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

ПРЕПРИНТ

В.Ю.ЛУНИН, А.Г.УРЖУМЦЕВ

ПОВЫШЕНИЕ РАЗРЕШЕНИЯ КАРТ ЭЛЕКТРОННОЙ ПЛОТНОСТИ БЕЛКОВ ПУТЕМ УТОЧНЕНИЯ МОДЕЛЬНОЙ СТРУКТУРЫ. I. ОПИСАНИЕ МЕТОДА

ПУЩИНО • 1981

УДК 548.73

Приводится описание численного эксперимента по расширению набора определяемых фаз при рентгеноструктурном исследовании белка. Расширение набора было произведено путем построения и уточнения пробной модели структуры.

Работа будет направлена в журнал «Кристаллография».

Задачи в этой области столь сложны, что даже не очень четкая и половинчатая идея оправдывает затраченное на нее время, и можно то и дело возвращаться к одной и той же задаче, приближаясь понемногу к ее точному решению.

Р.Фейнман, Фейнмановские лекции по физике

Введение

При расшифровке кристаллической структуры белка методом рентгеноструктурного анализа на первом этапе работы ищется функция распределения электронной плотности в кристалле. В дальнейшем эта функция «интерпретируется» — исходя из картины расположения ее максимумов строится атомная модель структуры. (Описание методики рентгеноструктурного анализа белков см., например, в /1/). Сложность расчета функции распределения электронной плотности заключается в том, что эксперимент дает непосредственно лишь значения модулей коэффициентов Фурье этой функции, а задача определения фаз составляет так называемую «фазовую проблему рентгеноструктурного анализа». Далее мы будем пользоваться принятой в рентгеноструктурном анализе терминологией, именуя коэффициенты Фурье функции распределения электронной плотности структурными факторами.

До настоящего времени по существу единственным способом расчета фаз структурных факторов белка является метод изоморфного замещения (см. /2/), дополненный использованием данных по аномальному рассеянию. Недостатком метода является то, что он, зачастую, позволяет определить значения фаз лишь при разрешении порядка 3 Å, и надежность определения этих фаз не слишком велика. Это приводит к тому, что карты электронной плотности содержат недостаточное для построения полной молекулярной модели число деталей и сложны в интерпретации. В то же время экспериментальный набор модулей структурных факторов, как правило, бывает шире, чем полученный набор фаз, то есть не для всех экспериментально по-

лученных модулей мы в состоянии определить фазу. Применение «прямых» методов (см. /1/) уточнения и расширения набора фаз к белковым кристаллам сопряжено со значительными трудностями, так как для этих методов «стартовый» набор фаз должен быть существенно шире, чем обычно имеется в наличии.

В 1978 г. Агарвал и Айзекс в работе /3/ высказали идею, как, стартуя с набора фаз среднего разрешения (например, 3 Å) и набора модулей структурных факторов для более высокого разрешения (например, 2 Å), определить фазы для тех структурных факторов, для которых они не были определены. Было предложено действовать следующим образом:

а) По функции электронной плотности, рассчитанной по стартовому набору фаз, строится некоторая фиктивная модель структуры. Эта модель строится автоматически специальной программой и не претендует, вообще говоря, на соответствие истинной структуре.

б. Параметры атомов этой фиктивной модели уточняются программой кристаллографического уточнения в обратном пространстве (алгоритм уточнения см. в /4/).

в) Рассчитывается новая функция распределения электронной плотности (более высокого разрешения) с использованием экспериментальных значений модулей структурных факторов и рассчитанных по уточненной модели фаз.

В работе /3/ эта идея была проверена на данных по инсулину и были получены обнадеживающие результаты.

Содержанием данной работы является дальнейшее развитие, попытка практической реализации и проверка на тесте вышеизложенной идеи. В первом параграфе мы излагаем результаты применения этой методики к тестовому белку. Здесь же излагаются некоторые аспекты практической реализации метода. Второй параграф этой работы посвящен обсуждению вопроса: «Почему этот метод работает?». Здесь в рамках вероятностного подхода (см. /5/) делается попытка объяснить упомянутые в первом параграфе эффекты, возникшие по ходу численного эксперимента.

В работе по уточнению модели была использована программа Агарвала и Айзекса (см. /4/), любезно предоставленная ее авторами Ю.Н.Чиргадзе, которая была затем исправлена и адаптирована к ЭВМ серии ЕС в НИВЦ АН СССР А.Г.Уржумцевым. Все расчеты велись на ЭВМ ЕС-1040.

Авторы благодарны Г.Н.Борисюк за консультации по вопросам теории вероятностей и статистики, а также сотрудникам Института белка АН СССР Ю.Н.Чиргадзе и Ю.В.Сергееву за ценные обсуждения практических вопросов методики. Особо авторы хотят выразить признательность операторам и программистам группы сопровождения задач НИВЦ АН СССР и ее руководителю В.А.Королеву за помощь в организации и проведении трудоемких расчетов.

§ 1. Численный эксперимент по расширению набора фаз

1.1. В этом параграфе мы опишем ход и результаты численного эксперимента по расширению набора фаз.

Для теста был взят белок актинидин, кристаллизующийся в пространственной группе $P2_12_12_1$ с параметрами ячейки $a=78.2$, $b=81.8$, $c=33.05$ Å. В независимой части ячейки находится одна молекула, содержащая 218 аминокислотных остатков, состоящих из 1039 атомов углерода, 268 — азота, 337 — кислорода и 9 — серы. (Выбор именно этого объекта был чисто случаен и не несет никакой специальной нагрузки). По координатам атомов, взятым из банка белковых молекул, были рассчитаны модули и фазы структурных факторов до разрешения 2 Å (мы будем далее их обозначать $F(s)$ и $\varphi(s)$, где $s=(h, k, l)$ — целочисленный вектор обратного пространства). При этом всем атомам был приписан изотропный температурный фактор $B_t=20$ Å^2. Мы будем далее называть этот набор координат моделью M_0, а соответствующий набор фаз — набором Φ_0.

