АКАДЕМИЯ НАУК СССР
НАУЧНЫЙ ЦЕНТР БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ
НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

ПРЕПРИНТ

В.Ю. ЛУНИН

ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИИ О РАСПРЕДЕЛЕНИИ ЗНАЧЕНИЙ ЭЛЕКТРОННОЙ ПЛОТНОСТИ В БЕЛКАХ. I. ВОССТАНОВЛЕНИЕ НЕДОСТАЮЩИХ СТРУКТУРНЫХ ФАКТОРОВ

ПУЩИНО · 1986

УДК 548.73

Предложен новый тип информации о распределении электронной плотности в кристаллах биологических макромолекул — квазигистограмма изображения функции распределения электронной плотности при конечном разрешении. Показано, как такая информация может быть использована для восстановления значений низкоугловых структурных факторов, модули которых не были измерены в процессе рентгеновского эксперимента.

Введение

В процессе рентгеноструктурного исследования пространственной организации макромолекул при конечном разрешении d_{min} ищется функция $\varrho(\vec{r})$, являющаяся суммой отрезка ряда Фурье:

$$\varrho(\vec{r}) = \frac{1}{V} \sum_{|\vec{s}| \leqslant 1/d_{min}} F(\vec{s}) e^{i\varphi(\vec{s})} e^{-2\pi i(\vec{s},\vec{r})}. \qquad (1)$$

Мы будем называть эту функцию $\varrho(r)$ «изображением» (или более полно, «изображением функции распределения электронной плотности при разрешении d_{min}»), чтобы подчеркнуть ее отличие от «истинного» распределения электронной плотности, отвечающего бесконечному ряду в (1). На практике точное определение изображения затруднено по двум причинам. Во-первых, фазы $\varphi(\vec{s})$ структурных факторов определяются с ошибкой (некоторые фазы могут быть вообще не определены), во-вторых, из-за конкретных условий рентгеновского эксперимента часть модулей $F(\vec{s})$ структурных факторов может оказаться не измеренной. В данной работе мы остановимся на ошибках в определении функции $\varrho(\vec{r})$, вызванных отсутствием информации о значениях модулей части структурных факторов. Эффект исключения из суммы (1) около 18% слагаемых продемонстрирован на рис. 1. (Набор исключенных слагаемых определялся конкретным экспериментом; в основном это были отражения центральной зоны.)

Пусть S_d — обозначает множество индексов $\vec{s}$, для которых известны значения структурных факторов $F^0(\vec{s})e^{i\varphi^0(\vec{s})}$, а S_u — обозначает множество индексов, для которых модули структурных факторов в (1) неизвестны. Обозначим L — множество всевозможных изображений $\varrho(\vec{r})$, имеющих предписанные значения структурных факторов $F^0 e^{i\varphi^0}$ для $\vec{s} \in S_d$ и произвольные структурные факторы с $\vec{s} \in S_u$. Существуют различные подходы к доопределению неизвестных структурных факторов, т.е. к выбору конкретного изображения из класса L. Наиболее распространенный — положить неизвестные структурные факторы

