БИОФИЗИКА, 2018, том 63, вып. 2, с. 213–224

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 57.087, 519.6

МЕТОД SEM: МОДЕЛИРОВАНИЕ СТРУКТУРНЫМИ УРАВНЕНИЯМИ В МОЛЕКУЛЯРНОЙ БИОЛОГИИ

© 2018 г. А.А. Иголкина, М.Г. Самсонова

Санкт-Петербургский политехнический университет Петра Великого, 195251, Санкт-Петербург, Политехническая ул., 29

E-mail: igolkinanna11@gmail.com

Поступила в редакцию 17.01.17 г.

Моделирование структурными уравнениями – это метод многомерного анализа второго поколения для оценки причинных взаимодействий, сочетающий в себе целый ряд статистических подходов (регрессионный анализ, анализ путей и факторный анализ). В обзоре рассмотрены все основные типы моделей при моделировании структурными уравнениями и методы оптимизации параметров моделей. Кроме того, детально разобраны характерные практические примеры моделирования структурными уравнениями в областях молекулярной биологии, охватывающие моделирование биохимических процессов, моделирование влияния генетических маркеров на заболевания и моделирование взаимодействий в генных сетях.

Ключевые слова: SEM, структурные уравнения, молекулярная биология, методы оптимизации, генетические сети, генетические маркеры.

Моделирование структурными уравнениями (Structural Equation Modeling, SEM) – это метод многомерного анализа для оценки причинных взаимодействий, сочетающий в себе идеи регрессионного анализа, анализа путей (path analysis) и факторного анализа (factor analysis). Этот метод был впервые предложен генетиком С. Райтом в начале XX века [1,2] и сегодня является мощной техникой для анализа сложных структур взаимодействий компонент различной природы, включая циклические структуры.

Математическое описание модели SEM состоит из двух частей: структурной и измерительной. Необходимость моделирования двух частей вызвана присутствием в такой модели латентных (скрытых) переменных – некоторых абстрактных конструктов, которые не могут быть измерены явно, но могут быть введены в модель через некоторые наблюдаемые переменные [3,4]. В качестве примера латентной переменной в области молекулярной биологии можно привести «стресс клетки» или «белковый комплекс». Каждую из этих переменных нельзя измерить напрямую, но можно определить, например, во втором случае – через переменные концентраций субъединиц белкового комплекса.

Важной составляющей моделирования структурными уравнениями является используемый метод оценки параметров моделей. Основной и исторически первый принцип оценки параметров моделей SEM заключается в сравнении выборочной матрицы ковариаций для наблюдаемых переменных и матрицы ковариаций, выраженной через параметры модели [3,4]. Однако широкое использование SEM в последние десятилетия повлекло разработку новых методов оптимизации параметров и подходов к построению моделей SEM.

Настоящий обзор структурирован следующим образом: в первом разделе приведено описание модели SEM, во втором – рассмотрены методы оптимизации параметров модели, в третьем – описаны критерии, используемые для оценки качества моделей, и в заключительном, четвертом разделе приведены примеры применения SEM в области молекулярной биологии.

1. МОДЕЛИ SEM

Рассмотрим модель многомерной линейной регрессии (Multiple Linear regression, MLR) для описания взаимодействий между зависимыми переменными (откликами) Y_i, $i = \overline{1, n}$ и независимыми (факторами, индикаторами) X_j, $j = \overline{1, m}$:

Сокращения: SEM – моделирование структурными уравнениями, LISREL – SEM-модель с латентными переменными, RMSEA – среднеквадратичная погрешность аппроксимации, SRMR – стандартизированный среднеквадратичный остаток, SNP – однонуклеотидный полиморфизм, MAPK – митоген-активируемая протеинкиназа.

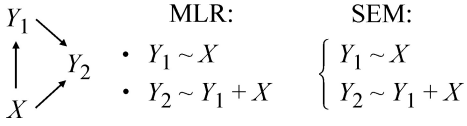

Рис. 1. Пример задачи, в которой одна из переменных (Y_1) является индикатором для другой переменной (Y_2) и в то же время является откликом для третьей (X). Согласно теории многомерной линейной регрессии оценка параметров взаимодействий между переменными X, Y_1 и Y_2 производится независимо в двух моделях. SEM-модель для рассматриваемой задачи заключает в себе все структурные уравнения, оценка всех параметров модели производится одновременно.

$$Y = BX + \varepsilon, \tag{1.1}$$

где B – матрица коэффициентов размера ($n \times m$), ε – вектор случайных ошибок. Эта модель представляет собой пример так называемой модели первого поколения, так как в ней зависимые и независимые переменные разделены (находятся с противоположных сторон от знака равенства). Если в задаче существуют такие переменные, которые в одних взаимодействиях участвуют как отклики, а в других – как индикаторы, то оценка параметров взаимодействий в такой задаче согласно теории многомерной линейной регрессии будет выполнена в рамках нескольких независимых моделей (рис. 1).

Оценка параметров взаимодействий в такого рода задаче с помощью SEM не требует независимого анализа нескольких моделей и поэтому является примером моделей второго поколения. В простейшем случае математическая модель SEM записывается в следующем виде:

$$Y = BY + \Gamma X + \zeta, \tag{1.2}$$

где Y – вектор эндогенных переменных, X – вектор экзогенных переменных с матрицей ковариаций cov(X) = Ф, ζ – вектор случайных ошибок, независимых от X, с матрицей ковариаций cov(ζ) = Ψ, матрица B содержит параметры линейных взаимодействий между эндогенными переменными, матрица Γ содержит параметры линейных взаимодействий экзогенных и эндогенных переменных. Полный набор параметров модели θ представляет собой набор всех элементов матриц B, Γ, Ψ и Ф.

Традиционные методы оценки параметров модели SEM базируются на представлении матрицы ковариаций между наблюдаемыми переменными через параметры модели. В случае SEM модели (1.2) матрица ковариаций вектора $\begin{bmatrix} Y \\ X \end{bmatrix}$ выражается через параметры модели следующим образом:

$$\Sigma(\theta) = \begin{bmatrix} (1-B)^{-1}(\Gamma\Phi\Gamma^T + \Psi)(1-B)^{-1T} & (1-B)^{-1}\Gamma\Phi \\ \Phi\Gamma^T(1-B)^{-1T} & \Phi \end{bmatrix}$$

Существует целый ряд методов оценки параметров θ, которые так или иначе минимизируют разницу между матрицей $\Sigma(\theta)$ и выборочной матрицей ковариаций вектора $\begin{bmatrix} Y \\ X \end{bmatrix} - S$.

SEM позволяет работать с моделями, в которых часть переменных – латентные, т.е. не могут быть измерены напрямую, а только лишь представлены через ряд наблюдаемых переменных. Классическая SEM-модель с латентными переменными называется LISREL (1.3) и, как всякая модель с латентными переменными, состоит из двух частей – структурной и измерительной [5]. Структурная часть описывает взаимодействия между латентными переменными, а измерительная отражает то, через какие наблюдаемые переменные измеряются латентные.