Далее в нашей работе набор всех модулей до разрешения 2 Å (15009 рефлексов) имитирует набор экспериментально измеренных модулей структурных факторов нативного белка, а набор фаз $\varphi(s)$ с $d=\lambda/2\sin\Theta\geqslant 3$ Å (4641 рефлекс) имитирует «стартовый» набор фаз, полученный по какой-либо методике. Фазы с $2\text{ Å}\leqslant d<3$ Å считаются неизвестными и служат далее только для контроля за ходом работы. В этом тесте мы игнорируем ошибку экспериментального определения величин $F(s)$ и неточности в определении фаз «стартового» набора. Наша задача состоит в том, чтобы, используя лишь фазы с $d\geqslant 3$ Å и все модули $F(s)$, определить значения всех фаз $\varphi(s)$ с $d\geqslant 2$ Å.

1.2. Первый этап работы заключался в построении синтеза (функции распределения) электронной плотности с разрешением 3 Å по «стартовому» набору фаз и затем автоматическом построении атомной модели белка (мы будем обозначать ее M_1) по этому синтезу. Синтез строился на сетке с числом делений по осям 112, 120, 48. Максимальное значение плотности на синтезе при этом оказалось равным 2.17 $e/$Å^3.

Для построения модели M_1 был выбран простейший алгоритм. Считалось, что все атомы в структуре аппроксимируются атомами азота с различными значениями изотропного температурного фактора B_t. Далее считалось, что на синтезе 3 Å изображение электронной плотности атома азота с температурным фактором B_t аппроксимируется гауссовой кривой:

$$\varrho_{B_t}(r) = C\left(\frac{4\pi}{B+B_t}\right)^{3/2} exp\left\{-\frac{4\pi^2R^2(r)}{B+B_t}\right\}, \qquad (1.1)$$

где $R(r)$ — расстояние в Å от начала координат до точки с координатами $r=(x, y, z)$, B и C — некоторые константы, выбор

которых описан в приложении 1. В нашем случае значения этих констант были определены как $C=12.1$ *e*, $B=71.4$ Å^2.

Процесс построения модели шел следующим образом:

а. На синтезе выбирался локальный максимум плотности, при этом его координаты считались координатами центра некоторого атома. Температурный фактор B_t для этого атома подбирался так, чтобы значение $\varrho_{B_t}(0)$, вычисленное согласно (1.1), совпадало со значением электронной плотности в этом максимуме (если выполнение этого условия было возможно лишь при $B_t<0$, для такого атома принудительно устанавливалось $B=0$).

б. Из синтеза вычиталась плотность определенного таким образом атома согласно формуле (1.1).

После этого происходил возврат к п. а до тех пор, пока максимальное значение в синтезе не становилось меньше некоторого заранее заданного уровня (в нашем случае $\varrho_{crit}=0.3$).

В результате работы программы по указанному алгоритму была создана модель M_1 из 2168 атомов. Отметим сразу, что, исходя из результатов численных экспериментов, мы считаем полезным в начале работы включать в модель больше атомов, чем их есть на самом деле в структуре. Это связано с двумя обстоятельствами: во-первых, при нашем способе построения модели часть атомов окажется «совсем не на месте», что будет обнаружено в процессе уточнения, и такие атомы мы в дальнейшем удалим из модели; во-вторых, поскольку мы ограничили максимум электронной плотности отдельного атома плотностью для азота при $B_t=0$, то возможно, что особо сильные пики на синтезе (например, для атомов серы) будут формироваться несколькими модельными атомами с близкими координатами центров. Подчеркнем, что при построении модели мы не накладываем никаких ограничений на межатомные расстояния в модели.

Для контроля по модели M_1 был рассчитан набор фаз (мы будем обозначать его Φ_1) с разрешением до 2 Å. Результаты сравнения фаз из Φ_1 с контрольными фазами из Φ_0 приведены в таблицах 1.2 и 1.3. В таблице 1.2 дана средняя ошибка $<|\varphi-\varphi^c|>$ (где φ^c принадлежит набору Φ_1, а φ — контрольное значение из Φ_0) для нецентросимметричных (нц/с) рефлексов в зависимости от $s^2=(2\sin\Theta/\lambda)^2$; в таблице 1.3 $<|\varphi-\varphi^c|>$ дано как функция «относительной силы рефлекса» $t=FF^c/\sigma_N^2$, где F и F^c — рассчитанные соответственно по моделям M_0 и M_1 модули структурных факторов, $\sigma_N^2=\sum_{j=1}^{N} f_j^2(s)$ — сумма по всем атомам структуры квадратов факторов атомного рассеяния.

Отметим сразу, что средняя ошибка определения фазы растет с ростом $|s|$ и убывает при возрастании силы рефлекса. Аналогичный эффект мы увидим далее и на последующих моделях структуры. Объяснение этому будет дано в § 2. Отметим еще для сравнения, что если генерировать модель датчиком случайных чисел, то средняя ошибка $<|\varphi-\varphi^c|>=90°$.

Таблица 1.1

План уточнения модельной структуры

Уточняемая модель	Зона уточнения	Количество рефлексов	Количество циклов по B	Уровень срезки по B	Количество циклов по X	Полученная модель	Количество атомов
M_1	3.0	4641	2	—	3	M_2	2168
M_2	3.0	4641	2	80	3	M_3	2142
M_3	2.7	6294	2	70	3	M_4	2043
M_4	2.5	7874	2	70	3	M_5	2033
M_5	2.2	11353	2	70	3	M_6	2026
M_6	2.0	15009	2	70	3	M_7	2022
M_8	3.0	4641	2	—	3	M_9	2182
M_9	2.5	7874	2	70	3	M_{10}	1878
M_{10}	2.2	11353	2	70	3	M_{11}	1875
M_{11}	2.0	15009	2	70	3	M_{12}	1868
M_{15}	3.0	4641	2	—	3	M_{16}	2933
M_{16}	2.5	7874	2	70	3	M_{17}	2872
M_{17}	2.2	11353	2	50	3	M_{18}	2560
M_{18}	2.0	15009	2	40	3	M_{19}	2295
M_{19}	2.0	15009	2	35	2	M_{20}	2080
M_{21}	2.0	15009	2	—	3	M_{22}	2205
M_{22}	2.0	15009	2	50	3	M_{23}	1782
M_{23}	2.0	15009	2	45	3	M_{24}	1681

Таблица 1.2

Зависимость средней фазовой ошибки (в градусах, для нецентросимметричных рефлексов) от $s^2 = (2 \sin \Theta/\lambda)^2$. Модели M_1—M_{12}

