

Секция «Вычислительная математика и кибернетика»

Модели данных микрочипов ДНК, учитывающие эффекты насыщения и кросс-гибридизации

Когадеева Мария Сергеевна

Студент

Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия

E-mail: m.kogadeeva@gmail.com

Микрочипы ДНК — перспективная технология, позволяющая идентифицировать экспрессию десятков тысяч генов одновременно. Микрочип представляет собой пластину с пробами — известными фрагментами одноцепочечной ДНК длиной 25 нуклеотидов. После нанесения на микрочип специально подготовленного образца, в котором с определёнными участками ДНК связаны флюоресцентные метки, происходит соединение комплементарных участков последовательностей (реакция гибридизации). Сканирование микрочипа позволяет выделить области с высокой интенсивностью свечения и таким образом определить участвующие в гибридизации пробы, оценить состав исследуемого образца и уровень экспрессии генов.

Для анализа данных, полученных при сканировании микрочипа ДНК, используется модель адсорбции Ленгмюра [1][3]. В этой модели интенсивность каждой пробы I_i представляется в виде $I_i = \frac{\alpha_i C_i}{C_i + \beta_i} + b_i$, где C_i — концентрация ДНК, кодирующей ген, соответствующий i -й пробе; β_i — параметр диссоциации в законе насыщения. Параметр нормализации α_i и фоновая поправка b_i зависят от технических характеристик конкретного чипа. Однако на интенсивность может влиять также кросс-гибридизация — связывание проб с участками ДНК, кодирующими неспецифичные гены. В рамках одной модели практически невозможно в явном виде учесть все биохимические и биофизические факторы, влияющие на гибридизацию [2].

В данной работе предлагается модель данных микрочипа ДНК, учитывающая кросс-гибридизацию проб и два вида технического шума: $I_i^t = d^t \sum_j A_{ij} C_j^t + b_i^t + g_i^t$, где t — номер чипа, i — номер пробы; суммирование идёт по всем генам, которые могут присутствовать в образце, $A_{ij} \geq 0$ — коэффициент склонности i -й пробы к взаимодействию с j -м геном, $C_j^t \geq 0$ — концентрация j -го гена на t -м чипе, d^t — параметр нормализации, b_i^t — шум типа пятен (например, пузыри воздуха в образце), g_i^t — гладкий шум (например, засветка при сканировании). Начальные значения A_{ij} задаются из биохимических соображений, но, вообще говоря, матрица (A_{ij}) неизвестна и сильно разрежена. Концентрации C_j^t и параметры модели A_{ij} , b_i^t , g_i^t предлагается оценивать путём минимизации регуляризованной среднеквадратичной невязки:

$$Q = \sum_t \left(\sum_i (I_{xy}^t - Y_{xy}^t)^2 + \lambda_1 \sum_{xy} (|\Delta_x b_{xy}^t| + |\Delta_y b_{xy}^t|) + \lambda_2 \sum_{xy} ((\Delta_x g_{xy}^t)^2 + (\Delta_y g_{xy}^t)^2) \right),$$

где x, y — координаты i -й пробы на чипе, Y_{xy}^t — наблюдаемая интенсивность, λ_1 и λ_2 — параметры регуляризации, $\Delta_x z_{xy} = z_{x+1,y} - z_{xy}$ и $\Delta_y z_{xy} = z_{x,y+1} - z_{xy}$ — разностные операторы. Рассматривается также нелинейная модель, учитывающая одновременно неспецифические взаимодействия и закон насыщения. Модели настраиваются и сравниваются на данных, полученных в ходе лабораторных экспериментов.

Литература

1. Held, G.A., Grinstein, G., Tu, Y. Relationship between gene expression and observed intensities in DNA microarrays - a modeling study // *Nucleic Acids Research*, 2006. No. 34(9), p.e70.
2. Koltai, H., Weingarten-Baror, C. Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction // *Nucleic Acids Research*, 2008. No. 36(7), p.2395-2405.
3. Ono, N., Suzuki, S. et.al. An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays // *Bioinformatics*, 2008. No. 24(10), p.1278-85.