LISREL structural part:

$$\eta = B\eta + \Gamma\xi + \zeta,$$

LISREL measurement part:

$$Y = \Lambda_y\eta + \varepsilon,\ X = \Lambda_x\xi + \delta. \tag{1.3}$$

В LISREL модели η и ξ обозначают соответственно вектора эндогенных и экзогенных латентных переменных; Y и X обозначают вектора наблюдаемых переменных, через которые выражаются эндогенные и экзогенные латентные переменные. Векторы случайных ошибок ζ, ε и δ предполагаются распределенными нормально со средним, равным нулю, и матрицами ковариаций Ψ, Θ_ε и Θ_δ соответственно; матрица ковариаций экзогенных латентных переменных полагается равной Ф. Матрицы B и Γ содержат параметры взаимодействий между латентными переменными; матрицы Λ_y и Λ_x содержат параметры связи между наблюдаемыми переменными и латентными (так называемые факторные нагрузки). Полный набор параметров модели θ представляет

собой набор всех элементов матриц B, Γ, Ψ, Θ_ε, Θ_δ, и Φ. Матрица ковариаций вектора $\begin{bmatrix} Y \\ X \end{bmatrix}$ выражается через параметры модели LISREL следующим образом:

$$\Sigma(\theta) = \begin{bmatrix} \Lambda_y(I-B)^{-1}(\Gamma\Phi\Gamma^T + \Psi)(1-B)^{-1T}\Lambda_y^T + \Theta_\varepsilon & \Lambda_y(I-B)^{-1}\Gamma\Phi\Lambda_x^T \\ \Lambda_x\Phi\Gamma^T(1-B)^{-1T}\Lambda_y^T & \Lambda_x\Phi\Lambda_x^T + \Theta_\delta \end{bmatrix} \quad (1.4)$$

Дальнейшее усложнение модели LISREL происходило путем синтеза структурной и измерительной частей модели, а также объединения эндогенных и экзогенных переменных. Так, например, была предложена обобщенная модель SEM (General MSEM) [6]:

structural part:

$$\eta = B\eta + \Gamma X + \zeta,$$

measurement part:

$$Y = \Lambda\eta + \mathrm{K}X + \varepsilon, \quad (1.5)$$

где η обозначает вектор латентных переменных (экзогенных и эндогенных), X – вектор экзогенных наблюдаемых переменных, Y – вектор наблюдаемых переменных, через которые проявляются латентные.

2. МЕТОДЫ ОПТИМИЗАЦИИ ПАРАМЕТРОВ SEM-МОДЕЛЕЙ

Методы, основанные на работе с матрицами ковариаций. Исторически первые методы оценки параметров модели SEM основаны на минимизации различий между выборочной матрицей ковариаций для наблюдаемых переменных S и матрицей ковариаций Σ(θ), выраженной через параметры модели.

Простейший метод оценки параметров – невзвешенный метод наименьших квадратов (Unweighted Least Squares, ULS). В качестве целевой функции для минимизации рассматривается сумма квадратов отклонений всех элементов матриц S и Σ(θ) и записывается следующим образом:

$$F_{\mathrm{ULS}}(\theta) = \frac{1}{2}\mathrm{tr}[(S - \Sigma(\theta))^2], \quad (2.1)$$

где tr( ) означает след матрицы. В случае проявления гетероскедастичности или наличия автокорреляции случайных ошибок рекомендуется использовать обобщенный метод наименьших квадратов (Generalized Least Squares, GLS) со следующей целевой функцией:

$$F_{\mathrm{GLS}}(\theta) = \frac{1}{2}\mathrm{tr}[(S - \Sigma(\theta)S^{-1})^2]. \quad (2.2)$$

Целевые функции (2.1) и (2.2) являются частными случаями минимизируемой функции в методе взвешенных наименьших квадратов (Weighted Least Squares, WLS)

$$F_{\mathrm{WLS}}(\theta) = \frac{1}{2}\mathrm{tr}[(S - \Sigma(\theta)W^{-1})^2]. \quad (2.3)$$

где W – некоторая весовая матрица, равная единичной в случае невзвешенного метода наименьших квадратов и равная S в случае обобщенного метода наименьших квадратов [3].

Метод максимального правдоподобия. Метод максимального правдоподобия для оценки параметров моделей SEM по своей идее отличается от методов, описанных в предыдущей части, однако по существу тоже сравнивает матрицу ковариаций для наблюдаемых переменных S с матрицей ковариаций, выраженной через параметры модели Σ(θ).

Метод максимального правдоподобия основывается на предположении, что вектор всех наблюдаемых переменных $\begin{bmatrix} Y \\ X \end{bmatrix}$ имеет многомерное нормальное распределение $N(0, \Sigma(\theta))$. Согласно этому предположению, выборочная матрица ковариаций наблюдаемых переменных подчинена распределению Уишарта [7]. Собрав независящие от параметров распределения множители в общую константу C, плотность распределения Уишарта можно записать в следующем упрощенном виде:

$$f_W(S, \Sigma(\theta), n) = \frac{C}{|\Sigma(\theta)|^{n/2}}\exp\left[-\frac{n}{2}\mathrm{tr}(S\Sigma(\theta)^{-1})\right], \quad (2.4)$$

где n обозначает размер выборки. В предположении нормальности наблюдаемых переменных для выборочной матрицы ковариаций определяется функция отношения правдоподобия (Likelihood Ratio, LR), равная отношению плотностей распределения Уишарта для матрицы S при условии, что наблюдаемые переменные распределены $N(0, \Sigma(\theta))$, и при условии, что наблюдаемые переменные распределены $N(0, S)$ (2.4):

$$LR(\theta) = \frac{\exp\left[-\frac{n}{2}\mathrm{tr}(S\Sigma(\theta)^{-1})\right]}{\exp\left[-\frac{n}{2}\mathrm{tr}(SS^{-1})\right]} \frac{|\Sigma(\theta)|^{-\frac{n}{2}}}{|S|^{-\frac{n}{2}}}, \quad (2.5)$$

Максимизация функции LR по параметрам θ означает максимизацию пропорции функции правдоподобия относительно идеальной модели. После применения логарифмирования к (2.5) максимизация LR заменяется на минимизацию следующей функции:

$$F_{\mathrm{ML}}(\theta) = \mathrm{tr}(S\Sigma(\theta)^{-1}) + \log|\Sigma(\theta)| + \log|S| - p, \quad (2.6)$$

где p – количество наблюдаемых переменных (= $\mathrm{tr}(SS^{-1})$). Последнее слагаемое равно следу единичной квадратной матрицы с размером, равным количеству наблюдаемых переменных. Если оценки параметров $\hat{\theta}$ приводят к равенству $\Sigma(\hat{\theta}) = S$, то $F_{\mathrm{ML}}(\hat{\theta}) = 0$, что является минимумом функции $F_{\mathrm{ML}}(\theta)$. Таким образом, метод максимального правдоподобия для оценки параметром моделей SEM тоже является в некотором смысле методом подгонки матрицы $\Sigma(\theta)$ к выборочной матрице ковариаций S [3,8].