Модель	Границы интервалов по s^2										
	.000—.025	.025—.050	.050—.075	.075—.100	.100—.125	.125—.150	.150—.175	.175—.200	.200—.225	.225—.250	.000—.250
M_1	21	20	19	27	47	63	68	70	72	76	57
M_2	17	15	16	23	43	59	67	70	70	75	54
M_3	16	15	15	23	43	59	64	70	69	74	53
M_4	18	15	16	24	40	55	62	65	66	71	51
M_5	17	14	15	23	37	51	59	64	63	69	49
M_6	17	14	15	22	34	46	55	58	58	63	45
M_7	17	14	14	22	31	43	51	53	53	58	42
M_8	26	24	24	32	42	52	63	66	70	71	54
M_9	21	19	19	28	39	52	61	66	70	72	52
M_{10}	21	19	19	26	35	47	55	61	64	68	48
M_{11}	21	19	18	24	33	44	52	56	59	63	45
M_{12}	21	18	18	24	32	41	49	51	55	59	42
Количество рефлексов	320	681	957	1078	1278	1475	1502	1691	1783	1948	12713

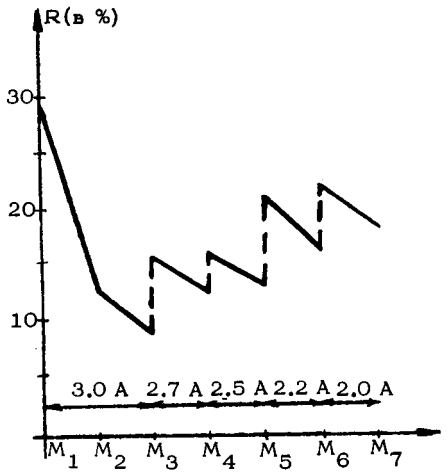
1.3 Следующим этапом работы явилось уточнение модели M_1. Для уточнения использовалась программа (Агарвала и Айзекса) уточнения в обратном пространстве (см. /4/), которая минимизирует критерий:

$$G = \sum_s (F(s) - F^c(s))^2$$

как функцию параметров $r = (x, y, z)$ и B_t для каждого атома.

Уточнение состояло из нескольких серий с постепенным увеличением от серии к серии числа рефлексов, включенных в уточнение. В каждой серии вначале проводилось несколько циклов уточнения B_t, а затем несколько циклов уточнения координат атомов. По ходу работы периодически после уточнения температурного фактора B_t из модели удалялись атомы со значениями $B_t > B_{crit}$ — такие атомы считались включенными в модель «совсем неправильно». Подробные сведения об этом этапе работы приведены в таблице 1.1. На рисунке 1.1 и в таблицах 1.2 и 1.3 приведены данные об изменении фактора $R = \sum_s |F - F^c| / \sum_s F$ и средней ошибки фаз, рассчитанных по модели, в ходе уточнения. Отметим то обстоятельство, что улучшение фаз φ^c при уточнении происходит, по существу, лишь для тех структурных факторов, амплитуды которых включены в уточнение.

1.4 Агарвал и Айзекс в своей работе /3/ рекомендуют после проведенного уточнения построить (используя для этого фазы, рассчитанные по последней модели) новый синтез (уже более



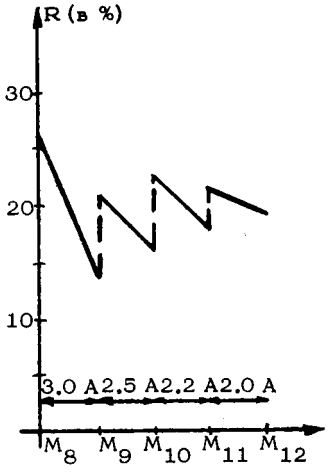
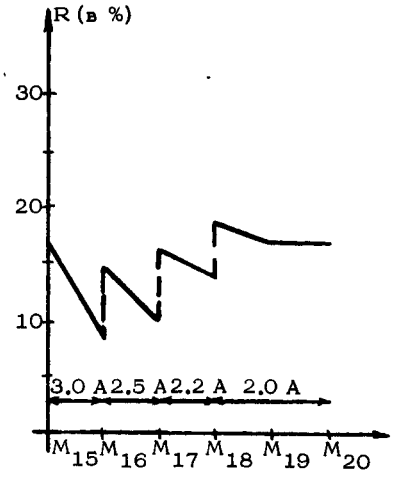


Рис. 1.1. Изменение R-фактора при уточнении модели M_1

Рис. 1.2. Изменение R-фактора при уточнении модели M_8



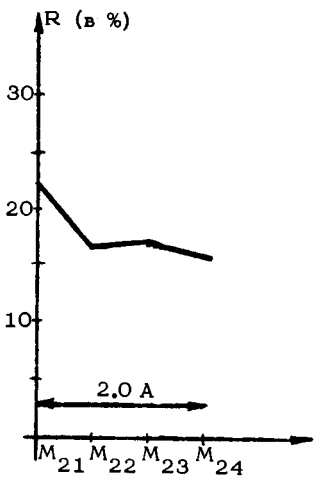


Рис. 1.3. Изменение R-фактора при уточнении модели M_{15}

Рис. 1.4. Изменение R-фактора при уточнении модели M_{21}

высокого разрешения, нежели «стартовый») и затем повторить работу, строя модель уже по новому синтезу. Наша попытка применить этот прием не привела к успеху. Нами был рассчитан синтез 2.5 Å, при расчете которого фазы до 3 Å брались из набора Φ_0 (точные), а для $2.5\ Å \leqslant d \leqslant 3\ Å$ из набора Φ_7 (фазы, рассчитанные по последней модели). Максимальное значение плотности на синтезе оказалось равным 2.54 $e/Å^3$. По этому синтезу была построена модель M_8 из 2182 атомов (при создании ее параметры в (1.1) имели значения: $B=42.3$, $C=11.1$, $\varrho_{crit}=0.32$). Далее модель M_8 уточнялась согласно плану, приведенному в таблице 1.1, что привело к результатам, отображенным на рисунке 1.2 и в таблицах 1.2 и 1.3. Как видно из этой таблицы, хотя для модели M_8 ошибка в фазах для высокоугловой зоны существенно меньше, чем для соответствующей модели M_1, но окончательный итог (ср. модели M_7 и M_{12}) одинаков. Возможно, что эта неудача вызвана «грубостью» алгоритма построения модели, не учитывающего тонких деталей синтеза 2.5 Å, и использование на этом этапе более изощренного алгоритма изменит ситуацию.