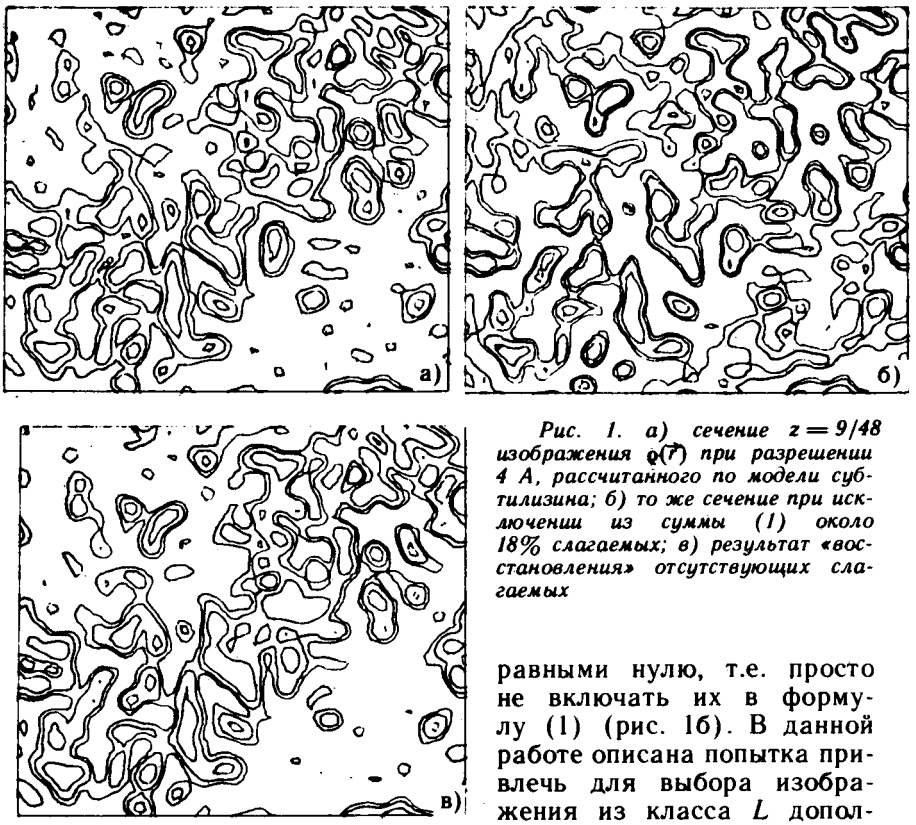


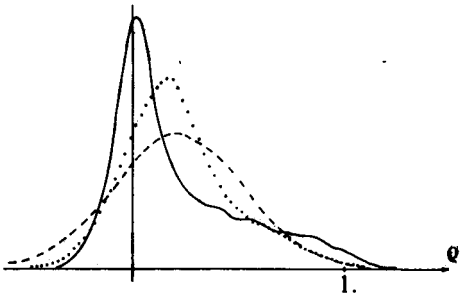
Рис. 1. а) сечение $z = 9/48$ изображения $\varrho(\vec{r})$ при разрешении 4 А, рассчитанного по модели субтилизина; б) то же сечение при исключении из суммы (1) около 18% слагаемых; в) результат «восстановления» отсутствующих слагаемых

равными нулю, т.е. просто не включать их в формулу (1) (рис. 1б). В данной работе описана попытка привлечь для выбора изображения из класса L дополнительную информацию, характеризующую изображения распределений электронной плотности в белках. Результат такого рода попытки иллюстрирован рис. 1в.

§ 1. Гистограмма, отвечающая изображению функции распределения электронной плотности, как вид дополнительной информации об объекте

1. В этой работе мы остановимся на простейшем свойстве изображений распределения электронной плотности в белках — наличии характерной гистограммы. Зададимся вопросом, какие значения и с какой частотой может принимать функция вида (1)? Пусть $\varrho(\vec{r})$ рассчитана в узлах некоторой сетки в элементарной ячейке и $\{\varrho_j\}$ — совокупность ее значений в этих узлах. Построим гистограмму для этих значений, т.е. разобьем интервал $(\varrho_{min}, \varrho_{max})$ на K равных частей (бинов) и определим частоты попадания значений ϱ_j в каждый из бинов:

Рис. 2. Гистограммы для функций: ϱ_e (——), ϱ_{sp} (— —), $\tilde{\varrho}$ (. . . .)



$$\nu_k = \frac{n_k}{\sum n_k}; \quad k = 1, \ldots, K.$$

Здесь n_k — число значений ϱ_j, попадающих в бин с номером k, т.е. таких, что

$$\varrho_{min} + (j-1)\frac{\varrho_{max} - \varrho_{min}}{K} \leqslant \varrho_j < \varrho_{min} + j\,\frac{\varrho_{max} - \varrho_{min}}{K}.$$

Анализ гистограмм для изображений уже известных белков приводит к выводу, что они имеют характерные особенности, отличающие их от гистограмм «случайно выбранных» функций вида (1). Этот факт иллюстрируется на рис. 2. На этом рисунке приведены гистограммы для трех функций вида (1), полученных следующим образом:

$\varrho_e(\vec{r})$ — определена по формуле (1), где в качестве $F(\vec{s})e^{i\varphi(\vec{s})}$ взяты точные структурные факторы, рассчитанные по атомной модели субтилизина;

$\varrho_{sp}(\vec{r})$ — в качестве $F(\vec{s})$ взяты точные значения, рассчитанные по модели субтилизина, а $\varphi(\vec{s})$ определены датчиком случайных чисел;

$\tilde{\varrho}(\vec{r})$ — из синтеза $\varrho_e(\vec{r})$ исключено около 18% рефлексов.