Байесовский подход к оптимизации параметров моделей SEM. В отличие от рассмотренных выше методов оптимизации параметров, в байесовском подходе вектор параметров θ рассматривается как случайная величина с так называемым априорным распределением $p(\theta)$. При заданной выборке значений наблюдаемых переменных Z апостериорное распределение значений параметров обозначается $p(\theta|Z)$. По теореме Байеса, с учетом того, что $p(Z) = \mathrm{const}$, имеем:

$$p(\theta|Z) \propto p(Y|\theta)p(\theta),$$

или иначе:

$$\log p(\theta|Z) \propto \log p(Y|\theta) + \log p(\theta).$$

В правой части последнего выражения первое слагаемое представляет собой функцию правдоподобия для выборки Z. При увеличении размера выборки первое слагаемое будет увеличиваться, тогда как второе – нет и при достаточно большом размере выборки станет несущественным. Таким образом, байесовский подход, направленный на максимизацию $\log p(\theta|Z)$, и метод максимального правдоподобия, направленный на максимизацию $\log p(Z|\theta)$, считаются асимптотически эквивалентными [9].

В моделировании SEM кроме наблюдаемых используются еще латентные переменные, выборочные значения которых обозначим через Ω. Байесовский подход к оценке параметров модели SEM заключается в получении формул для следующих апостериорных распределений: $p(\Omega|\theta,Z)$ и $p(\theta|\Omega,Z)$. Вывод формул для этих распределений приведен в четвертой главе монографии [9]. На основании полученных апостериорных распределений оценки для θ и Ω получаются с помощью метода Монте-Карло по схеме марковских цепей (Markov Chain Monte Carlo, MCMC). С помощью данного метода можно итеративно генерировать выборку из совместного распределения $p(\theta,\Omega|Z)$, основываясь на $p(\Omega|\theta,Z)$ и $p(\theta|\Omega,Z)$ и последовательно обновляя значения θ и Ω. Поскольку новые значения θ и Ω зависят от предыдущих, то итеративная процедура является марковской цепью. Если такая цепь достаточно длинная, то $p(\theta,\Omega|Z)$ сходится к стационарному распределению. Для обновления значений θ и Ω рекомендуется использовать алгоритм сэмплирования по Гиббсу (Gibbs sampling). Пусть на j-й итерации были получены значения $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_a^{(j)})$ и $\Omega^{(j)} = (\Omega_1^{(j)}, \ldots, \Omega_b^{(j)})$. Согласно алгоритму сэмплирования по Гиббсу обновление параметров происходит следующим образом [10]:

$$\begin{gathered}
\theta_1^{(j+1)} \text{ из } p(\theta_1|\theta_2^{(j)}, \ldots, \theta_a^{(j)}, \Omega^{(j)}, Z) \\
\theta_2^{(j+1)} \text{ из } p(\theta_1|\theta_1^{(j+1)}, \ldots, \theta_a^{(j)}, \Omega^{(j)}, Z) \\
\ldots \qquad\qquad \ldots \\
\theta_a^{(j+1)} \text{ из } p(\theta_1|\theta_1^{(j+1)}, \ldots, \theta_{a-1}^{(j+1)}, \Omega^{(j)}, Z) \\
\Omega_1^{(j+1)} \text{ из } p(\Omega_1|\theta^{(j+1)}, \Omega_2^{(j)}, \ldots, \Omega_b^{(j)}, Z) \\
\Omega_2^{(j+1)} \text{ из } p(\Omega_1|\theta^{(j+1)}, \Omega_1^{(j+1)}, \ldots, \Omega_b^{(j)}, Z) \\
\ldots \qquad\qquad \ldots \\
\Omega_b^{(j+1)} \text{ из } p(\Omega_1|\theta^{(j+1)}, \Omega_1^{(j+1)}, \ldots, \Omega_{b-1}^{(j+1)}, Z).
\end{gathered}$$

Байесовский подход к оптимизации параметров моделей SEM позволяет справиться с проблемой малого размера выборки с помощью запуска симуляции методом Монте-Карло по схеме марковских цепей несколько раз.

Регуляризация. Под регуляризацией в статистических методах и в машинном обучении понимается некоторое дополнительное условие во избежание некорректно поставленной задачи или переобучения. Так, при работе с многомерной линейной регрессией регуляризация вводится как дополнительное слагаемое (регуляризационная функция) в целевой функционал в виде штрафа за сложность модели. Выделяют три основные типа регуляризационых функций для многомерной линейной регрессии:

– Ridge-регуляризация [11]

$$R_2(\theta) = \alpha_2 \Sigma_{i=1}^{n} \theta^2;$$

– LASSO-регуляризация (Least Absolute Shrinkage and Selection Operator) [12]

$$R_1(\theta) = \alpha_1 \Sigma_{i=1}^{n} |\theta_i|;$$

– ElasticNet-регуляризация [13]

$$R_3(\theta) = \alpha_1 \Sigma_{i=1}^{n} |\theta_i| + \alpha_2 \Sigma_{i=1}^{n} \theta^2.$$

Таким образом, новый целевой функционал представим в виде $F_{new}(\theta) = F_{initial}(\theta) + R(\theta)$.

Подобный подход был предложен в работе [14], в которой авторы применили регуляризацию при оценке параметров моделей SEM. В качестве минимизируемого целевого функционала авторы ввели следующий: $F_{RegSEM}(\theta) = F_{ML}(\theta) + R(\theta)$, где $F_{ML}(\theta)$ рассчитывается по формуле (2.6), а $R(\theta)$ – одна из трех выше предложенных регуляризационных функций. Введенная регуляризация помогает не только справиться с переобучением, но и помочь с решением выбора наилучшей модели SEM, описывающей данные.

Зачастую при работе с моделями SEM требуется спецификация – улучшение первичной конфигурации путем добавления или удаления взаимодействий. Если n – количество возможных варьируемых взаимодействий, то количество всех возможны конфигураций модели – 2^n. Чтобы не производить моделирование всех возможных конфигураций, авторы работы [14] предлагают следующий подход. Необходимо последовательно увеличивать размер штрафа (α_1 и/или α_2) от нуля до такого значения, при котором оптимальные оценки всех параметров модели оказываются статистически неотличимы от нуля. Такое последовательное увеличение штрафа порождает серию SEM-моделей. Лучшей предлагается считать ту модель, которая в полученной серии имеет наилучшие значения критериев качества (см. раздел 3).