Нам удалось продвинуться дальше, применив другую методику, которая оказалась более подходящей для нашего случая. Вначале мы резко сократили число атомов в модели M_{12} (оставив всего 1140 атомов с $B_t<30\ Å^2$ — модель M_{13}). После этого было проведено 3 цикла уточнения координат модели M_{13} — модель M_{14}. Далее по модели M_{14} был построен разностный синтез с коэффициентами:

Таблица 1.3

Зависимость средней фазовой ошибки (в градусах, для нецентросимметричных рефлексов) от относительной силы рефлекса $t = FF/\sigma_N^2$. Модели $M_1 \div M_{12}$

Модель	Границы интервалов по t							
	0.00—0.50	0.50—0.75	0.75—1.00	1.00—1.50	1.50—2.00	2.00—5.00	5.00—10.00	10.00—∞
M_1	66	61	59	54	49	43	30	28
M_2	66	56	55	48	43	32	19	2
M_3	65	59	52	47	41	31	15	2
M_4	64	55	51	43	38	28	13	3
M_5	63	52	46	40	32	23	9	3
M_6	61	45	41	31	25	17	7	3
M_7	59	41	34	27	20	14	6	3
M_8	67	58	53	47	41	32	19	6
M_9	65	57	50	46	39	30	21	4
M_{10}	63	51	46	41	32	23	12	2
M_{11}	61	47	39	32	25	17	6	1
M_{12}	59	43	34	27	21	13	6	1
Среднее количество рефлексов	5000	1800	1400	1800	1000	1300	1000	10

Таблица 1.4

Зависимость средней фазовой ошибки (в градусах, для нецентросимметричных рефлексов) от $s^2 = (2 \sin \Theta/\lambda)^2$. Модели $M_{15} \div M_{24}$

Модель	Границы интервалов по s^2										
	.000—.025	.025—.050	.050—.075	.075—.100	.100—.125	.125—.150	.150—.175	.175—.200	.200—.225	.225—.250	.000—.250
M_{15}	16	16	16	22	33	47	55	58	59	63	45
M_{16}	14	13	14	21	33	45	53	57	59	64	44
M_{17}	14	13	14	21	30	42	50	53	53	60	41
M_{18}	15	14	15	20	29	38	45	49	49	55	38
M_{19}	16	15	16	21	27	35	42	45	44	51	36
M_{20}	17	14	17	22	27	35	40	43	43	50	35
M_{21}	21	17	16	23	29	37	43	46	46	51	37
M_{22}	17	14	12	18	24	33	36	40	42	47	33
M_{23}	18	14	14	19	24	32	35	38	40	46	32
M_{24}	17	14	13	19	24	29	34	37	39	44	31
Количество рефлексов	320	681	957	1078	1278	1475	1502	1691	1783	1948	12713

Таблица 1.5

Зависимость средней фазовой ошибки (в градусах, для нецентросимметричных рефлексов) от относительной силы рефлекса $t = FF^{\circ}/\sigma_N^2$. Модели $M_{15} \div M_{24}$

Модель	Границы интервалов по t							
	0.00—0.50	0.50—0.75	0.75—1.00	1.00—1.50	1.50—2.00	2.00—5.00	5.00—10.00	10.00—∞
M_{15}	62	48	41	36	27	19	9	3
M_{16}	61	46	41	32	26	17	6	1
M_{17}	58	42	35	27	22	15	6	1
M_{18}	56	37	32	24	19	13	6	3
M_{19}	54	35	27	21	17	11	5	3
M_{20}	54	34	27	21	17	11	7	2
M_{21}	54	35	26	22	16	11	7	5
M_{22}	49	30	23	19	14	10	6	3
M_{23}	50	30	25	20	15	10	7	1
M_{24}	49	28	23	18	14	9	6	1
Среднее количество рефлексов	5000	1800	1400	1800	1000	1300	1000	10

$$Fe^{i\tilde{\varphi}} = \begin{cases} Fe^{i\varphi} - F^*e^{i\varphi^*}, & d = 1/|s| \geqslant 3\ \text{Å}, \\ Fe^{i\varphi^c} - F^*e^{i\varphi^*}, & 2.5\ \text{Å} \leqslant d < 3\ \text{Å}, \end{cases}$$

где $\tilde{F}, \tilde{\varphi}$ — модуль и фаза коэффициентов Фурье для синтеза;
F, φ — «истинные модуль и фаза структурного фактора (рассчитанные по модели M_0);
F^*, φ^* — рассчитаны по модели M_{14};
F^c, φ^c — рассчитаны по модели M_{12}.

Значение электронной плотности в максимуме оказалось равным 2.32 $e/\text{Å}^3$. По разностному синтезу было определено программой построения модели дополнительно 1793 атома (параметры в (1.1): $B=42.8$, $C=11.05$, $\varrho_{crit}=0.40$), которые были добавлены к модели M_{14} — модель M_{15}. Далее было проведено уточнение модели M_{15} по плану, приведенному в таблице 1.1, и получены результаты, отображенные на рисунке 1.3 и в таблицах 1.5 и 1.4. Сравнение наборов фаз, рассчитанных по моделям M_7 и M_{20}, показывает эффективность разностных синтезов в указанной методике.

1.5 Ввиду достигнутого эффекта от применения разностного синтеза на предыдущем этапе этот прием был повторен за тем изменением, что на этот раз был построен разностный синтез разрешения 2 Å. План уточнения и достигнутые результаты приведены в таблицах 1.1, 1.4 и 1.5 и на рисунке 1.4.

1.6 Сделаем несколько заключительных замечаний. Основной итог работы по расширению набора фаз отражен в последних строках таблиц 1.4 и 1.5. Поскольку на практике надежность оп-

Таблица 1.6

Зависимость показателя достоверности m для фаз, рассчитанных по модели M_{24}, от разрешения

Границы по $s^2=\left(\frac{2\sin\Theta}{\lambda}\right)^2$	0.00-0.05	0.05-0.10	0.10-0.15	0.15-0.20	0.20-0.25	Все рефлексы
Границы по $d=\frac{1}{\lvert s\rvert}$	∞-4.5	4.5-3.2	3.2-2.6	2.6-2.2	2.2-2.0	∞-2.0
m	0.86	0.94	0.85	0.74	0.65	0.79
Число рефлексов	1001	2035	2753	3193	3731	12713

ределения фаз обычно оценивается не средней ошибкой, а показателем достоверности (см. /1/, /2/), то мы приведем еще в таблице 1.6 значения показателя достоверности для фаз, рассчитанных по последней модели M_{24} (способ определения показателя достоверности для фаз, рассчитанных по некоторой модели, изложен в следующем параграфе).