Для разных белков гистограммы для точных изображений $\varrho_e(\vec{r})$ могут несколько различаться и определение «эталонной» гистограммы $\{\nu_k^0\}$ для еще неизвестного объекта представляет собой отдельную задачу. Мы не будем в этой работе на ней останавливаться, ограничась важным частным случаем, когда эталонная гистограмма может считаться известной. Это случай, когда нам известна структура гомологичного белка, гистограмма для которого может быть взята в качестве эталона. Следует также отметить, что речь всегда идет о гистограмме для изображения распределения электронной плотности при конкретном разрешении d_{min}. Переход к изображению при другом разрешении меняет «эталонные» частоты $\{\nu_k^0\}$.

2. Допустим теперь, что нам известна для исследуемого объекта эталонная гистограмма $\{\nu_k^0\}$. Мы применим эту информацию для выбора функции из введенного ранее класса L. Пусть $\varrho^c(\vec{r})$ — некоторая функция из этого класса (т.е. она имеет заданные структурные факторы $F^0e^{i\varphi^0}$ для $\vec{s} \in S_d$, а структурные факторы для $\vec{s} \in S_u$ определены произвольно). Мы можем рассчитать гистограмму $\{\nu_k^c\}_{k=1}^K$ для функции $\varrho^c(\vec{r})$ и сравнить ее с эталонной:

$$Q(\varrho^c) = \frac{1}{K}\sum_{k=1}^{K}\frac{(\nu_k^c - \nu_k^0)^2}{\nu_k^0}.$$

Сформулируем критерий выбора функции $\varrho(\vec{r})$ из класса L следующим образом:

КРИТЕРИЙ 1. Среди всех функций класса L найти ту, для которой величина $Q(\varrho)$ минимальна.

Это означает, что мы разрешаем неизвестным структурным факторам принимать любые (не обязательно нулевые) значения, но требуем при этом, чтобы гистограмма получаемой функции $\varrho(\vec{r})$ была по возможности наиболее близка к эталонной.

Результат применения такого типа критерия иллюстрирован рис. 1в.

§ 2. «Вычислительная» постановка задачи квазигистограммы

1. Итак, мы хотим определить вещественные и мнимые части структурных факторов $f_R(\vec{s})$, $f_I(\vec{s})$ (для $\vec{s}\in S_u$) таким образом, чтобы была минимальна величина

$$Q = \frac{1}{K}\sum_{k=1}^{K}\frac{(\nu_k^c - \nu_k^0)^2}{\nu_k^0},$$

где $\{\nu_k^0\}$ — заданные величины, а величины $\{\nu_k^c\}$ вычисляются следующим образом:

а) вычисляются значения (на некоторой сетке в элементарной ячейке) функции $\varrho(\vec{r})$:

$$\varrho_j^c = \varrho^c(\vec{r}_j) = \frac{1}{V}\Big\{\sum_{s\in S_d} F^0(\vec{s})e^{i\varphi^0(\vec{s})}e^{-2\pi i(\vec{s},\vec{r})} + \\ + \sum_{\vec{s}\in S_u}[f_R(\vec{s}) + if_I(\vec{s})]e^{-2\pi i(\vec{s},\vec{r})}\Big\}; \qquad (2)$$

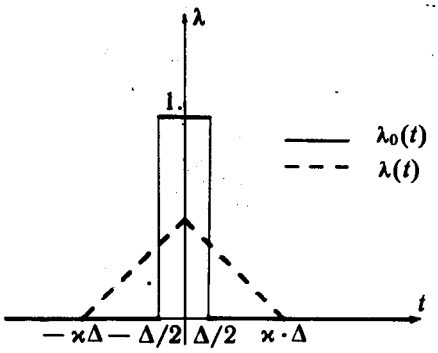
б) определяются частоты попадания величин ϱ_j^c в заданные бины:

$$\nu_k^c = \frac{1}{N}\sum_{j=1}^{N}\lambda^0(\varrho_j^c - t_k), \qquad (3)$$

где N — общее число узлов сетки, t_k — середина k-го бина ($t_k = \varrho_{min} + (k - 1/2)\cdot\Delta$), Δ — длина бина ($\Delta = (\varrho_{max} - \varrho_{min})/K$),

$$\lambda^0(t) = \begin{cases} 1, & \text{если } |t| \leqslant (\Delta/2), \\ 0, & \text{если, } |t| > (\Delta/2). \end{cases}$$