3. КРИТЕРИИ ОЦЕНКИ КАЧЕСТВА МОДЕЛЕЙ SEM

После того, как была произведена оптимизация параметров, необходимо оценить насколько хорошо полученная модель соответствует данным. Существует ряд критериев и статистик, позволяющих оценить такое соответствие. После оценки параметров модели SEM методами, которые так или иначе минимизируют различия между ковариационными матрицами $\Sigma(\theta)$ и S, используется статистика χ^2, которая по сути отражает разницу между этими матрицами. Другими критериями являются среднеквадратичная погрешность аппроксимации RMSEA (Root mean square error of approximation) и стандартизированный среднеквадратичный остаток SRMR (standardised root mean square residual).

При выборе наилучшей модели из ряда альтернативных используются значения следующих критериев соответствия модели данным (index of fit): информационный критерий Акайке (Akaike information criterion, AIC), байесовский информационный критерий (Bayesian information criterion, BIC), индекс критерия согласия (goodness of fit index, GFI), сравнительный индекс согласия (comparative fit index, CFI), коэффициент согласия (Tucker-Lewis index, TLI). Обычно в работах используют несколько критериев на усмотрение авторов [4].

Кроме анализа критериев, отражающих, насколько модель хорошо описала данные, для моделей SEM производится расчет статистической значимости для каждого параметра модели. Для каждого параметра с помощью теста Вальда проверяется нулевая гипотеза: параметр θ_i равен нулю. Статистика Вальда представляет собой отношение значения параметра к стандартному отклонению и является обобщением Z-теста Фишера. После применения теста Вальда для каждого параметра модели определяется p-value – вероятность ошибки первого рода.

4. ПРИМЕРЫ ПРИМЕНЕНИЯ SEM В МОЛЕКУЛЯРНОЙ БИОЛОГИИ

Работы, рассмотренные в этом разделе в качестве примеров, были выбраны с целью продемонстрировать широкий круг задач молекулярной биологии и типы данных, к которым применимы модели SEM.

Моделирование биохимических процессов. Использование SEM для моделирования биохимических реакций иллюстрируют две работы, в первой из которых авторы исследуют белок-белковые взаимодействия, а во второй – причинные связи между соединениями, участвующими в биохимических реакциях, и их параметрами.

В работе [15] авторы исследовали процесс деградации растительного фермента RuBisCO, играющего важнейшую роль в фиксации неорганического углерода и ремобилизации азота из вегетативных тканей в молодые растущие листья и семена. Несмотря на свою важность, этот фермент в растениях является относительно неэффективным, из-за чего растениям приходится интенсивно синтезировать RuBisCO, де-

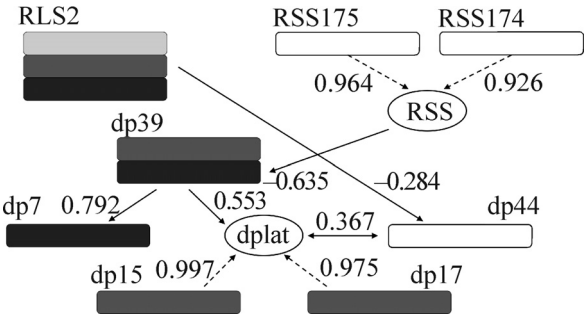

Рис. 2. Схема причинных связей между субъединицами RuBisCO и продуктами их деградации, отражающая финальную модель SEM. Обозначение «dp» в названии узлов схемы означает, что узел отвечает некоторому продукту деградации (degradation product, dp). Узлы, обведенные овалом, обозначают латентные переменные, пунктирные стрелки к латентным переменным проведены от тех наблюдаемых переменных, через которые были введены латентные. Одинаковые фрагменты аминокислотных последовательностей субъединиц RuBisCO, присутствующие в продуктах деградации, выделены одним оттенком серого. Схема построена на основе рис. 2 из работы [15].

лая его наиболее распространенным ферментом на планете. Одним из факторов, влияющим на эффективность ремобилизации азота, является динамика деградации RuBisCO, моделирование которой с помощью SEM было проведено в работе [16]. В качестве данных для моделирования были использованы результаты масс-спектрометрии и 2D-электрофореза, по которым судили об уровне экспрессии белков. На основании последовательностей, полученных с помощью масс-спектрометрии, были определены первичные и вторичные продукты распада субъединиц RuBisCO.

Схема деградации RuBisCO была построена на основании изменения протеома в трех временны́х точках и представляла собой последовательный каскад распада субъединиц фермента (большая (Rubisco Large Subunit, RLS2) и малые (Rubisco Small Subunits, RSS174 и RSS175)) на первичные и вторичные продукты. Базовая модель SEM для процесса деградации отражала построенную схему и затем была модифицирована добавлением новых связей между переменными в модели и удалением статистически не значимых взаимодействий (рис. 2). В модель также были введены латентные переменные, отражавшие белки, экспрессия которых была сильно скоррелирована (т.е. присутствовала мультиколлинеарность): RSS (соответствует двум малым субъединицам) и dplat (соответствует двум продуктам вторичной деградации dp15 и dp17). Результирующая модель имела RMSEA-индекс < 0,1 и p-value для каждого параметра, меньшее 0,05. Оценки параметров модели SEM были получены с помощью байесовского метода оптимизации.

Моделирование с помощью SEM процесса деградации RuBisCO подтвердило первоначальную гипотезу о многоэтапной деградации этого фермента и выявило особенности, которые ранее не были известны, а именно наличие перекрестных связей между первичными и вторичными продуктами деградации малых и больших субъединиц: между dplat и d44 и между dp39 и RSS.

В работе [17] авторы моделировали биохимические процессы в красных кровяных клетках пациентов с биполярным аффективным расстройством, прошедших Li^+-терапию. Препараты, содержащие Li^+, приводят к существенному снижению количества аффективных состояний у пациентов с биполярным расстройством. Такое положительное влияние ионов лития объясняется двумя гипотезами: (1) нарушением Na^+/Li^+-обмена через клеточную мембрану, (2) конкуренцией ионов Li^+ с ионами Mg^{2+} за сайты связывания с Mg^{2+} в малых биомолекулах. С помощью SEM авторы работы [17] построили возможный биохимический путь превращения Li, в котором сочетаются обе гипотезы. В качестве параметров модели были выбраны следующие одиннадцать величин: концентрации ионов Li^+ внутри красных кровяных клеток и в плазме ($[Li^+]_i$ и $[Li^+]_e$), кинетические параметры Na^+/Li^+-обмена (V_{std}, V_{max}, K_m), внутриклеточная концентрация Mg^{2+} ($[Mg^{2+}]_f$), константа связывания Li^+ с клеточной мембраной (K_{Li}) и параметры фосфолипидного состава мембраны. Авторы сформулировали шесть альтернативных гипотетических путей, описывающих взаимодействия между одиннадцатью рассматриваемыми переменными, в каждом из которых $[Li^+]_i$ выступал в качестве экзогенной переменной, так как пациенты проходили Li^+-терапию. Затем были сконструированы соответствующие модели SEM и проведена оценка параметров с помощью метода максимального правдоподобия. Наилучшая финальная модель была выбрана как единственная, у которой все взаимодействия между переменными были статистически значимыми (рис. 3). Чтобы добиться большего соответствия модели данным, в модель были добавлены дополнительные связи между переменными, и наилучшая конфигурация имела следующие значения критериев оценки качества модели: индекс критерия согласия – 0,93, сравнительный индекс согласия – 0,96.