Отметим далее, что описываемая методика применима не только к задаче расширения набора фаз, но и к задаче уточнения имеющихся фаз. Формально, проведенный численный эксперимент показывает для структурных факторов с $d \geqslant 3$ Å лишь то, с какой точностью мы можем автоматически восстановить значения их фаз, стартуя с «идеального» синтеза 3 Å. Однако, в силу «грубости» алгоритма построения модели по синтезу, мы полагаем, что начальная модель M_1 не будет намного хуже, даже если мы будем стартовать с синтеза, имеющего некоторые искажения (в связи с неточным определением стартового набора фаз). Поэтому данные таблиц 1.2, 1.3, 1.4 и 1.5 по зонам разрешения до 3 Å можно рассматривать как оценку точности, до которой мы можем уточнить значения фаз из этих зон.

Наконец, заметим, что ответа на естественно возникающий вопрос: «С какого разрешения начинает работать эта методика и до какого разрешения она работает?» мы пока не знаем. Для ответа на этот вопрос требуется проведение дополнительных экспериментов.

§ 2. Влияние ошибок в определении координат атомов на точность определения фаз

2.1 Для того чтобы понять, почему изложенная в § 1 методика работает, прежде всего надо попытаться ответить на вопрос: «Какую информацию о фазах несет в себе некоторая модель структуры?». Ясно, что если модель близка к истинной структуре, то и фазы, рассчитанные по ней, не будут сильно отличаться от истинных фаз. С другой стороны, для случайной модели фазы, рассчитанные по ней, будут далеки от истинных. Цель этого параграфа — дать количественные соотношения, позволяющие связать погрешность в определении координат атомов модели с ошибкой в рассчитанных фазах.

В основе рассмотрения будет лежать вероятностный подход, развитый в работах ряда авторов (см., например, /5/). Однако, в отличие от работы /5/, где истинная и модельная структуры рассматриваются как случайные, мы будем полагать модельную структуру вполне конкретной и точно известной нам. В этом параграфе мы рассмотрим простейший случай, когда модель содержит столько же атомов, что и истинная структура, и отличается только положениями атомов (вопросы, связанные с различным числом атомов в модельной и истинной структурах и с неточным заданием факторов атомного рассеяния для модельной структуры, здесь рассматриваться не будут).

Пусть истинная структура имеет симметрию группы P_1 и состоит из N атомов с координатами r_j^* и факторами атомного рассеяния $f_j(s)$ $(j=1, \ldots, N)$. Пусть модельная структура также состоит из N атомов с факторами атомного рассеяния $f_j(s)$, но с координатами атомов r_j (здесь по-прежнему r_j и r_j^* — трехмерные векторы координат: $r_j=(x_j, y_j, z_j)$; $s=(h, k, l)$ — целочисленный вектор в обратном пространстве; $|s|=2\sin\Theta/\lambda$).

Будем обозначать структурные факторы истинной и модельной структур $F(s)$ и $F^c(s)$ соответственно:

$$\begin{aligned} \vec{F}(s) &= F(s)\,e^{i\varphi(s)} = \sum_{j=1}^{N} f_j(s)\,e^{2\pi i(s, r_j^*)}, \\ \vec{F}^c(s) &= F^c{}_{(s)}\,e^{i\varphi^c(s)} = \sum_{j=1} f_j(s)\,e^{2\pi i(s, r_j)}. \end{aligned} \tag{2.1}$$

Поскольку часто будет рассматриваться отдельный структурный фактор, то для краткости мы будем опускать в формулах аргумент s там, где это не вызывает недоразумений.

Мы будем считать, что

$$r_j^* = r_j + \Delta_j, \quad j = 1, \ldots, N, \tag{2.2}$$

где Δ_j (ошибки в координатах атомов модельной структуры) — случайные величины. Основное предположение о характере ошибок Δ_j, которое мы будем использовать, формулируется следующим образом.

УСЛОВИЕ 1 Случайные величины Δ_j независимы и имеют одинаковое распределение вероятностей, симметричное относительного нуля.

Таким образом, мы считаем модельную структуру точно известной, а модули и фазы «истинных» структурных факторов F и φ являются случайными величинами, выражающимися через величины Δ_j при помощи соотношений (2.1) и (2.2). Заметим, что значение случайной величины F мы можем непосредственно получить из рентгеновского эксперимента. Поэтому информация о фазах, обусловленная модельной структурой, может быть представлена в виде плотности $P(\varphi|F)$ условного распределения вероятностей случайной величины φ при условии, что случайная величина F приняла данное (полученное из эксперимента) значение.

2.2 Оказывается (см. приложение 2), что для «большинства» рефлексов плотность распределения вероятностей величины $\alpha=\varphi-\varphi^c$ (при условии, что величина F приняла данное значение) можно приближенно представить в виде:

$$P(\alpha) = \frac{1}{2\pi I_0(\lambda t)}\,e^{\lambda t\cos\alpha} \tag{2.3}$$

где I_0 — модифицированная функция Бесселя нулевого порядка; $\alpha = \varphi - \varphi^c$; $t = FF^c/\sigma_N^2$; $\lambda = 2\mu/(1-\mu^2)$; а важная для нас в дальнейшем величина $\mu = \mu(s)$, характеризующая ошибку в определении координат атомов, определяется как

$$\mu(s) = \langle \cos 2\pi (s, \Delta_j) \rangle$$

(скобки $\langle \ldots \rangle$ здесь и далее означают среднее значение случайной величины).

Отметим два предельных случая. Если ошибки в построении модели столь велики, что Δ_j можно считать равномерно распределенными, то $\mu(s) = 0$ и $P(\alpha) = const$, что означает полную неопределенность фазы. С противоположной стороны, если плотность распределения вероятностей величины Δ_j стремится к δ-функции, то $\mu(s)$ стремится к единице и функция $P(\alpha)$ стремится к $\delta(\alpha)$, то есть фаза определена однозначно.