Рис. 3. Функции $\lambda^0(t)$ и $\lambda(t)$



Минимизация определенной таким образом функции Q представляется весьма сложной вычислительной задачей. В первую очередь это связано с тем, что функция Q является кусочно постоянной — небольшие изменения структурных факторов не приводят, вообще говоря, к «переходу» значений ϱ_j из одного бина в другой, т.е. к изменению частот ν_k^c. Это не позволяет использовать для минимизации функции Q никакие градиентные методы (градиент Q равен нулю почти при всех значениях переменных). Чтобы обойти эту сложность, мы несколько изменим постановку задачи, введя понятия квазичастот и квазигистограмм.

2. «Плохие» свойства функции Q вызваны тем, что при расчете частот ν_k по формуле (3) используется ступенчатая функция $\lambda^0(t)$.

ОПРЕДЕЛЕНИЕ. Пусть $\lambda(t)$ — произвольная функция такая, что $\lambda(t) \geqslant 0$ и $\int_{-\infty}^{\infty} \lambda(t)dt = \Delta$. Будем называть квазичастотами величины

$$\hat{\nu}_k = \frac{1}{N} \sum_{j=1}^{N} \lambda(\varrho_j - t_k). \tag{4}$$

Совокупность квазичастот $\{\hat{\nu}_k\}$ будем называть квазигистограммой.

Главный смысл введения квазичастот — вклад отдельной точки ϱ_j распределяется теперь по нескольким бинам, создавая зависимость квазичастот $\hat{\nu}_k$ не только от количества значений ϱ_j, попавших в данный бин, но и от количества значений, попавших в соседние бины. При этом эту зависимость можно сделать гладкой, что делает возможным применение градиентных методов для минимизации функционала

$$\hat{Q} = \frac{1}{K} \sum_{k=1}^{K} \frac{(\hat{\nu}_k^c - \hat{\nu}_k^0)^2}{\hat{\nu}_k^0}. \tag{5}$$

Здесь $\hat{\nu}_k^0$ — эталонные квазичастоты (определяемые в нашем рассмотрении из анализа изображения $\varrho(\vec{r})$ для гомологичного белка), а величины $\hat{\nu}_k^c$ вычисляются по формулам (4) и (2) через вещественные и мнимые части $f_R(\vec{s})$, $f_I(\vec{s})$ «свободных» структурных факторов.

Таким образом, окончательная постановка задачи восста-

новления изображения $\varrho(\vec{r})$ по неполному набору структурных факторов такова:

ЗАДАЧА 1. Среди всех функций класса L найти такую функцию $\varrho^*(\vec{r})$, для которой величина $\hat{Q}(\varrho^*)$ минимальна (т.е. квазичастоты которой наиболее близки к эталонным).

В этой работе мы будем использовать простейшую функцию $\lambda(t)$ вида (рис. 3):

$$\lambda(t)=\begin{cases} -\frac{1}{\varkappa^2\Delta}|t|+\frac{1}{\varkappa} & \text{при } |t|<\varkappa\Delta, \\ 0 & \text{при } |t|\geqslant\varkappa\Delta. \end{cases} \qquad (6)$$

§ 3. Тестирование метода

В качестве тестового объекта взята атомная модель белка субтилизина, которая была размещена в элементарной ячейке размерами 73×64×48 Å в пространственной группе $P\,2_12_12_1$. При этом ориентация молекулы была взята произвольно, а положение центра масс было выбрано таким образом, чтобы молекулы не налезали друг на друга. По координатам атомов рассчитаны структурные факторы и построен синтез разрешения 4 Å (рис. 1а). По этому синтезу определены эталонные квазичастоты (отрезок (—0.5, 1.5) был разбит на 30 бинов, использовалась функция $\lambda(t)$ вида (6) с $\varkappa=5$).

