**А.А. Дунаев, А.Э. Кель, И.В. Лобив, Ф.А. Мурзин,**
**О.Н. Половинко, Е.С. Черемушкин**

# ВИЗУАЛИЗАЦИЯ ГЕНЕТИЧЕСКОЙ ИНФОРМАЦИИ*

## ВВЕДЕНИЕ

В настоящее время проведены основные экспериментальные работы по секвенированию нуклеотидных последовательностей. Для хранения получаемой первичной информации созданы и постоянно пополняются такие специализированные банки данных, как EMBL и GenBank. В то же время, несмотря на наличие большого количества отсеквенированных последовательностей, наши представления о принципах их организации весьма ограничены. Поэтому одним из ведущих направлений молекулярной биологии в последнее время становится компьютерный анализ генетических текстов [1,2].

Проблематика идентификации структурно-функциональной организации генома наряду с такими вопросами, как распознавание интронов, экзонов или сайтов сплайсинга, включает в себя и круг задач, связанных с регуляцией транскрипции генов позвоночных [3].

Ввиду больших объемов генетических текстов возникает необходимость в визуализации генетической информации. Визуализация генетических текстов может стать необходимым шагом в процессе решения различных генетических задач, например задач распознавания специфичных участков ДНК (генов, сайтов и т.д.) [4]. Визуальный анализ биологических последовательностей [5, 6, 7] дает возможность определить структуру информации, закодированной в геноме, а также корректно выбрать метод для анализа этой структуры.

## 1. АЛГОРИТМЫ ВИЗУАЛИЗАЦИИ

Авторами были разработаны несколько алгоритмов для представления генетических текстов в графической форме и пакет программ, реализующий

* Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 01-01-794) и Министерства образования РФ.

данные возможности. Ниже описаны некоторые из реализованных алгоритмов.

### 1.1. Визуализация частот нуклеотидов

Пусть $\overline{S}$ — последовательность в четырехбуквенном алфавите A, C, G, T. Обозначим $k$-й элемент последовательности как $s_k$, и длину последовательности — $M$.

Пусть дано $N$. Обозначим через $BL_N[i]$ подпоследовательность $\overline{S}$ длины $N$, начинающуюся с $i$-й позиции, т.е. $BL_N[i] = s_i \ldots \ s_{N+i-1}$.

Пусть $n_A[i,N]$, $n_C[i,N]$, $n_G[i,N]$, $n_T[i,N]$ — количества букв A, C, G, T рассматриваемой подпоследовательности $BL_N[i]$ соответственно. Если $i, N$ заранее известны, то будем писать для краткости $n_A$, $n_C$, $n_G$, $n_T$.

Легко видеть, что $n_T = N - (n_A + n_C + n_G)$. Это означает, что достаточно изучать поведение трех компонентов. Отсюда могут быть вычислены частоты $p_A = n_A / N$, $p_C = n_C / N$, $p_G = n_G / N$.

Введем $\overline{p}_A = f(p_A)$, $\overline{p}_C = f(p_C)$, $\overline{p}_G = f(p_G)$, где $f(x) = \mathrm{int}(255 \times x)$. Тогда тройка $\langle \overline{p}_A,\ \overline{p}_C,\ \overline{p}_G \rangle$ может быть рассмотрена как вектор компонентов цвета $\langle R,\ G,\ B \rangle$ соответственно.

Цветное изображение может быть задано тремя матрицами $S = (S_R, S_G, S_B)$, $S_R = S_R(i,j)$, $S_G = S_G(i,j)$ $S_B = S_B(i,j)$, $0 \le i \le n-1$, $0 \le j \le m-1$. Обычно значения $S_R(i,j), S_G(i,j), S_B(i,j)$ лежат в диапазоне от 0 до 255. Набор троек $\{ (r,g,b) : 0 \le r,g,b < 255 \}$ называется цветовым кубом. Наша задача состоит в построении изображения, отражающего адекватность данных частот.

Предположим, что даны две позиции $i_1$, $i_2$ на последовательности $\overline{S}$, $i_1 - i_2 \le n \cdot m$ и $i_1 \le k \le i_2$. Далее, рассматриваемое окно $BL_N[k]$ движется вдоль данной последовательности $\overline{S}$. Затем мы получаем соответствующую тройку $\langle \overline{p}_A,\ \overline{p}_C,\ \overline{p}_G \rangle$ для каждой позиции $k$.