В промежуточных случаях мы можем из функции $P(\alpha)$ определить «показатель достоверности» m определения фазы φ^c (по Блоу и Крику, /2/) из соотношения:

$$m = |\langle e^{i\varphi} \rangle| = \left| \int_0^{2\pi} e^{i\varphi} P(\varphi - \varphi^c)\, d\varphi \right| = I_1(\lambda t)/I_0(\lambda t) \quad (2.4)$$

(где I_0 и I_1 — модифицированные функции Бесселя нулевого и первого порядков). Кроме того, в качестве характеристики, определяющей надежность фазы φ^c, можно использовать, например, среднюю ошибку определения фазы:

$$\langle |\Delta\varphi| \rangle = \langle |\varphi - \varphi^c| \rangle = \quad (2.5)$$

$$= \int_{-\pi+\varphi^c}^{\pi+\varphi^c} |\varphi - \varphi^c| P(\varphi - \varphi^c)\, d\varphi = \frac{1}{\pi I_0(\lambda t)} \int_0^{\pi} \varphi e^{\lambda t \cos\varphi} d\varphi.$$

Соотношения (2.4) и (2.5) однозначно определяют m как функцию от $\langle |\Delta\varphi| \rangle$. График этой функции приведен на рисунке 2.1.

Из определения величины $\mu(s)$ видно, что если плотность распределения вероятностей величины Δ_j «стягивается» к нулю, то есть ошибки в определении координат модели уменьшаются, то значение μ возрастает, что приводит к обострению плотности $P(\alpha)$, то есть к уменьшению ошибки определения фазы.

Заметим также, что при одних и тех же ошибках Δ_j значение $\mu(s)$, вообще говоря, убывает с ростом $|s|$. Это означает, что при фиксированной модели ошибка в определении фазы будет возрастать с ростом $|s|$. Это обстоятельство уже отмечалось при описании результатов эксперимента, изложенного в параграфе 1.

Наконец, из (2.3) видно, что ошибка в определении фазы будет убывать с ростом величины FF^c, то есть для «сильных» рефлексов ошибка меньше. Этот факт мы также отмечали при описании результатов эксперимента в § 1.

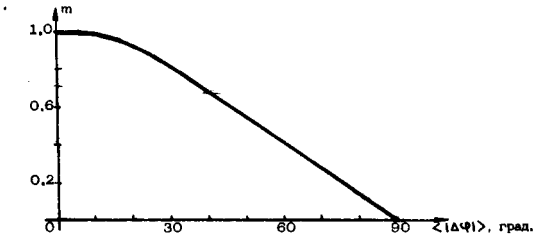


Рис. 2.1. Связь показателя достоверности m со средней ошибкой определения фазы для нецентросимметричных рефлексов

2.2 В этом пункте мы введем некоторые количественные характеристики, вытекающие из представления (2.3). Для определенности будем считать в этом пункте, что случайные ошибки Δ_j распределены нормально, то есть плотность распределения вероятностей величины Δ_j есть:

$$P(\Delta) = \frac{1}{(2\pi\nu^2)^{3/2}} exp\left\{ - \frac{|\Delta|^2}{2\nu^2} \right\}.$$

Нам будет удобно характеризовать это распределение средней величиной ошибки в определении координат

$$\omega = <|\Delta|> = \iiint_{R^3} |\Delta| P(\Delta)\, d\Delta_x d\Delta_y d\Delta_z = \frac{4}{\sqrt{2\pi}}\nu.$$

Как показывает прямое вычисление, в этом случае

$$\mu(s) = \mu_\omega(s) = exp\left\{ - \frac{\pi^3}{4} (\omega|s|)^2 \right\},$$

и мы можем определить из распределения (2.3) показатель достоверности m и среднюю ошибку в определении фазы $<|\Delta\varphi|>$ как функции ω, $|s|$ и $t = FF^c|\sigma_N^2$.

Относительно диапазона значений t заметим, что формально t может меняться от 0 до $(\Sigma f_j^2(s))^2/\sigma_N^2 \sim N$. Однако известно (см. /5/), что если рассматривать координаты атомов модели как независимые случайные величины, равномерно распределенные в элементарной ячейке, то $<F^{c2}/\sigma_N^2> = 1$. Поэтому «в среднем» значение величины $t = FF^c/\sigma_N^2$ будет порядка единицы. На

рисунках 2.2 и 2.3 приведены графики зависимости $<|\Delta\varphi|>$ и m от $\omega = <|\Delta|>$ для нескольких характерных значений t и $|s|$.

2.4 Возвращаясь к результатам численного эксперимента, изложенным в первом параграфе, подчеркнем два момента, опре-

