen J.A., Garg R., Hansch C. // Chem. Rev. — 1999. — Vol. 99. — P. 723-744.*
5. *Sippl W. // J. Comp.-Aided Molecular Design. — 2000. — Vol. 14. — P. 559-572.*
6. *Yu S.J., Keenan S.M., Tong W., Welsh W.J. // Chem. Res., Toxicol. — 2002. — Vol. 15. — P. 1229-1234.*
7. *Jain A.K., Duin R.P.W. // IEEE Transact. on Pattern Analysis and Machine Intelligence. — 2000. — Vol. 22, №1. — P. 4-38.*
8. *Филимонов Д.А., Поройков В.В., Караичева Е.И. и др. // Экспер. клинич. фармакол. — 1995. — Т. 58, №2. — С. 56-62.*
9. *Mc.Lachlan G.J. Discriminant analysis and statistical pattern recognition. — New York: John Wiley & Sons, 1992. — 387 p.*
10. *Leo A.J. // Chem. Rev. — 1993. — Vol. 93. — P. 1282-1306.*

Надійшла до редакції 19.09.2003 р.