ПРИМЕНЕНИЕ МЕТОДА PLS-DA ДЛЯ РАЗДЕЛЕНИЯ ПРОМОТОРНЫХ И НЕПРОМОТОРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ РАЗЛИЧНЫХ КЛАССОВ

Темлякова Е.А., Камзолова С.Г., Джелядин Т.Р., Сорокин А.А.

Институт Биофизики Клетки РАН, Россия, 142290, Пущино, Институтская ул. 3

Задача пределения промоторных участков на хромосоме прокариотических организмов до сих пор не имеет удовлетворительного решения. Используя дискриминантный анализ при помощи проекции на латентные структуры (PLS-DA) нами были построены три модели, обученные разделять разные классы последовательностей *Escherichia coli* K12, а именно:

- **Модель 1:** промоторные последовательности и случайные нескоррелированные бернуллиевские последовательности;
- **Модель 2:** промоторные последовательности и кодирующие и межгенные участки;
- **Модель 3:** промоторные последовательности и промоторные "островки" [2].

PLS-DA представляет собой эффективный инструмент для разделения различных классов данных с большой размерностью [1]. Каждый исследуемый объект описывается набором дискрипторов, формирующих матрицу данных X (любые характеристики, которые можно вычислить или определить экспериментально), и набором откликов, формирующих матрицу Y (класс объекта). PLS-DA определяет новую систему координат для матриц X и Y одновременно, как линейную комбинацию дискрипторов и, таким образом, позволяет уменьшить размерность исходных данных, сохраняя большую часть полезной информации. В качестве дискрипторов в данной работе использовался профиль электростатического потенциала, рассчитанный при помощи оригинального метода [3].

В результате было показано, что точность предсказаний на тестовых выборка составляла примерно 75% для каждой из моделей. При этом, если в качестве дополнительного критерия использовать текстовый анализ исследуемых последовательностей [2], то точность предсказаний повышается до 89-95%.

Работа была поддержана грантом РФФИ №11-04-01436-а.

Литература.

1. *Sarker M., Rayens W.* Partial least squares for discrimination// J.Chemom., 2003, V.17, p.166
2. *Shavkunov K.S., Masulis I.S., Tutukina M.N., Deev A.A., Ozoline O.N.* Gains and unexpected lessons from genome-scale promoter mapping // Nucleic Acids Research, V.37, 2009, pp.4919-4931
3. *Polozov R.V., Dzhelyadin T.R., Sorokin A.A., Ivanova N.N., Sivozhelezov V.S., Kamzolova S.G.* Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences// J. Biomol. Struct. Dyn. 16, 6, 1999, pp.1135 - 1143.