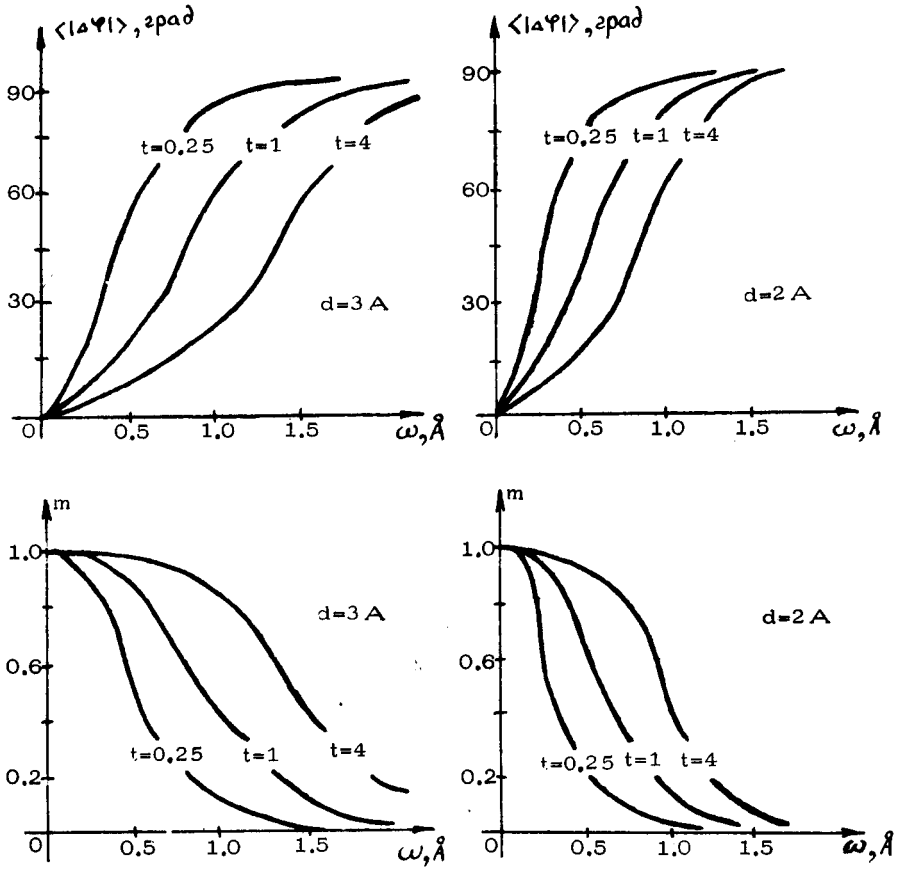


Рис. 2.2. Зависимость фазовой ошибки от погрешности $\omega = <|\Delta|>$ определения координат $(t = FF^c/\sigma_N^2)$

Рис. 2.3. Зависимость показателя достоверности m от погрешности $\omega = <|\Delta|>$ определения координат $(t = FF^c/\sigma_N^2)$

деливших успех работы. Во-первых, как следует из графиков, приведенных на рисунках 2.2 и 2.3, даже очень грубая модель дает некоторую оценку искомых фаз. Так, ошибки при расчете фаз по модели M_1, построенной программой, описанной в п.1.2, соответствуют средней ошибке в определении координат модели

M_1 порядка 0.9 Å (при дополнительной гипотезе о нормальности распределения ошибки в координатах этой модели). Ясно, что такую точность модели может дать даже столь грубый алгоритм, как описанный нами. Во-вторых, как было указано в п.2.2, уменьшение средней ошибки в определении координат атомов модели, достигнутое за счет уточнения модели, хотя не привело к точной структуре (при предположении о нормальности распределения ошибки в координатах ее средняя величина для модели M_{24} оценивается величиной порядка 0.5 Å), но позволило снизить среднюю ошибку фазы.

Дальнейшее развитие методики может идти как по пути совершенствования алгоритма построения модели по синтезу среднего разрешения (что позволит более точно определять фазы уже на первом этапе работы), так и по пути дальнейшего развития средств уточнения атомной модели.

ПРИЛОЖЕНИЕ 1

Расчет параметров изображения атома на синтезе среднего разрешения

Пусть мы рассчитываем на некоторой сетке синтез по множеству рефлексов S:

$$\varrho(r) = \frac{1}{V} \sum_{s \in S} F(s)\, e^{i\varphi(s)}\, e^{-2\pi i (s,r)}.$$

Если некоторый атом имеет фактор рассеяния $g(s)$ и температурный фактор B_t, то, будучи помещенным в начало координат, он дает вклад в синтез:

$$q(r) = \frac{1}{V} \sum_{s \in S} g(s)\, e^{-B_t \frac{s^2}{4}}\, e^{-2\pi i (s,r)} =$$

$$= \int_0^1 \int_0^1 \int_0^1 q_0(r-t)\, \mu_{B_t}(t)\, dt_x dt_y dt_z, \qquad (\text{П } 1)$$

где

$$q_0(r) = \frac{1}{V} \sum_{s \in S} g(s)\, e^{-2\pi i (s,r)},$$

$$\mu_{B_t}(r) = \frac{1}{V} \sum_{s \in Z^3} e^{-B_t \frac{s^2}{4}}\, e^{-2\pi i (s,r)}.$$

Пусть $R(r)$ — расстояние в Å точки с координатами $r = (x,y,z)$

от начала координат. Функцию $\mu_{B_t}(r)$ можно приближенно считать равной

$$\mu_{B_t}(r) \cong \left(\frac{4\pi}{B_t}\right)^{3/2} e^{-\frac{4\pi^2 R^2(r)}{B_t}}$$

(это связано с тем, что, в силу быстрого убывания экспоненты, интеграл по всему пространству в равенстве

$$\iiint_{R^3} \mu_{B_t}(r)\, e^{2\pi i (s,r)}\, dV_r = e^{-B_t \frac{s^2}{4}}$$

можно приближенно заменить интегралом по элементарной ячейке).

Пусть функция $q_0(r)$ аппроксимирована функцией

$$\bar{q}_0(r) = C\left(\frac{4\pi}{B}\right)^{3/2} e^{-\frac{4\pi^2 R^2(r)}{B}}, \qquad (\text{П } 2)$$

тогда, заменяя интеграл в (П 1) на интеграл по всему пространству, получаем использованную ранее аппроксимацию (1.1).

После логарифмирования функций $q_0(r)$ и $\bar{q}_0(r)$ и введения параметров $\beta = 4\pi^2/B$, $\alpha = ln[VC(4\pi/B)^{3/2}]$ и переменной $u = R^2(r)$ задача определения параметров аппроксимации (П 2) сводится к линейной задаче нахождения параметров аппроксимации

$$ln \sum_{s \in S} g(s)\, e^{-2\pi i (s,r)} \simeq \alpha + \beta u.$$

Параметры α и β определялись из условия минимума величины

$$\sum_j \left[ln \sum_{s \in S} g(s)\, e^{-2\pi i (s,r)} - (\alpha + \beta u)\right]^2,$$

где внешнее суммирование проводилось по всем точкам сетки, лежащим внутри сферы радиуса 2 Å с центром в начале координат, а фактор рассеяния $g(s)$ для атома азота был взят так же, как и в работе /4/.

