Письма в ЖЭТФ, том 16, вып. 3 стр. 183 – 186 *5 августа 1972 г.*

ОБРАТНАЯ ЗАДАЧА ДЛЯ ДНК

М. Я. Азбель

1. Для изучения характера записи информации в ДНК важно выяснить закономерность следования в ней нуклеотидов. В ДНК большая поверхностная свободная энергия V на границе между длинными расплавленными и спиральными состояниями (рс и сс) позволяет сделать это по кривой плавления ДНК со скрепками. При $V = \infty$ плавлению ДНК соответствовал бы фазовый переход 1 рода при температуре $T_п$, вблизи которой $E_\sigma = \sigma a_\sigma (- T_п + T)$, где E_σ – свободная энергия на одно звено; $\sigma = +1, -1{,}0$ отвечают сс, рс и точке плавления, а a_σ и $T_п$ зависят от состава ДНК; независящее от σ слагаемое в E_σ опущено. При большой V в начале плавления, при $T > T_{AT}$ [1] (где T_{AT} – температура плавления гомо-AT), существенны (длинные) участки, настолько обогащенные легкоплавкими AT-парами, что данная T совпадает с их температурой плавления $\tilde{T}$, при которой (с учетом поверхностной энергии) равны свободные энергии рс и сс. (Назовем эти участки п-участками – пу, а это их состояние – пс. Все остальные участки находятся ниже "своей" температуры плавления и, с точностью до энтропийного слагаемого – см. ниже – спиральны). Вероятность такой, немалой флуктуации состава с ростом длины участка быстро падает, поэтому в основном приближении (оп) крылья кривой плавления определяются непосредственно участками определенной длины и состава. Это и позволяет восстанавливать их вероятность. Существует, однако, предельная точность подобного восстановления (определяемая относитель-

[1] АТ и ГЦ – пары азотистых оснований аденин – тимин и гуанин – цитозин соответственно; „скрепки" – низкомолекулярные примеси в растворе (см. [1]).

ным удалением от $T_п$, т. е. величиной $1/V$), с которой термодинамические величины не зависят от детального вида последовательности нуклеотидов, а решение обратной задачи устойчиво.

2. Будем поэтому рассматривать лишь крылья кривой (для конкретности, при $T < T_п$) и лишь с указанной точностью. На крыльях (см. выше) изменение полной свободной энергии E' (на одно звено) по сравнению с $\tilde{E}$ для "чистого" сс ДНК связано только с появлением пу. Если вероятность пу длины λ есть w_λ, причем $w = \Sigma w_\lambda << 1$, то в оп такие участки независимы, не пересекаются и образуют на ДНК слабый "твердый" (так как последовательность нуклеотидов неизменна) раствор. (Так как $\lambda >> 1$, то $w_\lambda << 1$ всегда, даже при $T_п$). Значит, при $m_\lambda w_\lambda << |\ln w|$ (m_λ – характерная длина корреляции на участке λ, w_λ^{-1} – среднее расстояние между участками λ), с относительной точностью w

$$E = E - \tilde{E} = \Sigma w_\lambda (E_\lambda^п - E_\lambda^с), \qquad (1)$$

где $E_\lambda^{п, с, р}$ отвечают пс, сс, рс для пу длины λ. Поверхностные члены (см. выше) включены в $E_\lambda^р$ (входящую в $E_\lambda^п$).Так как выплавление между сс, участка длины λ обуславливает (на звено), помимо поверхностной энергии $- 2V/\lambda$ (зависимостью которой от состава рс и сс с точностью λ^{-1} пренебрегаем), неаддитивную энтропию $- S(\lambda) \sim - \ln \lambda$ (см. [1]), то $E_\lambda^с = E_\lambda^р$ при температуре плавления $T_\lambda = T - 2\tilde{V}[\lambda(a_+ + a_-)]^{-1} \equiv$ $\equiv T - r/\lambda$; $\tilde{V} = V + S(\lambda)/2$, причем $\lambda \geqslant \lambda_0$, где $T_{\lambda_0} = T_{AT}$. Состав пу определяется $T_\lambda < \tilde{T} < T_{\lambda+1}$. Поэтому, если $c(\lambda, \tilde{T})\, d\tilde{T}$ – вероятность на участке длины λ состава, обеспечивающего температуру плавления в интервале $d\tilde{T}$, то $w_\lambda = c(\lambda, T_\lambda) r/\lambda^2$. Так как $w_\lambda << 1$, то $E_\lambda^п$ достаточно записать в нулевом приближении, т. е. при $V \to \infty$, а значит, и при $\lambda \to \infty$ (для $E_\lambda^с$ это также возможно, но нецелесообразно). В результате находим:

$$- E' = \int_0^\infty a_+ r^2 c(\lambda, T_\lambda) \frac{d\lambda}{\lambda^3} \sim c(\lambda^*, T^*); \quad T^* = T - \frac{r(\lambda^*)}{\lambda^*};$$

$$\lambda^* = \max(\lambda_0, \bar{\lambda}); \quad dc/d\bar{\lambda} = 0 \qquad (2)$$

Точность этой формулы, в связи с переходом $\lambda^* \to \infty$ в $E_\lambda^п$, определяется $\beta = m_{\lambda^*}/\lambda^*$. Если $\beta \gtrsim 1$ и $c(\lambda, \tilde{T})$ быстро убывает с ростом $\tilde{T} - T_\lambda$, E' находится по порядку величины, а $\ln|E'|$ – с относительной точностью $|\ln|E'||^{-1}$. (В более общем случае $E' \sim \max_\lambda \int_{T_\lambda}^{T_{AT}} (\tilde{T} - T_\lambda) c(\lambda, \tilde{T})\, d\tilde{T} = f(\lambda_m, T_{\lambda_m})$ – ср. с (2)). Так как $w \lesssim 1$, такая точность сохраняется при любых T. (В частном случае линейной зависимости E_σ от концентрации AT-пар, при $\bar{\lambda} > \lambda_0$ для случайной последовательности с этой точностью получаются результаты [2], и $\ln E' \sim T - T_п$). При $V \to \infty$ существенны $\lambda \to \infty$, что отве-

чает случайной последовательности и экспоненциальному стремлению $c(\lambda, x_\lambda)$ к нулю. Поэтому естественно положить $c(\lambda, x_\lambda) = \exp\{-\phi(\lambda, x_\lambda)\}$; это позволяет вычислить предэкспоненциальный множитель в формуле для E'.

3. Пусть $E' = E'(T, \tau)$ известна. Тогда, учитывая определение $\bar{\lambda}$, находим физически ясную формулу $\lambda' = -[\partial E'/\partial T][\partial E'/\partial \tau]^{-1} = h(T, \tau)$. Разрешая это уравнение совместно с $\tilde{T} = T - \tau(h)/h$ относительно T и τ как функций $\bar{\lambda}$ и $\tilde{T}$, найдем в нулевом приближении $c(\lambda, \tilde{T})$, как результат решения алгебраической задачи. Дальнейшие члены разложения E' дадут следующие приближения для c. Само $\tilde{T}$ определяется только сс и рс и известным образом выражается, в зависимости от точности определения, через концентрации элементов (AT и $ГЦ$), пар соседних элементов и т. д. Поэтому только точность измерений лимитирует, при указанных ограничениях, детальность получаемой о структуре ДНК информации. Выяснение же вида функции $E''(T, \tau)$ сводится к замене переменных в экспериментально найденной зависимости свободной энергии ДНК от T и концентрации ξ низкомолекулярных примесей – "скрепок" – в растворе, поскольку скрепки, имеющие всюду в ДНК одинаковый химический потенциал, совпадающий с их химическим потенциалом $\mu(T, \xi)$ в растворе, очевидным образом перенормируют a_+ и a_-; для невзаимодействующих скрепок

$$E_{\sigma\alpha} \to E_{\sigma\alpha} - \ln[1 + \exp(\mu - \mathcal{E}_{\sigma\alpha})/T] \qquad (3)$$

($\mathcal{E}_{\sigma\alpha}$ – энергия, привнесенная скрепкой в звено типа α). Для получения большей информации о структуре ДНК нужно учесть зависимость энергии взаимодействия соседних пар от их вида, и, значит, использовать соответствующее число различных скрепок одновременно. При $\bar{\lambda} < \lambda_0$ нахождение $c(\lambda, T_{AT})$ очевидно и не требует использования скрепок. Метод весьма чувствителен к структуре ДНК. Так, если $c(\lambda, T_{AT}) = 0$ при $\lambda > \lambda_g$, то при $\lambda_0 > \lambda_g$ происходит резкое падение $|E'|$.

