

АЛГОРИТМЫ И МЕТОДЫ ИССЛЕДОВАНИЯ ТРЕХМЕРНЫХ АТОМИСТИЧЕСКИХ МОДЕЛЕЙ МОЛЕКУЛ БЕЛКОВ НА ОСНОВЕ АНАЛИЗА КАРТИНЫ РАССЕЯНИЯ МОЩНОГО РЕНТГЕНОВСКОГО ЛАЗЕРНОГО ИЗЛУЧЕНИЯ

Шайтан К.В.^{1,2}, Шайтан А.К.¹, Багров Д.В.¹, Блинов В.Н.^{1,3*},
Новоселецкий В.Н.¹, Соколова О.С.¹, Шуров Д.Л.^{1,2},
Турченков Д.А.¹, Кирпичников М.П.¹

¹Московский государственный университет имени М.В.Ломоносова, Москва

²Институт химической физики имени Н.Н.Семенова РАН, Москва

³Московский институт электроники и математики при НИУ ВШЭ, Москва

shaytan49@yandex.ru

Поступила 15.08.2013

Обсуждаются методы и алгоритмы обработки информации в области представления и анализа экспериментальных данных о пространственной структуре и распределении электронной плотности в биомолекулах и их комплексах, алгоритмам анализа и интеграции данных рентгеновского рассеяния, электронной микроскопии и методов молекулярного моделирования для исследования структуры и динамики биомолекул и их комплексов.

УДК 577

Введение

Определение пространственного строения биомолекул и их комплексов является необходимым шагом в решении многих практических и теоретических вопросов биологии, фармакологии, биотехнологии и медицины. В настоящее время в мире расшифровано порядка 80000 пространственных структур белков, в их числе, ферменты, мембранные белки, ионные каналы, аквапорины, рибосомальные субъединицы и др.

* Работа поддержана Программой фундаментальных научных исследований НИУ ВШЭ.

Признанными методами для изучения структуры и конформационных изменений в белковых молекулах являются рентгеноструктурный анализ (РСА), ядерный-магнитный резонанс (ЯМР), просвечивающая электронная микроскопия (ПЭМ) и криомикроскопия (таблица 1). Каждый из этих методов имеет свои преимущества и ограничения.

Несмотря на существенные успехи, достигнутые в этом направлении с помощью современных методов РСА, ЯМР, электронной микроскопии (ЭМ) достаточно большой класс биологических объектов не удастся исследовать с необходимой атомной точностью. К таким объектам, прежде всего, относятся большие биомакромолекулы и их комплексы, не поддающиеся кристаллизации или формирующие кристаллы малых размеров и/или плохого качества. Решительный скачок в экспериментальных возможностях ожидается в связи с разработкой и введением в эксплуатацию источников рентгеновского излучения высокой мощности четвертого поколения – лазеров на свободных электронах, которые обладают уникальным набором характеристик, включая тонкую временную структур импульсов и пиковую светимость, которая более, чем на 10 порядков больше, чем у источников предыдущего поколения – синхротронов. Такие характеристики позволяют проводить ряд экспериментов и получать данные нового типа, в частности, детектировать картину рассеяния на единичных макромолекулах и нанокристаллах.

Ожидается, что на первом этапе в большинстве случаев качество получаемых экспериментальных данных не будет достаточным для восстановления структуры молекул с атомной точностью стандартными методами, принятыми в современном РСА. Таким образом, формируется запрос на создание новых эффективных алгоритмов анализа экспериментальных данных, а также разработку алгоритмов интеграции различных экспериментальных данных с применением современных представлений о строении вещества, методов молекулярного моделирования и новых математических идей.