ПРИЛОЖЕНИЕ 2

Вывод формулы для плотности $P(\varphi|F)$

Мы можем вывести формулу для плотности распределения вероятностей $P(\varphi|F)$, повторив, по существу, вывод, приведенный в работе /5/ (глава 5). Отличие будет в том, что мы будем считать F^c и φ^c конкретными параметрами, а не случайными величинами, как в работе /5/. Кроме того, чтобы лучше уяснить законность аппроксимации (2.3), мы проведем вычисления до конца, не производя преждевременно отбрасы-

вания каких-либо величин. Повторяя выкладки работы /5/, основанные на применении центральной предельной теоремы теории вероятностей, мы придем к тому, что совместное распределение величин F и φ дается формулой:

$$P(F,\varphi) = C\,exp\left\{\frac{[F^2-2\mu FF^c\cos(\varphi-\varphi^c)+\mu^2F^{c\,2}]}{\sigma_N^2(1-\mu^2)(1-\varepsilon^2)} + \right.$$

$$\left. + \frac{\varepsilon[F^2\cos(2\varphi+\tilde{\varphi}) - 2\mu FF^c\cos(\varphi+\tilde{\varphi}+\varphi^c) + \mu^2F^{c\,2}\cos(2\varphi^c+\tilde{\varphi})]}{\sigma_N^2(1-\mu^2)(1-\varepsilon^2)}\right\}$$

где введены обозначения

$$\sigma_N^2 = \sum_{j=1}^{N} f_j^2(s), \quad \varepsilon = \frac{\mu^2(s)-\mu(2s)}{1-\mu^2(s)} \cdot \frac{F(s)}{\sigma_N^2(s)} \qquad (\Pi\ 3)$$

а $Fe^{i\tilde{\varphi}} = \sum_{j=1}^{N} f_j^2(s)\,e^{2\pi i(s,2r_j)}$ — структурный фактор модифицированной структуры (с атомами, имеющими факторы атомного рассеяния $f_j^2(s)$ и координаты атомов $2r_j$).

Отсюда нетрудно видеть, что искомое условное распределение представляется в унифицированном виде (см. /6/):

$$P(\varphi|F) = \varkappa\,exp\{A\cos\varphi + B\sin\varphi + C\cos2\varphi + D\sin2\varphi\},$$

где $\varkappa$ — нормирующий множитель, а коэффициенты A, B, C, D в данном случае есть:

$$A = \frac{2\mu FF^c}{\sigma_N^2(1-\mu^2)(1-\varepsilon^2)}[\cos\varphi^c + \varepsilon\cos(\varphi^c+\tilde{\varphi})],$$

$$B = \frac{2\mu FF^c}{\sigma_N^2(1-\mu^2)(1-\varepsilon^2)}[\sin\varphi^c + \varepsilon\sin(\varphi^c+\tilde{\varphi})], \qquad (\Pi\ 4)$$

$$C = -\varepsilon F^2\cos\tilde{\varphi}/[\sigma_N^2(1-\mu^2)(1-\varepsilon^2)],$$

$$D = \varepsilon F^2\sin\tilde{\varphi}/[\sigma_N^2(1-\mu^2)(1-\varepsilon^2)].$$

Заметим теперь, что, как правило, величина ε, определенная в (П 3) и входящая в формулы (П 4), будет мала, и для плотности $P(\varphi|F)$ справедливо упрощенное представление

$$P(F,\varphi) \cong \frac{F}{\pi\sigma_N^2(1-\mu^2)}\,exp\left\{-\frac{F^2-2\mu FF^c\cos(\varphi-\varphi^c)+\mu^2F^{c\,2}}{\sigma_N^2(1-\mu^2)}\right\}, \qquad (\Pi\ 5)$$

использованное в работах /5/, /7/. В общем случае для величины F верхняя грань есть σ_N^2. Однако, для независимых слу-

чайных координат r_j, равномерно распределенных в единичном кубе, среднее значение (см. /5/) $<F> = \frac{1}{2}\sqrt{\pi \sum f_j^4(s)}$ и дисперсия равна $\left(1 - \frac{\pi}{4}\right)\sum_j f_j^4(s)$. Поэтому, как правило, величина F/σ_N^2 будет порядка $1/\sqrt{N}$. Кроме того, считая ошибки Δ_j распределенными по нормальному закону, мы получим (см. п.2.3), что

$$(\mu^2(s) - \mu(2s))/(1-\mu^2(s)) = exp\left\{-\frac{\pi^3}{2}(\omega|s|)^2\right\},$$

(где $\omega = <|\Delta_j|>$ — средняя ошибка в определении координат), то есть при больших $\omega|s|$ величина ε мала еще и за счет этого множителя. Наконец, отметим, что, поскольку при уменьшении ошибок Δ_j величины F^c стремятся к F, а φ^c к φ, то при малых ошибках будет еще мало и все выражение, стоящее в формуле для $P(F,\varphi)$ во вторых квадратных скобках в показателе экспоненты.

Из формулы (П 5) вытекает упрощенное представление условной плотности $P(\varphi|F)$:

$$P(\varphi|F) \cong \varkappa exp\left\{\frac{2\mu F F^c cos(\varphi - \varphi^c)}{\sigma_N^2(1-\mu^2)}\right\},$$

которое мы и использовали в § 2.

ЛИТЕРАТУРА

1. *Бландел Т., Джонсон Л.* Кристаллография белка. М., Мир, 1979.
2. *Blow D.M., Crick F.H.C.* The Treatment of Errors in the Isomorphous Replacement Method. — Acta Cryst., 1959, v. 12, p. 794.
3. *Agarwal R.C., Isaacs N.W.* Method for obtaining a high resolution protein map starting from a low resolution map. — Proc. Natl Acad. Sci. USA, 1977, v. 74, p. 2835.
4. *Agarwal R.C.* A New Least-Squares Refinement Techniques Based on the Fast Fourier Transform Algorithm, — Acta Cryst., 1978, v. A34, p. 791.
5. *Сринивасан Р., Партасарати С.* Применение статистических методов в рентгеновской кристаллографии. М., Мир, 1979.
6. *Hendrickson W.A., Lattman E.E.* Representation of Phase Probability Distributions for Simplified Combination of Independ Phase Information. — Acta Cryst., 1970, v. B26, p. 136.
7. *Luzzati V.* Traitement statistique des erreurs dans la determination des structures cristallines. — Acta Cryst., 1952, v. 5, p. 802.

СОДЕРЖАНИЕ

Владимир Юрьевич Лунин
Александр Георгиевич Уржумцев

ПОВЫШЕНИЕ РАЗРЕШЕНИЯ КАРТ ЭЛЕКТРОННОЙ ПЛОТНОСТИ БЕЛКОВ ПУТЕМ УТОЧНЕНИЯ МОДЕЛЬНОЙ СТРУКТУРЫ
Препринт

Подписано в печать 14.05.81 г. Т08934.
Уч.-изд. л. 1,3. Формат 60×90/16. Тираж 100 экз.
Бумага офсетная. Заказ 1264Р. Изд. № 150

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