Поэтому запишем

$$\langle \overline{p}_A,\ \overline{p}_C,\ \overline{p}_G \rangle = \langle \overline{p}_A(k),\ \overline{p}_C(k),\ \overline{p}_G(k) \rangle = \langle R(k),\ G(k),\ B(k) \rangle .$$

Теперь мы можем создать следующее изображение

$$S_R(i,j)=\begin{cases} R(i_1+m\cdot i+j-1),\ i_1+m\cdot i+j-1\le n\cdot m;\\ 0,\ i_1+m\cdot i+j-1>n\cdot m;\end{cases}$$

$$S_G(i,j)=\begin{cases} G(i_1+m\cdot i+j-1),\ i_1+m\cdot i+j-1\le n\cdot m;\\ 0,\ i_1+m\cdot i+j-1>n\cdot m;\end{cases}$$

$$S_B(i,j)=\begin{cases} B(i_1+m\cdot i+j-1),\ i_1+m\cdot i+j-1\le n\cdot m;\\ 0,\ i_1+m\cdot i+j-1>n\cdot m.\end{cases}$$

Осуществляется последовательное заполнение изображения пикселями в процессе обозрения компонентов $\langle R,\ G,\ B\rangle$.

Вначале мы заполняем верхний ряд, т.е. $i=0$, затем первый и т.д. Аналогично, двигаясь вдоль последовательности $\overline{S}$, мы можем получить второе изображение, третье и т.д. Как результат, получаем последовательность изображений, которые образуют видеоряд и могут быть представлены в виде AVI-файла.

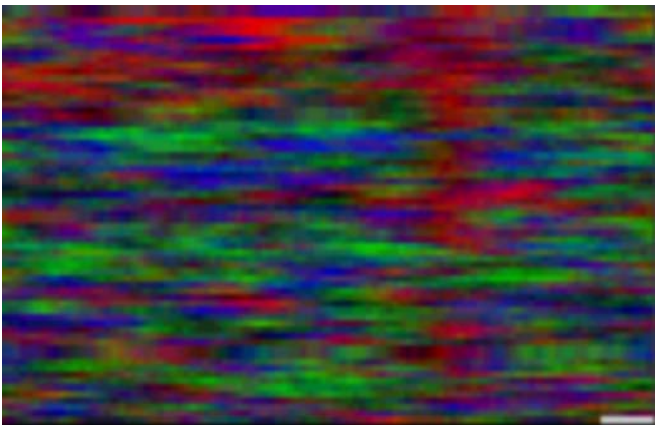

*Рис. 1.* Последовательное заполнение изображения в процессе обозрения компонентов $\langle R,\ G,\ B\rangle$

Этот процесс напоминает действие сглаживающего одномерного фильтра на последовательность, а потом вывод по строкам. На рисунке видна нерегулярная структура ДНК, но, тем не менее, прослеживаются некоторые закономерности.

Регулируя параметры алгоритма (например, ширину изображения или длину окна), можно увидеть некоторые общие участки. Неплохо прослеживается неоднородность GC-состава ДНК. Например, в центре больше зеленого и синего, что указывает на известный факт, что кодирующие области более GC-богаты, чем некодирующие.

### 1.2. Визуализация структуры нуклеотидной последовательности

Пусть дана функция $g : \{A,C,G,T\} \to \{i : 0 \le i \le 255\}$. Тогда наша последовательность $\overline{S}$ создает последовательность целых чисел по следующему правилу

$$g\left[\overline{S}\right] = g(s_1)g(s_2)g(s_3)\ldots$$

Каждые 3 числа, стоящие рядом, могут рассматриваться, как компоненты цвета, т.е. мы имеем следующую последовательность троек

$$\langle g(s_1)g(s_2)g(s_3)\rangle, \langle g(s_4)g(s_5)g(s_6)\rangle, \langle g(s_7)g(s_8)g(s_9)\rangle\ldots =$$
$$= \langle R_1, G_1, B_1\rangle,\ \langle R_2, G_2, B_2\rangle,\ \langle R_3, G_3, B_3\rangle,\ldots$$

Аналогично, двигаясь вдоль последовательности, получаем последовательность изображений.