Наилучшая модель не только продемонстрировала статистическую значимость причинных связей между параметрами, описывающими

две рассматриваемые гипотезы, но также позволила лучше понять возможный биохимический механизм, лежащий в основе Li^+-терапии пациентов с биполярным синдромом.

Моделирование влияния однонуклеотидных полиморфизмов, ассоциированных с экспрессией (eQTL), на болезни. Полногеномный поиск ассоциаций (Genome-wise association studies, GWAS) является популярным методом для нахождения генетических вариантов, ассоциированных с тем или иным биологическим или молекулярным фенотипом. Использование SEM позволяет эффективно агрегировать генетические варианты, соответствующие различным генам, и описывать сложные взаимодействия между генетическими факторами и интересующими характеристиками, например состоянием пациента.

В работе [18] моделирование SEM было использовано для анализа связи между генетическими вариантами и четырьмя основными признаками ожирения: мерой жировой прослойки в области живота и спины, индексом массы тела и гипертонией. Предварительный анализ показал, что единственный сигнальный путь из базы данных KEGG, связанный с каждым из четырех рассматриваемых признаков – это MAPK-сигнальный путь (MAPK – mitogen-activated protein kinase, митоген-активируемая протеинкиназа). Среди однонуклеотидных полиморфизмов (single nucleotide polymorphisms, SNPs), локализованных внутри генов, участвующих в MAPK-пути, были выбраны те, которые оказались статистически значимо связаны с каждым из четырех рассматриваемых признаков. Выбранные полиморфизмы были агрегированы в одну латентную переменную, отвечающую за MAPK-сигнальный путь. Для четырех признаков ожирения и введенной латентной переменной были предложены две структурные модели (рис. 4). Оценки параметров моделей были произведены с помощью метода максимального правдоподобия, и наилучшая из двух моделей была выбрана согласно наибольшему значению индекса критерия согласия, равному 0,79, и наименьшему значения критерия Акайке, равному 1132,00.

Согласно наилучшей модели (рис. 4а), все три меры ожирения и генетические факторы влияют на гипертонию. Однако генетические маркеры, связанные с MAPK-сигнальным путем, а также объем жировых прослоек влияют на гипертонический статус скорее косвенно, посредством индекса массы тела.

В работе [19] авторы построили более сложную модель взаимодействий между генетическими вариантами (однонуклеотидными полиморфизмами) и целым рядом метаболических коморбидных болезней – коронарной сердечной недостаточностью, диабетом второго типа, подагрой, болезнью почек и инсультом. Под коморбидностью понимают одновременное присутствие у пациента нескольких заболеваний, имеющих в качестве причины общий механизм. Была сконструирована модель SEM, которая отражала такой механизм. Переменные в модели, отражающие заболевание, были бинарными. В модель были включены два биохимических показателя в качестве непрерывных переменных: показатель гликированного гемоглобина (HbA1c) и уровень содержания мочевой кислоты в крови. Кроме того, в модели присутствовали три латентные переменные – фактор ожирения, фактор дислипидемии и фактор кровяного давления. Отобранные на основании статистических тестов генетические варианты входили в модель в виде бинарных переменных. Спецификация модели, т.е. уточнение структуры взаимодействий между переменными, была проведена итеративно с помощью алгоритма обратного удаления (backwards elimination): стартуя со структуры полносвязного графа, алгоритм постепенно удалял по одному из взаимодействий, если такое удаление приводило к улучшению модели по критериям сравнительного индекса согласия, коэффициента согласия и RMSEA. Для финальной модели значения сравнительного индекса согласия и коэффициента согласия составляли $\geq 0{,}95$, RMSEA $\leq 0{,}06$. Оценку параметров моделей проводили с помощью метода максимального правдоподобия.

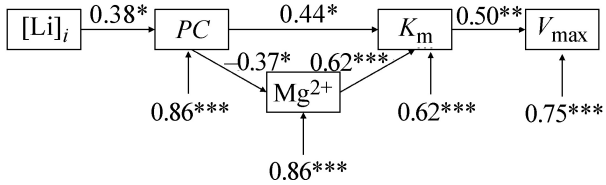

Рис. 3. Схема финальной модели SEM взаимодействий между биохимическими показателями (переменными) у пациентов с биполярным расстройством. $[Li^+]_i$ – концентрации ионов Li^+ внутри красных кровяных клеток; *PC* – параметр фосфолипидного состава мембраны; Mg^{2+} – внутриклеточная концентрация Mg^{2+}; K_m – константа Михаэлиса–Ментен Na^+/Li^+-обмена; V_{max} – максимальная скорость реакции Na^+/Li^+-обмена. Числа над стрелками, соединяющими переменные, отражают значения параметров связи между ними (*B*), числа в нижней строке схемы отражают параметры случайных ошибок Ψ. Сноски около чисел означают статистическую значимость параметров: * – *p*-value < 0,05; ** – *p*-value < 0,01; *** – *p*-value < 0,01. Схема построена на основе рис. 3 из работы [17].

