ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ПРИ КЛАССИФИКАЦИИ БИОЛОГИЧЕСКИ АКТИВНЫХ СОЕДИНЕНИЙ НА ОСНОВЕ МОЛЕКУЛЯРНЫХ ТОПОЛОГИЧЕСКИХ ДЕСКРИПТОРОВ

Кабанкин А.С., Габриелян Л.И.
Центр теоретических проблем физико- химической фармакологии РАН, г. Москва

Кабанкин А.С.
Габриелян Л.И.
Центр теоретических проблем физико- химической фармакологии РАН

Из практики расчетов известно, что многие молекулярные дескрипторы, используемые при поиске связей структура-биологическая активность, особенно дескрипторы одного типа, например, топологические индексы молекулярной связности различных порядков для путей, достаточно сильно коррелируют между собой. Поскольку применение коррелирующих параметров, особенно в регрессионных задачах, считается недопустимым, иногда вместо исходных топологических дескрипторов используют их линейные комбинации – главные компоненты (ГК). Преимущество метода главных компонент состоит в том, что, во-первых, число параметров (главных компонент) существенно меньше числа исходных дескрипторов, во-вторых, эти параметры ортогональны, т.е. статистически независимы друг от друга. Как правило, ограничиваются рассмотрением лишь нескольких первых ГК (обычно это ГК, соответствующие собственным значениям корреляционной матрицы ≥ 1), на которые приходится максимальная доля общей дисперсии дескрипторов, а остальные ГК отбрасывают. При этом предполагается, что компоненты с наибольшей дисперсией обладают и наибольшей информативностью в отношении принадлежности соединения к той или иной группе. Однако, проведенные нами расчеты показали, что в классификационных задачах, в отличие от регрессионных задач, ситуация, вообще говоря, неоднозначна.

Методом главных компонент нами проведено исследование массива, содержащего 30 производных адамантана, каждое из которых было описано набором из 100 топологических дескрипторов. Ранее [1] для этого массива методом линейного дискриминантного анализа найдены дискриминантные функции, позволяющие на основе комбинации из 4-6 топологических дескрипторов прогнозировать гепатопротекторное детоксицирующее действие производных адамантана с вероятностью 0,7-0,8. Для корреляционной матрицы, отвечающей представленному массиву данных, число ненулевых собственных значений и ненулевых соответствующих главных компонент (ГК) оказалось равным 29, а число значимых главных компонент KL, вычисленное по формуле, предложенной в [2], равно 17. Однако для последующего анализа было оставлено 18 ГК, поскольку различие между средними значениями 18-й ГК в активной и неактивной группах статистически значимо по критерию D2.

Методом линейного дискриминантного анализа была определена индивидуальная информативность каждой из ГК при распознавании обучающей выборки, содержащей 12 активных и 18 неактивных соединений. Различие в средних значениях 11-й, 12-й и 18-й ГК для активной и неактивной групп статистически значимо, поскольку величина D2-статистики превышает критическое значение 3,84 при уровне значимости $\alpha = 0,05$, тогда как среди исходных топологических дескрипторов был лишь один статистически значимый дескриптор - wundsmax. Эти же ГК дают наименьшую ошибку при классификации соединений обучающей выборки.

Далее методом скользящего контроля была исследована предсказательная способность линейных дискриминантных функций, вычисленных для различных комбинаций ГК, в отношении гепатопротекторного детоксицирующего действия. Дискриминантные функции, построенные на первых ГК, включающих максимальную долю дисперсии исходной матрицы данных , не могут быть использованы для прогнозирования гепатопротекторного действия в ряду адамантанов. С другой стороны, дискриминантные функции даже на основе трех ГК (11-й, 12-й и 18-й), учитывающих очень малую часть общей дисперсии, обладают значительной предсказательной способностью (83,3% в целом). Добавление некоторых других, не первостепенных ГК увеличивает предсказательную способность соответствующих дискриминантных функций в отношении гепатопротекторного детоксицирующего действия до 90% в целом, что превышает аналогичные показатели для дискриминантных функций на основе наилучших комбинаций топологических дескрипторов. Ма-

лая информативность первых ГК для классификации соединений по активности, возможно, связана с тем, что эти ГК включают нечто общее, присущее всему массиву соединений, а различия между группами проявляются лишь в ГК с малым вкладом в общую дисперсию.

Таким образом, не отбрасывание, а учет главных компонент, вносящих малый и очень малый, но ненулевой вклад в общую дисперсию исходной матрицы топологических дескрипторов, позволяет получить эффективные дискриминантные функции, предсказательная способность которых в отношении гепатопротекторного детоксицирующего действия производных адамантана заметно выше по сравнению с дискриминантными функциями на основе наилучших комбинаций первоначальных топологических дескрипторов.

Литература

1. *Пирузян Л.А., Кабанкин А.С., Габриелян Л.И. и др.* Гепатопротекторное действие производных адамантана и анализ связи структура- детоксицирующая активность// Химико- фармацевтический журнал. 2004. №3. С.19-25.
2. *Todeschini R.* Data correlation, number of significiant principal components and shape of molecules. The *K* correlation index// Anal. Chim. Acta. 1997. v.348. P.419-430.