Наличие блочной структуры [1] легко учитывается введением функции распределения блоков по концентрациям и приводит лишь к замене T_λ на его наименьшее по блокам значение. Для решения обратной задачи удобно, конечно, сначала произвести разделение ДНК на блоки одного типа и именно для них производить измерения, используя для изучения характера записи информации ансамбль одинаковых блоков, взятых из ДНК различных особей одного биологического вида, рода и т. д.

5. Учтем теперь энтропийную добавку к E', которая при $T < T_{AT}$ (и $T > T_{ГЦ}$) является единственной. Она определяется (так как V велико) малым числом коротких р-участков. В области, где энергетический выигрыш отсутствует или пренебрежимо мал, р-участки редко сталкиваются и не стремятся (как при $T_{AT} < T$) объединяться в длинные участки. Поэтому вероятности $q_{\lambda m} = c_{\lambda m}\exp\{(\lambda E - E_{\lambda p})T^{-1}\}$ р-участков длины $\lambda \geqslant 1$ с концентрациями $x = m/\lambda$ ($E_{\lambda p}$ отвечает рс) можно считать независимыми, а вероятность (заведомо очень длинного) с-участка равной $q_s = \exp(+E'/T)$. Отсюда $\Sigma q_{\lambda m} q_s = 1 - q_s$ (для гомополимера, где есть только энтропийная добавка, q_λ всегда

строго независимы, и полученная формула является точной). По $E' \sim \exp(-2V)$ (ср. с пунктом 3) можно восстановить $|\ln|E'||$ первых коэффициентов $c_{\lambda m}$. В сущности, именно возможность при $V \gg 1$ разделить чисто "энтропийную" и чисто "энергетическую" области обуславливает конструктивность решения.

6. Необходимость устойчивости линейного считывания информации в ДНК обуславливает, вероятно, значительную избыточность. Отсюда может следовать существование "морфем" – отрезков, на которых реализуются максимумы $c_{\gamma_1\gamma_2\ldots\gamma_k}/c_{\gamma_1\ldots\gamma_{k-1}}$ и $c_{\gamma_1\ldots\gamma_k}/c_{\gamma_2\ldots\gamma_k}$ ($c_{\gamma_1\ldots\gamma_k}$ – концентрация последовательности $\gamma_1\ldots\gamma_k$) при локально наибольшем значении k. При достаточной избыточности их можно выделить, изучая кривую плавления. (В "идеальном случае" $c_{\lambda m} = c_{\lambda+1, m+1}$ или $c_{\lambda m} = c_{\lambda+1, m}$). Изучение достаточно больших λ позволило бы выяснить вероятность соседства разных морфем. Возникающие аналогии с языковым текстом обсуждены в [3]. В связи со сказанным представляет интерес машинный расчет кривой плавления последовательности пар, задаваемой записанным в двоичной системе текстом достаточно длинной книги, и исследование обратной задачи для этого случая.

Я глубоко признателен И.М.Лифшицу за исключительно полезные дискуссии. Я благодарен также А.И.Ларкину и М.Д.Франк-Каменецкому.

Институт теоретической физики
им. Л.Д.Ландау
Академии наук СССР

Поступила в редакцию
19 июня 1972 г.

Литература

[1] А.А.Веденов, А.М.Дыхне, М.Д.Франк-Каменецкий. УФН, **105**, 479, 1971.

[2] А.А.Веденов, А.М.Дыхне. ЖЭТФ, **55**, 357, 1968.

[3] M. Ya. Azbel. Ideen des Exakten Wissens, №12, 1972.