Для того, чтобы проверить, насколько квазигистограмма чувствительна к изменению ориентации молекулы, был проведен аналогичный расчет для молекулы субтилизина, помещенной в другой ориентации. Введенный выше формулой (5) критерий соответствия двух квазигистограмм принял значение 0.1×10^{-4}, что определило предел, до которого имеет смысл опускать значение минимизируемого критерия при решении задачи 1.

Далее была проиммитирована ситуация отсутствия информации о части модулей структурных факторов. Около 18% структурных факторов были объявлены неизвестными и была поставлена задача определения их путем минимизации функции (5). В качестве стартовых значений для неизвестных структурных факторов были взяты нулевые значения, что отвечает синтезу, изображенному на рис. 1б. Стартовое значение минимизируемого критерия (5) составило 0.33×10^{-2}.

Далее была проведена минимизация критерия $\hat{Q}$ методом градиентного спуска. Для этой цели была создана специальная программа минимизации, предназначенная для решения задачи 1. В результате 5 шагов спуска по антиградиенту значение минимизируемого критерия снизилось до 0.14×10^{-4}. Изображение $\varrho^*(\vec{r})$, рассчитанное с использованием определен-

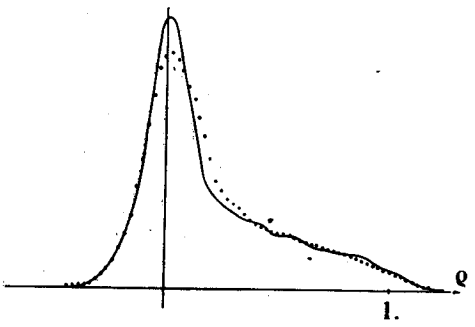


Рис. 4. Гистограммы для функций $\varrho_c(\vec{r})$ (——) *и* $\varrho^*(\vec{r})$ (. . .)

ных в процессе минимизации «неизвестных» структурных факторов, иллюстрировано рис. 1в. На рис. 4 приведены гистограммы для «точного» и «восстановленного» в результате решения задачи 1 синтезов. (Гистограмма «стартового» синтеза дана на рис. 2).

Приведем некоторые формальные критерии, отражающие точность определения «неизвестных» структурных факторов. Значение фактора расходимости составило:

$$\frac{\sum |F^\circ - F^*|}{\sum |F^\circ|} = 0.50,$$

где F° — точные значения модулей структурных факторов, F^* — значения, полученные в результате минимизации, сумма взята по всем 352 рефлексам, считавшимся неизвестными. Средняя ошибка определения фаз по 170 нецентросимметричным рефлексам составила 36°, из 182 центросимметричных рефлексов неверно был определен знак у 34 рефлексов.

§ 4. Оценка точности выделения области по искаженным изображениям

Определенные каким-либо образом изображения $\varrho(\vec{r})$ распределения электронной плотности используются на разных стадиях расшифровки структуры белка для определения положения молекулы, хода полипептидной цепи, положения боковых групп. Во всех этих случаях в элементарной ячейке выделяется и затем изучается область

$$\Omega_{\varrho^*} = \{r : \varrho(r) \geqslant \varrho^*\}$$

с некоторым значением уровня ϱ^*. Таким образом, качество полученного приближенного изображения $\tilde{\varrho}(\vec{r})$ определяется прежде всего тем, насколько точно анализ функции $\tilde{\varrho}(\vec{r})$ позволяет воспроизвести области Ω_{ϱ^*} для «точного» изображения $\varrho(\vec{r})$. Мы введем критерий, характеризующий точность выделения областей Ω_{ϱ^*} следующим образом.

Пусть $\varrho(\vec{r})$ — «точное» изображение (1), а $\tilde{\varrho}(\vec{r})$ — некоторое его приближение. Для заданного значения $t \in (0, 1)$ определим критические значения ϱ^* и $\tilde{\varrho}^*$ таким образом, чтобы области

$$\Omega^t = \{r : \varrho(r) \geqslant \varrho^*\} \quad \text{и} \quad \tilde{\Omega}^t = \{r : \tilde{\varrho}(r) \geqslant \tilde{\varrho}^*\}$$