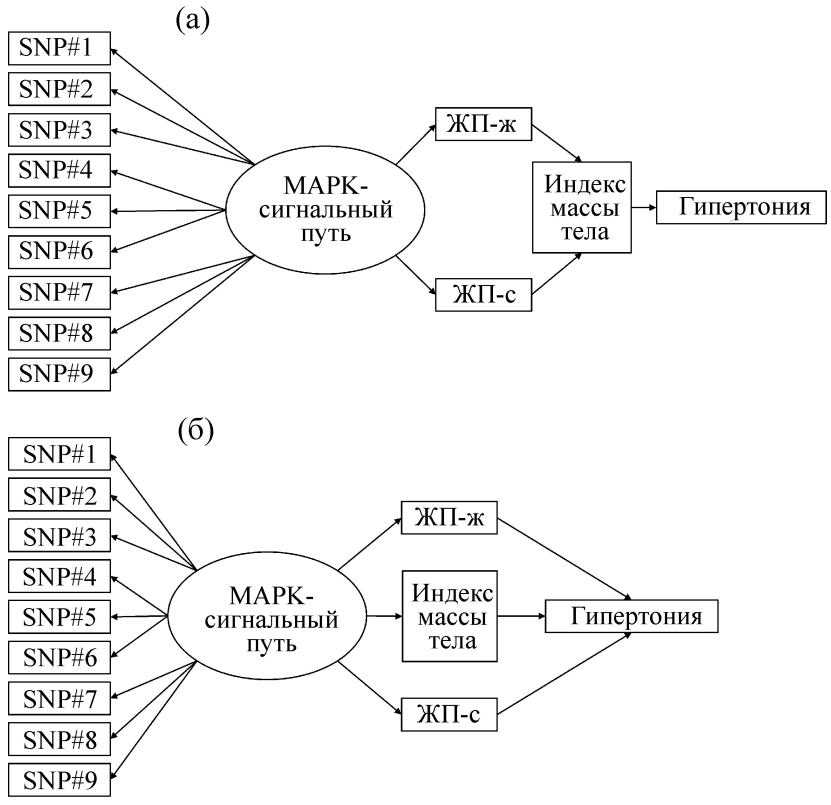

Рис. 4. Две схемы SEM-моделей, описывающих влияние генетических факторов через меры ожирения на гипертонию. Прямоугольники с названиями «SNP-номер» означают различные генетические SNPs-маркеры. Точные идентификаторы SNP представлены в работе [18]. Овал с названием «MAPK» означает латентную переменную, отвечающую за MAPK-сигнальный путь. Прямоугольники «ЖП-ж» и «ЖП-с» обозначают меры жировой прослойки в области живота и спины соответственно. Схема построена на основе рис. 1 из работы [18].

На основании полученной картины взаимодействий (рис. 5) авторы работы [19] подчеркивают важность фактора ожирения, показателя гликированного гемоглобина и уровня содержания мочевой кислоты в крови как движущих элементов в развитии метаболического синдрома. Поэтому авторы заключили, что метаболический синдром не является болезнью в биологическом смысле, но может использоваться для клинической диагностики как индикатор, указывающий на совместное присутствие факторов риска коморбидных метаболических заболеваний – показателя гликированного гемоглобина и уровня содержания мочевой кислоты в крови, факторов ожирения, дислипидемии и кровяного давления.

Моделирование ген-генных взаимодействий в сетях. Снижение стоимости секвенирования транскриптомов (РНК-секвенирование) в последние несколько лет привело к возможности анализировать в одной работе большое количество профилей экспрессии генов одновременно. Этот тренд привел к росту работ по анализу взаимодействий генов в сигнальных и генных сетях с помощью SEM [20–26].

Генная сеть представляет собой ориентированный граф, узлы которого описывают белки (продукты генов) или белковые комплексы, а ребра – направленные взаимодействия между белками. Количественное описание взаимодействий в генных сетях на основании данных об экспрессии генов для компонент сети – еще одна задача, с которой справляется SEM-моделирование.

В работе [26] авторы исследовали с помощью моделирования структурными уравнениями отличия во взаимодействии генов, контролирующих сигнальные пути, у здоровых людей и у больных пациентов, страдающих такими нейродегенеративными заболеваниями, как лобно-височная лобарная дегенерация с убиквитинированными включениями и рассеянный склероз.

Перед SEM-моделированием авторы идентифицировали те сигнальные пути, которые были наиболее обогащены дифференциально экспрессирующимися генами, т.е. такими генами,

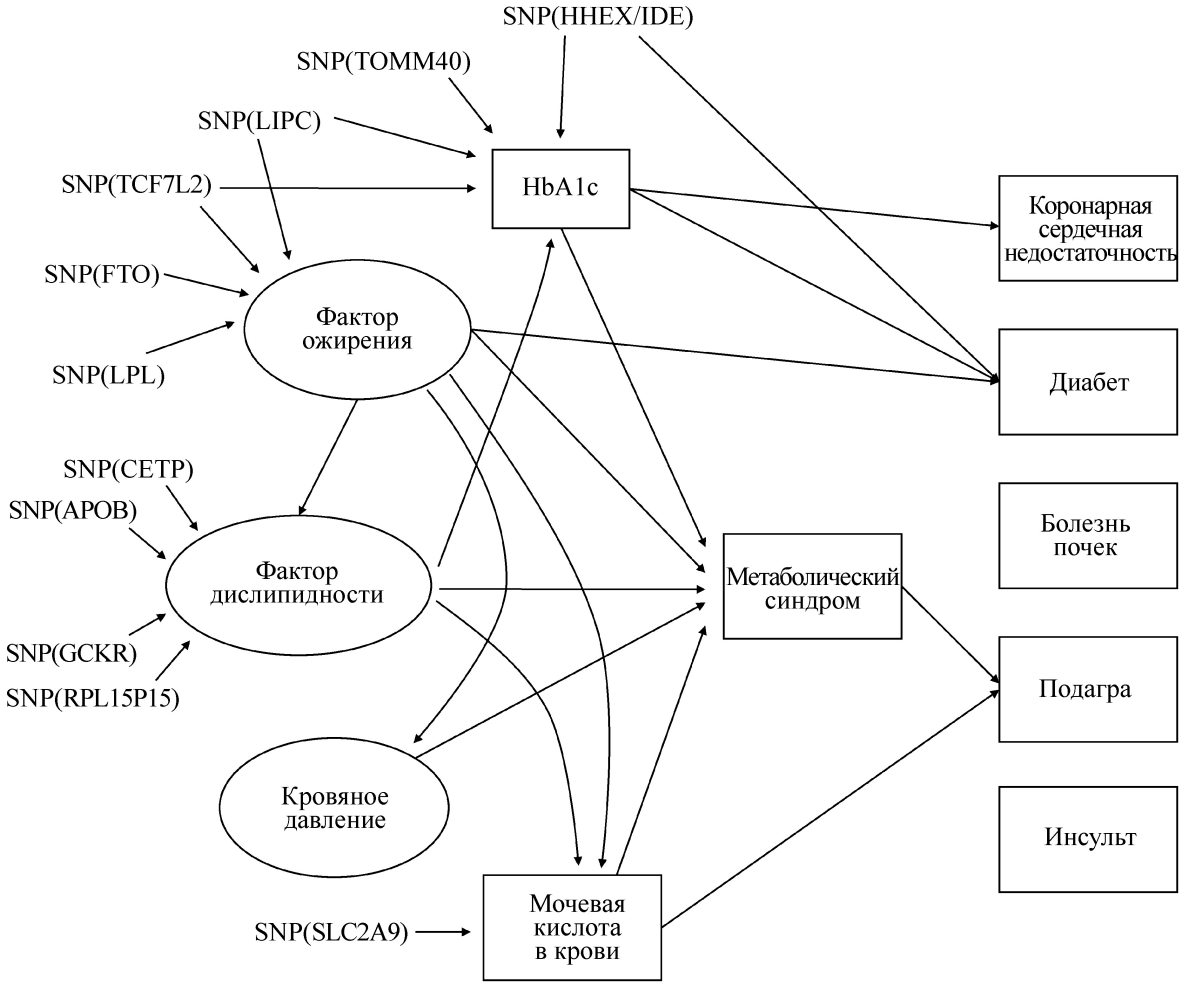

Рис. 5. Схема взаимодействий между генетическими маркерами (обозначены «SNP(название гена)») и коморбидными заболеваниями. Схема построена на основе рис. 2 из работы [19]. Значения параметров модели и их статистическая значимость представлены на исходной схеме.