Также можно рассмотреть другие функции

$$g_2 : \{A,C,G,T\}^2 \to \{i : 0 \le i \le 255\} \text{ или } g_3 : \{A,C,G,T\}^3 \to \{i : 0 \le i \le 255\}.$$

В этом случае рассматриваются пары и тройки на последовательностях. Как известно, они более информативны.

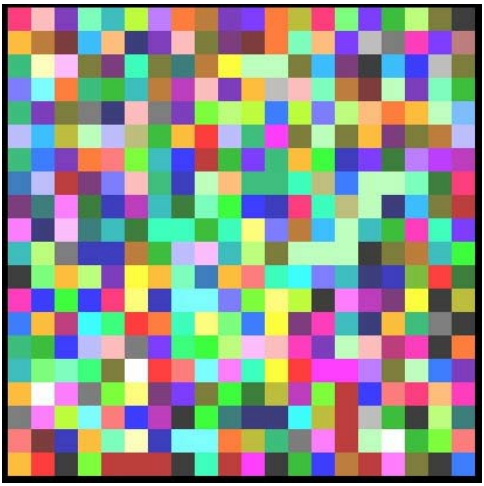

*Рис. 2.* Визуализация функции $g : \{A, C, G, T\} \to \{i : 0 \le i \le 255\}$

Отчетливо видна нерегулярная структура ДНК. По большей части генетическая информация похожа на высокочастотный шум, так что для анализа такой информации целесообразно использовать соответствующие методы высокочастотного анализа в совокупности с методами, опирающимися на реальные экспериментальные данные.

### 1.3. Визуализация в трехмерном пространстве

Рассмотрим последовательность троек, описанную в первом алгоритме, $\langle \overline{p}_A(k),\ \overline{p}_C(k),\ \overline{p}_G(k) \rangle,\ k \ge 1$. Они могут быть представлены координатами в трехмерном пространстве.

Предположим, что дана функция $h : [0,1]^3 \to \{i : 0 \le i \le 255\}^3$. Понятно, что она может быть представлена в виде

$$h(x, y, z) = \langle h_R(x, y, z),\ h_G(x, y, z),\ h_B(x, y, z) \rangle .$$

В итоге получаем трехмерное изображение, которое позволяет лучше увидеть структуру последовательности $\overline{S}$. Были использованы различные формы функции визуализации.

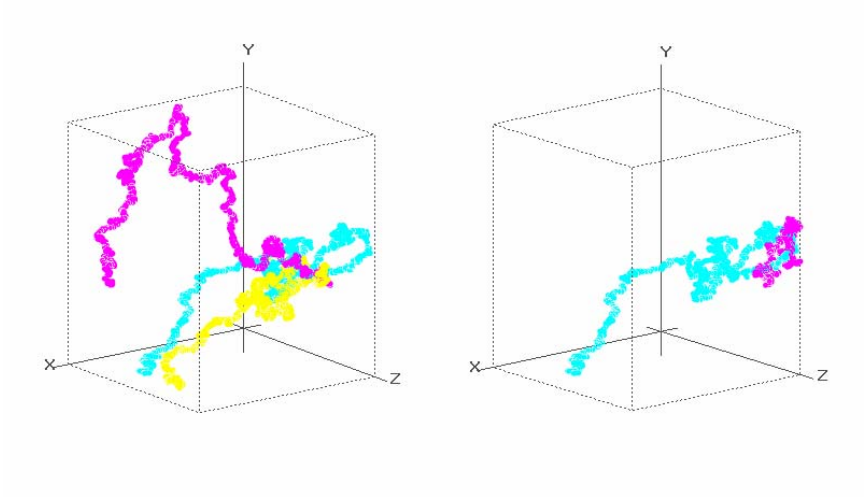


*Рис. 3.* Визуализация промоторов с помощью функции $h : [0,1]^3 \to \{i : 0 \le i \le 255\}^3$

Слева изображены три промотора для одного и того же гена *c-myc* у разных организмов: человека, мыши и крысы. Видна схожесть в поведении этих трех кривых. Справа — промоторы разных генов *a` actin* и *c-myc* у человека. Видно, что поведение кривых различается. В данном случае координатами точек являются наши $p_A$, $p_C$, $p_G$. Кубы на рисунках — единичные.