Среди наиболее актуальных задач можно выделить задачу реконструкции пространственной структуры биомакромолекул и их комплексов на основе карт электронной плотности низкого разрешения, которые, как ожидается, будут получаться в результате рассеяния лазерного рентгеновского излучения на отдельных частицах или нанокристаллах. Основным направлением для решения этой задачи является развитие алгоритмов и методик, позволяющих комбинировать экспериментальные данные низкого разрешения о строении биомакромолекулярного комплекса с информацией о структуре отдельных субъединиц этого комплекса, которую зачастую можно получить путем анализа баз данных строения белковых структур, в том числе, путем моделирования строения на основе гомологии аминокислотной последовательности с уже известными структурами.

1. Перспективы использования рентгеновских лазеров на свободных электронах для изучения структур биообъектов

Первые разработки лазеров на свободных электронах (ЛСЭ) были выполнены в Стэнфорде в 1976 г. Джоном Мэйди. До недавнего времени, данные приборы работали в инфракрасном либо ультрафиолетовом диапазонах. Однако, применение синхротронных радиационных источников позволило построить приборы, работающие в рентгеновском диапазоне. Данные приборы начали активно применяться для исследования структур биологических объектов и изучения промежуточных состояний реагентов в химических реакциях.

Общий принцип работы ЛСЭ показан на рис. 1. Поток заряженных частиц из электронной пушки разгоняется в ускорителе до высоких энергий и проходит через

Таблица 1. Сравнение современных методов определения структуры биомолекулярных объектов

Метод		Образцы	Разрешение	Достоинства	Ограничения
Рентгеноструктурный анализ		Кристаллы биомолекул	атомное	Высокое разрешение. Возможность применения роботов для получения кристаллов	Только для объектов, формирующих кристаллы. Требуется высокая концентрация белка (>3 мг/мл). Нативная конформация молекулы может быть нарушена при формировании кристалла.
Ядерно-магнитный резонанс		Относительно небольшие молекулы белков в растворе	атомное	Высокое разрешение. Конформация молекулы в растворе.	Молекулярная масса <100 кДа; Большая концентрация белка. Изотопное мечение.
ПЭМ белковых молекул	Двумерные кристаллы	Молекулы	Атомное/молекулярное	Мембранные белки. Применение роботов для кристаллизации.	Требуется наличие кристаллов. Поворот сетки на заданные углы. Медленно и очень сложно.
	Симметричные макромолекулярные структуры	Вирусы, спиральные молекулы (актин, тубулин и др.)	Вторичная структура	Нативная структура в растворе; способность дифференцировать разные конформации	Ограничения в разрешении (7-25Å).
	Ассиметричные молекулы	Большие комплексы, рибосома	Молекулярное		
ЭМ томография		Целые клетки, большие комплексы, макромолекулы	Общий вид структур	Взаимодействие субъединиц в комплексах, строение отдельных макромолекул	Ограничения в разрешении; радиационное разрушение образца.