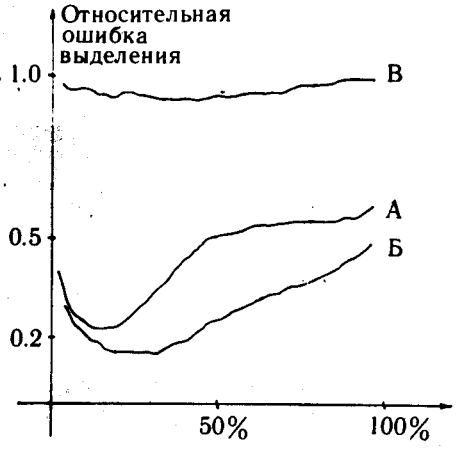


Рис. 5. Относительная точность выделения областей для функции $\varrho_e(\vec{r})$: А — по функции $\tilde{\varrho}(\vec{r})$; Б — по функции $\varrho^*(\vec{r})$; В — по функции $\varrho_{sp}(\vec{r})$

имели одинаковый объем, равный tV (где V — объем элементарной ячейки). Мы можем охарактеризовать точность совпадения областей Ω' и $\tilde{\Omega}'$, подсчитав объем области, состоящей из точек, принадлежащих Ω', но не принадлежащих области $\tilde{\Omega}'$ (или, что то же самое, объем области, состоящей из точек, попавших в $\tilde{\Omega}'$, но не принадлежащих Ω'). Обозначим через $M(t)$ объем этой области. Понятно, что области Ω' и $\tilde{\Omega}'$ тем точнее совпадают, чем меньше значение $M(t)$.

Введем нормировку величины $M(t)$ следующим образом. Допустим, что мы выбираем область Ω'_p объема tV совершенно случайно. Тогда, «в среднем», в область Ω' не попадет объем $(1-t)tV$, т.е. для случайно выбранной области Ω'_p мера ее несовпадения с областью Ω' будет $M_0(t)=(1-t)tV$. Мы введем теперь относительную ошибку выделения области Ω' при помощи функции $\tilde{\varrho}(\vec{r})$ как

$$\mu(t)=\frac{M(t)}{(1-t)tV}.$$

Величина $\mu(t)$ показывает, таким образом, во сколько раз точнее позволяет функция $\tilde{\varrho}(\vec{r})$ восстановить область Ω', нежели просто случайный выбор надлежащего числа точек.

На рис. 5 кривая В показывает точность определения областей Ω' для точного изображения $\varrho_e(\vec{r})$ синтеза разрешения 4 Å для субтилизина при помощи изображения $\varrho_{sp}(\vec{r})$, построенного по точным модулям и случайным фазам структурных факторов. Видно, что такой синтез не несет никакой информации об искомом объекте.

На рис. 5 кривая А показывает точность выделения областей Ω' по изображению $\tilde{\varrho}(\vec{r})$, рассчитанному по неполному набору рефлексов. Из рисунка видно, что синтез, построенный по неполному набору рефлексов, более менее удовлетворительно передает области высоких значений плотности (10—20% объема ячейки), но значительно хуже более широкие области.

Кривая Б на рис. 5 относится к попытке выделения областей Ω' при помощи изображения $\varrho^*(\vec{r})$, полученного в результате решения задачи 1, описанного в § 3. Здесь налицо существенный прогресс.

Владимир Юрьевич Лунин

ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИИ О РАСПРЕДЕЛЕНИИ ЗНАЧЕНИЙ ЭЛЕКТРОННОЙ ПЛОТНОСТИ В БЕЛКАХ
I. ВОССТАНОВЛЕНИЕ НЕДОСТАЮЩИХ СТРУКТУРНЫХ ФАКТОРОВ

Препринт

Отредактировано и подготовлено к печати в ОНТИ НЦБИ АН СССР

Редактор Р.Г.Цветницкая
Технический редактор Т.М.Печенкина
Корректоры Т.К.Крейцер, Л.М.Орлова

Подписано в печать 14.08.86 г. Т17853. Уч.-изд.л. 0,45. Усл.печ.л. 0,7. Формат 60x90/16. Тираж 200 экз. Бумага офсетная. Заказ 7409Р. Бесплатно. Изд. № 293.

Набрано на фотонаборном автомате ФА-1000. Отпечатано на ротапринте в Отделе научно-технической информации Научного центра биологических исследований АН СССР в Пущине