## 2. ВИЗУАЛЬНЫЙ АНАЛИЗ ВЫБОРКИ ПРОМОТОРОВ И РАСПОЗНАВАНИЕ ССТФ

Далее можно проанализировать промоторы [2]. В силу зависимости между сайтами связывания транскрипционных факторов (ССТФ), относящихся к похожим транскрипционным факторам, мы использовали специальную формулу для вычисления степени похожести промоторов, затем отсортировали их и выявили наиболее статичные участки.

Мы взяли выборку промоторов генов, специфично экспрессирующихся в печени ( $\overline{S}^1 \dots \overline{S}^K$ ) одинаковой длины L = 100 и рассчитали попарную похожесть при помощи нашей метрики. Относительное положение промоторов мы вычисляли, пользуясь известным положением старта транскрипции.

Потом находим путь ($p_1..p_{K-1}$) со свойством $\sum_{i=1}^{K-1} sim\left(\overline{S}^{p_i}, \overline{S}^{p_{i+1}}, L\right) \to \max$, где $sim$ — похожесть между $p_i$ и $p_{i+1}$.

Применяя правило ближайшего соседа, получаем приближенное решение. Потом на всех парах последовательностей $\overline{S}^{p_i}, \overline{S}^{p_{i+1}}$ мы ищем T непересекающихся фрагментов ($B_1 \ldots B_T$) длины $P < L$ с максимальной похожестью $sim^*(i,j) = sim\left(\overline{S}^{p_i}, \overline{S}^{p_{i+1}}, j, P\right)$, где $i$ — номер последовательности в полученном пути ($p_1..p_{K-1}$) и $j$ — старт фрагмента. На следующем рисунке показана визуализация этих данных.

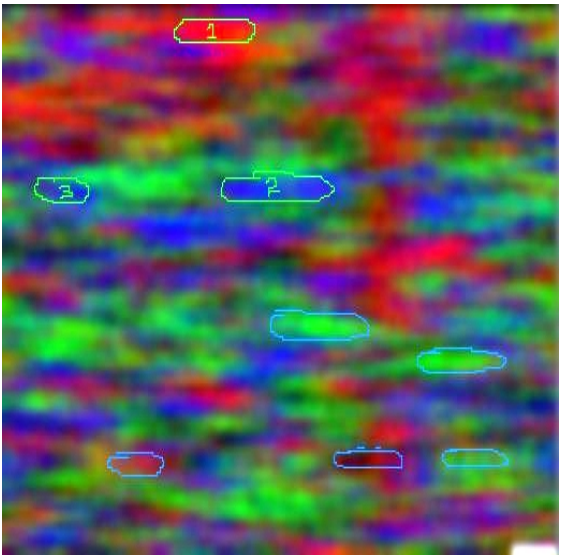


*Рис. 4.* Визуализация полученных результатов

Взята выборка промоторов генов, специфично экспрессирующихся в печени. После их сортировки с целью минимизации описанного выше функционала мы визуализировали их, используя первый алгоритм визуализации. Обведенные участки соответствуют высокогомологичным областям.

## 3. КРАТНОМАСШТАБНЫЙ АНАЛИЗ

Кратномасштабный анализ представляет собой широко известный математический метод, базирующийся на применении вейвлет-преобразования и позволяющий, в частности, эффективно исследовать одномерные сигналы [8].

В зависимости от конкретного приложения, исходные данные могут быть представлены в различных форматах. С другой стороны, для выполнения преобразования наиболее удобным является формат представления данных, при котором отсчеты записаны последовательно в виде чисел с плавающей запятой в двоичном формате. В таком случае становится возможным выполнять вычисления непосредственно после чтения фрагмента файла.

Теперь рассмотрим подготовку к обработке нуклеотидной последовательности. Нуклеотидная последовательность является, по сути, словом, состоящим из букв «генетического алфавита» — нуклеотидов A, C, T и G. Очевидно, такое представление малопригодно для какого-либо численного анализа, поэтому выполняется преобразование последовательности нуклеотидов к одномерному массиву чисел. В простейшем случае каждой букве алфавита сопоставляют число (иногда буквы группируют по две или по три, такой метод называется методом простого сопоставления). Полученная последовательность чисел уже может быть рассмотрена в качестве исходных данных для применения численных методов. Преобразованные данные записываются в файл в естественном формате, который после исчерпания данных в исходном файле дополняется нулями до оптимального размера, зависящего от конкретного вейвлета, применяемого в данный момент.