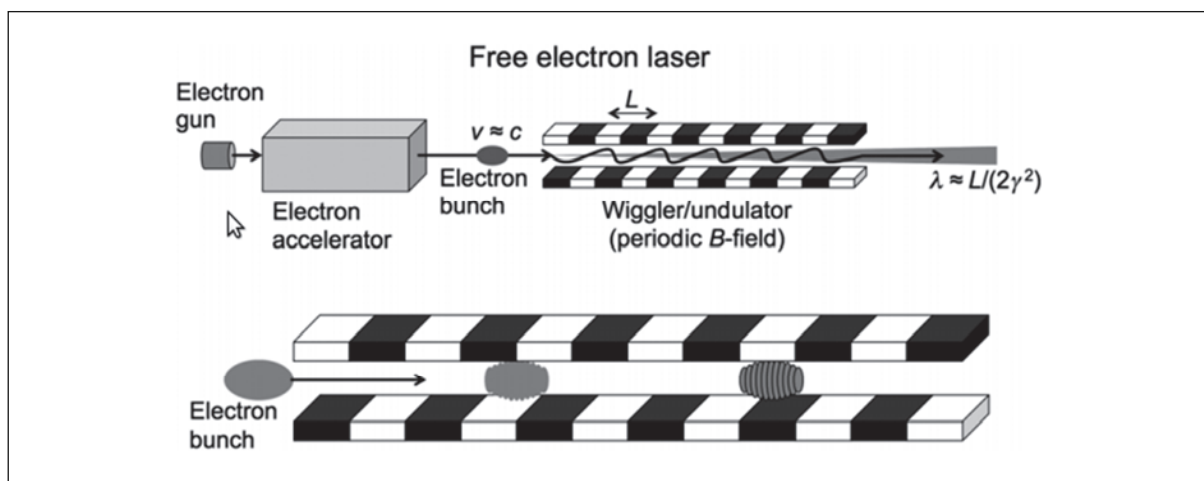


Рис. 1. Схема лазера на свободных электронах

ондуляторы – магниты, создающие синусоидальное магнитное поле. При этом электрон теряет энергию, испуская её в виде фотонов, которые собираются системой зеркал, образуя мягкое рентгеновское излучение. Данный эффект был обнаружен П.Л. Капицей и П.А.М. Дираком в 1933 г. Далее, в 1947 г. В.Л. Гинзбургом был предложен принцип ондулятора. Изначально прибор, работающий на этом принципе, предполагалось использовать для исследования космических лучей.

Особенность ЛСЭ состоит в том, что меняя энергию пучка и параметры ондулятора (энергию поля и расстояние между магнитами) можно в широких пределах менять частоту излучения. Например, в первых исследованиях Мэйди использовался 24 МэВ пучок и 5-ти метровый вигглер, эффективность прибора составляла всего 0.01%.

Полученная длина волны составляла 3 мк. Для получения рентгеновского излучения необходимо ускорить электроны на входе до скоростей, близких к скорости света.

Важное применение ЛСЭ – кристаллография. Метод основывается на дифракции рентгеновских лучей на исследуемом объекте. Высокая концентрация энергии и малая длительность импульса (несколько фемтосекунд), а также высокая пространственная когерентность излучения позволяют получить устойчивую дифракционную картину даже для одиночной молекулы.

На данный момент уже функционирует несколько установок – LCLS в Оксфорде, SwissFEL в Швейцарии, SACLA в Японии, FLASH Германии. Новый лазер на свободных электронах – XFEL, строящийся в Германии (Гамбург) является на сегодняшний день наиболее совершенной установкой такого типа. XFEL предполагается использовать для исследований динамики изменений состояния молекул при различных химических реакциях, а также структуры биообъектов. Особенностью данного проекта является наличие вокруг XFEL развитой инфраструктуры для биологических исследований, в том числе, и с участием ряда российских лабораторий.

Одной из важных проблем в современной биологии является изучение мембранных белков и мультимерных белковых комплексов, что помогает в понимании процессов жизнеобеспечения на надмолекулярном уровне. Эти объекты, подвергнутые очистке современными методами, недостаточно стабильны и редко доступны в виде кристаллов хорошего качества.

Лазеры на свободных электронах позволяют получать дифракционные картины не только для нанокристаллов, но и для одиночных молекул, целых клеток, а также анализировать изменение состояния вещества в ходе химических реакций. Как известно, определенные трудности применения рентгеноструктурных методов состоят также в повреждении кристалла в процессе исследования и в получении кристаллов необходимого размера. Использование фемтосекундных импульсов рентгеновских лазеров позволяет получить дифракционную картину достаточно высокого разрешения до разрушения исследуемого кристалла [1]. Лазеры на свободных электронах позволяют, в принципе, исследовать молекулярную структуру как больших, так и малых биологических систем.