средний уровень экспрессии которых значимо отличался между группами здоровых людей и больных пациентов. Такими путями для лобно-височной лобарной дегенерации и рассеянного склероза оказались соответственно глутаматергический синапс и фагоцитоз, опосредованный Fc-фрагментом иммуноглобулина. Для построения моделей SEM для этих путей был использован формализм генных сетей. Сконструированные генные сети затем были редуцированы так, чтобы содержать лишь кратчайшие пути между дифференциально экспрессирующимися генами в графе.

Вслед за построением SEM-моделей и их редукцией была проведена их дальнейшая модификация – последовательное добавление узлов и ребер между узлами в графе сети согласно улучшению критериев качества моделей χ^2, RMSEA, SRMR (рис. 6). Для финальных моделей генных сетей глутаматергического синапса и фагоцитоза, опосредованного Fc-фрагментом иммуноглобулина, значение критерия качества SRMR было равно 0,092 и 0,098 соответственно.

Оценка параметров моделей SEM была произведена независимо для групп здоровых и больных пациентов. Межгрупповое сравнение моделей было выполнено с помощью тестирования двух гипотез. Первая касалась матрицы ковариаций наблюдаемых переменных и постулировала отсутствие различий между матрицами, H_0: $\Sigma(\hat{\theta}_1) = \Sigma(\hat{\theta}_2)$ где $\hat{\theta}_1$ и $\hat{\theta}_2$ – оценки параметров моделей для двух групп. Вторая нулевая гипотеза была сформулирована для каждого параметра взаимодействия по отдельности, она постулировала отсутствие различий между параметрами. Выявленные статистически значимые различия во взаимодействиях между генами в финальной модели оказались дейст-

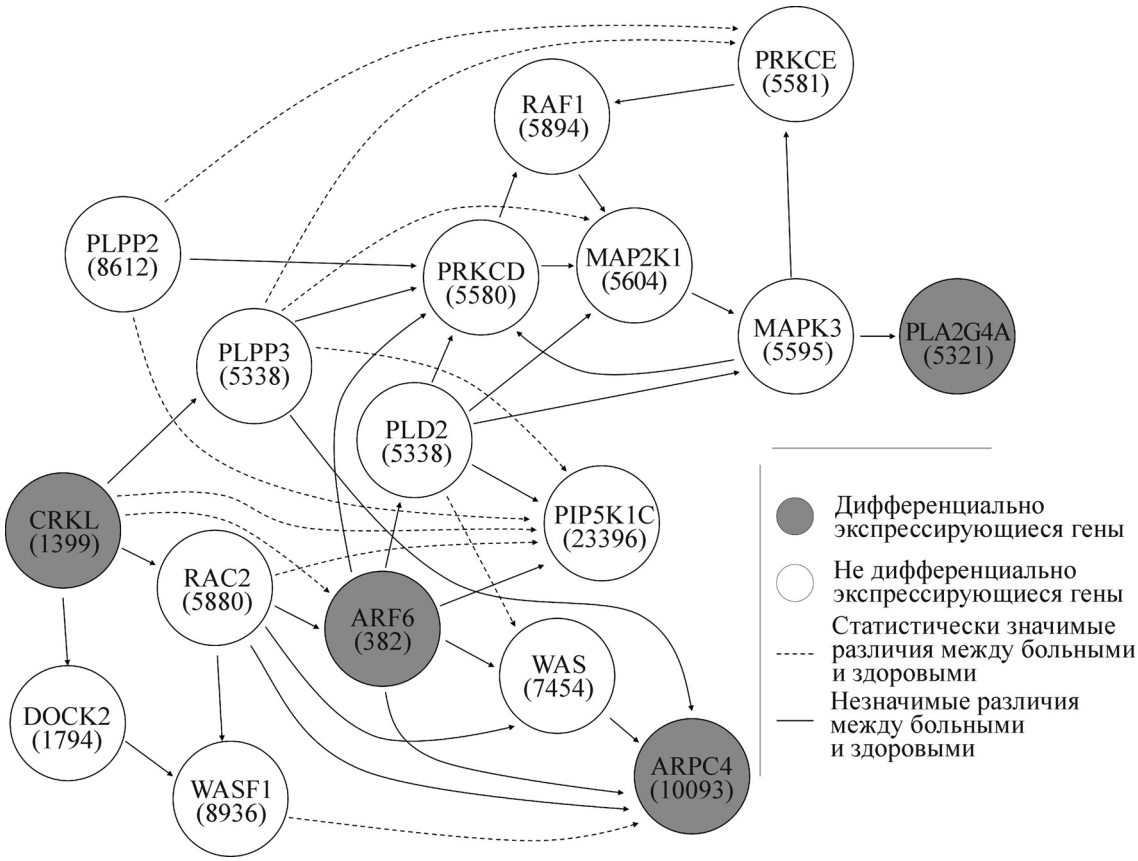

Рис. 6. Схема взаимодействий между генами в генной сети глутаматергического синапса, соответствующая финальной модели SEM. Серыми кружками обозначены дифференциально экспрессирующиеся гены, белыми кружками – не дифференциально экспрессирующиеся гены. Стрелки означают направленные взаимодействия между генами. Пунктирными линиями обозначены статистически значимые отличия в значениях предсказанных параметров между группами больных и здоровых пациентов (p-value < 0,05), сплошными линиями – статистически незначимые. Схема построена на основе рис. 3 из работы [26]. Значения параметров модели представлены на исходной схеме.

вительно связанными с рассматриваемыми нейродегенеративными заболеваниями.

ЗАКЛЮЧЕНИЕ

Моделирование структурными уравнениями – мощная техника многомерного анализа, включающая в себя широкий спектр постановок задач и методов оптимизации параметров моделей. В настоящем обзоре рассмотрены все основные модели SEM и методы оптимизации, по которым достигнут консенсус в научном сообществе и которые применяются на практике. Но моделирование с помощью структурных уравнений – динамически развивающаяся область исследований. Наибольшее внимание на сегодняшний уделяется разработке нелинейных и динамических моделей SEM и использованию более адекватных гипотез о типе распределения данных, нежели чем предположение о нормальности распределения [27–30]. Усложнение SEM-моделей неизбежно порождает разработку новых методов оптимизации, большинство из которых использует байесовский подход [31].