Для анализа данных используется видоизмененное быстрое вейвлет-преобразование, опирающееся на метод кратномасштабного анализа, разработанного Малла и Мейером, известного также, как пирамидальный алгоритм Малла. Были реализованы несколько вычислительных модулей, представляющих различные классы вейвлетов (вейвлеты Добеши: DB4, DB6, DB8; вейвлеты Хаара).

Результат вычислений — несколько векторов, являющихся приближениями одного и того же исходного вектора. Применительно к исследованиям нуклеотидных цепочек существуют несколько методов визуализации информации подобного рода. В настоящей работе был выбран наиболее наглядный, с точки зрения авторов, способ, который заключается в следующем (ниже показано главное окно программы).

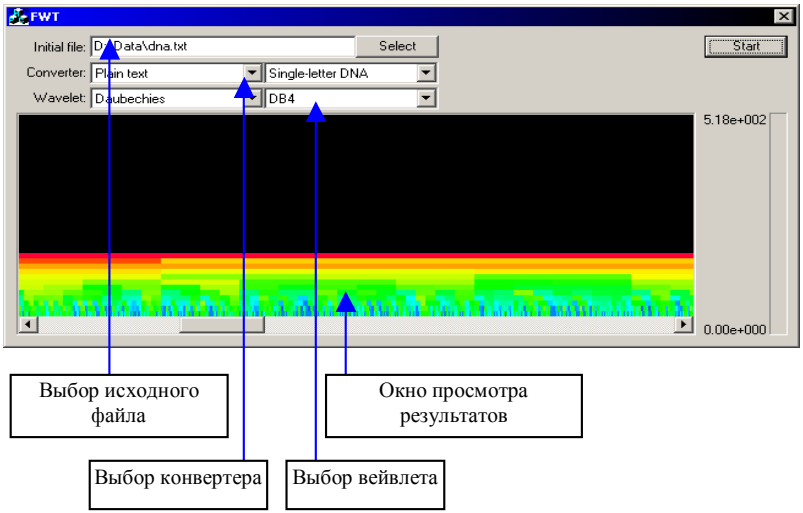


*Рис. 5.* Результат применения вейвлет-преобразования DB4

Среди всех значений, содержащихся в полученных массивах, выбирается минимальное и максимальное значения. После этого строится цветовая шкала соответствия значения оттенку цвета H в системе цветовых координат HSV. Минимальному значению соответствует цвет с оттенком 0, максимальному — с оттенком 360. После этого массивы отображаются на плоскости рядами цветных точек; цвет точки соответствует значению элемента массива. Такой способ отображения дает возможность визуально выделять характерные участки в массиве данных.

Реализованная программа позволяет работать с файлами объемом до 300 Мбайт. Проведенные предварительные исследования показали, что визуализация результатов вейвлет-преобразования, примененного к сигналам, ассоциированным с генетической последовательностью, может оказаться значительно более информативным методом визуализации, по сравнению с рассмотренными ранее.

## СПИСОК ЛИТЕРАТУРЫ


1. **Doolittle R. F.** Microbial genomes opened up // Nature. — 1997. — Vol. 392. — P. 339–342.
2. **Maley L. E., Marshall C. R.** The coming of age of molecular systematics // Science. — 1998. — Vol. 279. — P. 505–506.
3. **Ulyanov A., Stormo G.** Multi-alphabet consensus algorithm for identification of low specificity protein-DNA interactions // Nucleic Acids Res. — 1995. — Vol. 23. — P. 1434–1440.
4. **Kel A.E., Kondrakhin Y.V., Kolpakov Ph.A., Kel O.V., Romashenko A.G., Wingender E., Milanesi L., Kolchanov N.A.** Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences // Proc. Third Internat. Conf. Intelligent Systems Molec. Biol. — 1995. — P.197–205.
5. **Jeffrey H. J.** Chaos game representation of gene structure // Nucleic Acids Res. — 1990. — Vol. 18. — P. 2163–2170.
6. **Burma P. K., Raj A., Deb J.K., Brahmachari S. K.** Genome analysis: a new approach for visualization of sequence organization in genomes // J. Biosci.. — 1992. — Vol. 17. — P. 395–411.
7. **Solovyev V. V.** Fractal graphical representation and analysis of DNA and protein sequences // Biosystems. — 1993. — Vol. 30. — P. 137–160.
8. **Астафьева Н. М.** Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук. — 1998. — Т. 166, № 11. — С. 1145–1170.