Несмотря на большое количество методов оптимизации параметров SEM-моделей, на практике наиболее популярными до сих пор остаются традиционный методы, основанные на сравнении выборочной и полученной из модели матриц ковариаций – метод максимального правдоподобия и обобщенный метод наименьших квадратов. Скорее всего, использование байесовских подходов вытеснит традиционные подходы, так как первые требуют меньшего количества наблюдений и являются более гибкими к предположению о характере распределения параметром моделей [9].

При практическом использовании моделирования структурными уравнениями наибольшее внимание уделяется не методам оптимизации параметров модели, а ее спецификации, т.е. определению структурной части модели.

Зачастую исследователям известна только некоторая опорная модель, которая затем модифицируется с целью получить увеличение интересующих критериев качества модели. В настоящем обзоре продемонстрированы два основных подхода к спецификации моделей – когда в опорную модель последовательно вносят добавления ребер (пример моделирования отличий во взаимодействии генов, контролирующих сигнальные пути, у здоровых людей и у больных, страдающих нейродегенеративными заболеваниями [26]) или удаления ребер (пример моделирования взаимодействий между генетическими вариантами (однонуклеотидными полиморфизмами) и метаболическими коморбидными болезнями [19]). Разумной альтернативой такого последовательного улучшения моделей может стать SEM-моделирование с использованием регуляризатора при оценке параметров моделей.

Авторы выражают благодарность С.А. Руколайне за ценные замечания и рекомендации по материалу и содержанию статьи.

Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации, задание № 1.8697.2017/БЧ.

СПИСОК ЛИТЕРАТУРЫ

1. S. Wright, Genetics **3**, 367 (1918).
2. S. Wright, J. Agric. Res. **20**, 557 (1921).
3. K. A. Bollen, *Structural Equations with Latent Variables* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 1989). doi:10.1002/9781118619179.
4. R. B. Kline, *Pronciples and practice of Structural Equation Modeling*, 3rd ed. (The Gulford Press, 2011).
5. K. G. Jöreskog, U. H. Olsson, and F. Y. Wallentin, *Multivariate Analysis with LISREL* (Springer International Publishing, Cham, 2016). doi:10.1007/978-3-319-33153-9.
6. K. J. Preacher, M. J. Zyphur, and Z. Zhang, Psychological Methods **15**, 209 (2010). doi:10.1037/a0020141.
7. J. Wishart, Biometrika **20A**, 32 (1928). doi:10.1093/biomet/20A.1-2.32.
8. L. A. Hayduk, *Structural equation modeling with LISREL: Essentials and advances* (MD: Johns Hopkins University, Baltimore, 1988).
9. S.-Y. Lee, *Structural Equation Modeling: A Bayesian approach* (John Wiley & Sons, Ltd, Chichester, UK, 2007). doi:10.1002/9780470024737.
10. F. Yanuar, J. of Physics: Conf. Ser. **495**, 12047 (2014). doi:10.1088/1742-6596/495/1/012047.
11. A. E. Hoerl and R. W. Kennard, Technometrics **12**, 55 (1970). doi:10.1080/00401706.1970.10488634.
12. R. Tibshirani, J. Royal Stat. Soc.: Ser. B **73**, 273 (2011). doi:10.1111/j.1467-9868.2011.00771.x.
13. H. Zou and T. Hastie, J. Royal Stat. Soc.: Ser. B **67**, 301 (2005). doi:10.1111/j.1467-9868.2005.00503.x.
14. R. Jacobucci, K. J. Grimm, and J. J. McArdle, Struct. Equation Modeling **23**, 555 (2016). doi:10.1080/10705511.2016.1154793.
15. C. Tétard-Jones, A. M. R. Gatehouse, J. Cooper, et al., PLoS One **9** (2014). doi:10.1371/journal.pone.0087597.
16. U. Feller, I. Anders, T. Mae, J. Exp. Botany **59**, 1615 (2008). doi:10.1093/jxb/erm242.
17. N. Williams, B. T. Layden, J. Suhy, et al., Bipolar Disorders **5**, 320 (2003). doi:054 pii].
18. J.-Y. Kim, J.-H. Namkung, S.-M. Lee, and T.-S. Park, Genomics & Informatics **8**, 150 (2010). doi:10.5808/GI.2010.8.3.150.
19. R. Karns, P. Succop, G. Zhang, et al., Obesity **21**, 745 (2013). doi:10.1002/oby.20445.
20. D. Pepe and J. H. Do, J. Comput. Biol. **23** (2015). doi:10.1089/cmb.2015.0156.
21. J. M. Fear, M. N. Arbeitman, M. P. Salomon, et al., BMC Systems Biol. **9**, 1 (2015). doi:10.1186/s12918-015-0200-0.
22. X. Mi, K. Eskridge, D. Wang, et al., Stat. Applications in Genetics and Mol. Biol. **9** (2010). doi:10.2202/1544-6115.1552.
23. R. Li, S. Tsaih, K. Shockley, et al., PloS Genetics **2** (2006). doi:10.1371/journal.pgen.0020114.
24. D. L. Remington, Genetics **181**, 1087 (2009). doi:10.1534/genetics.108.092668.
25. N. L. Nock and L. X. Zhang, BMC Proceedings **5**, S47 (2011). doi:10.1186/1753-6561-5-S9-S47.
26. D. Pepe and M. Grassi, BMC Bioinformatics **15**, 132 (2014). doi:10.1186/1471-2105-15-132.
27. J. Chen and X. Tan, J. Multivariate Analysis **100**, 1367 (2009). doi:10.1016/j.jmva.2008.12.005.
28. N. Umbach, K. Naumann, H. Brandt, and A. Kelava, J. Statistical Software **77** (2017). doi:10.18637/jss.v077.i07.
29. T. Asparouhov, E. L. Hamaker, and B. Muthén, Structural Equation Modeling **24**, 257 (2017). doi:10.1080/10705511.2016.1253479.
30. S. J. Finney and C. DiStefano, In *Structural Equation Modeling: A Second Course*, Ed. by G. R. Hancock & R. O. Mueller (IAP Information Age Publishing, 2013), pp. 439–492.
31. X. Mi, K. Eskridge, D. Wang, et al., Gen. Res. **92**, 239 (2010). doi:10.1017/S0016672310000236.

SEM: Structural Equation Modeling in Molecular Biology

A.A. Igolkina and M.G. Samsonova

Peter the Great Saint Petersburg Polytechnic Univeristy, Polytekhnicheskaya 29, St. Petersburg, 195251 Russia

Structural Equation Modelling is a second-generation multivariate method which is used to estimate causal interactions among a set of variables and includes, as special cases, regression analysis, path analysis and confirmatory factor analysis. In this review, we focused on all main structural equation modelling models and on different methods for model parameter estimation. We also discussed representative examples of the utility of structural equation modelling within molecular biology: modelling of biochemical processes, modelling of relationships between genetic markers and diseases, modelling of interactions within gene networks.

Keywords: SEM, structural equations, molecular biology, maximum likelihood and Bayesian approaches, gene networks, genetic markers