Во многих случаях, крупные белковые комплексы образуют небольшие, слабосвязные кристаллы, не очень удобные для анализа даже на синхротронах 3-го поколения. Например, волокно мышечных саркомеров гибко и эластично, что затрудняет работу с их кристаллами. Результаты исследований [1] на LCLS показали, что исследовать такие кристаллы возможно, используя фемтосекундные рентгеновские лазеры. Особое место занимают мембранные белки – группа гидрофобных белков не растворимых в воде. Они выступают в различных качествах: транспортные, структурные белки, рецепторы, ферменты, медиаторы, молекулы межклеточной связи, активаторы. Примерно треть всех белков, кодируемых эукариотическими клетками – мембранные белки. Ионные каналы в организме человека, а также вторичные активные транспортные белки, вызывают интерес у исследователей-медиков и биологов, т.к. на их основе возможна разработка новых лекарств. Однако, в базе данных Protein Data Bank [2] из 73000 структур, доступных для анализа, мембранных белков достаточно мало. Значительная часть их (порядка 300) доступна в базе Membrane Protein Data Bank [3]. Структуры получены методами рентгеновской и электронной дифракции, ЯМР-спектроскопии и крио-электронной микроскопии. Значительная часть структур имеет приложение в медицине: Na/K-ионный канал, Fo-F1 АТФаза, G белок-связанный рецептор, β_2 адронергический рецептор, аквапорин, лейкотриен C₄-синтаза, дофаминовый рецептор и др. Половина исследуемых сейчас лекарств воздействует именно на мембранные белки [4].

Ионные каналы и другие белки, интегрированные в мембрану – интересные объекты для исследования методами одномолекулярной и нанокристаллической дифракции, т.к. они образуют крупные макромолекулярные комплексы, достаточно редко встречаются и трудно кристаллизуются. Получаемые кристаллы также не дают достаточно четкую дифракционную картину. Тем не менее, на XFEL для фотосистемы 1 были получены структуры MDa комплекса из 36 белков. Во многих случаях данные, полученные с XFEL можно использовать для повышения точности и комбинировать с данными, например, криоэлектронной микроскопии.

2. Особенности определения структуры с использованием лазеров на свободных электронах для нанокристаллов и единичных макромолекул

Ожидается, что характеристики рентгеновского излучения, которое будет получено с помощью лазеров на свободных электронах, позволят проводить качественно новые типы экспериментов [5]. В частности, большие надежды возлагаются на новые возможности по разрешению пространственной структуры нанокристаллов и единичных макромолекул. Предварительные результаты экспериментов, проведенных на приборе LCLS (Калифорния, США) показали принципиальные возможности метода, выявили особенности и дальнейшие направления совершенствования технологии [1,6,7]. В данном разделе мы обсудим особенности определения структуры нанокристаллов и единичных молекул с использованием лазеров на свободных электронах в следующих направлениях: схема эксперимента, получение и доставка образцов в пучок, особенности проведения эксперимента и получаемых данных.

Схема типичного эксперимента

Одним из основных преимуществ лазеров на свободных электронах перед синхротронами третьего поколения является на порядки ($\sim 10^{10}$) большая пиковая светимость при малой длительности импульса и, как следствие, возможность получения статистически значимых дифракционных картин от объектов малого размера – нанокристаллов и даже от единичных макромолекул. Такие возможности значительно снижают требования к качеству образцов и их размеру (при использовании синхротронного излучения, для получения значимых картин рассеяния требуются кристаллы достаточного качества и размера). Однако, взамен возникают задачи, связанные с правильной подготовкой образцов, доставкой их в пучок, а также методикой проведения и анализа самого эксперимента.

Нанокристаллография – область, которая стала возможной, именно, благодаря лазерам на свободных электронах. В методах традиционной кристаллографии, чтобы получить дифракционный сигнал от маленького кристалла зачастую необходимо настолько увеличить мощность излучения, что кристалл подвергается радиационной деградации еще до того, как дифракционная картинка может быть записана [8]. Особенно сложно получить большие, хорошо дифрагирующие кристаллы для мембранных белков – к настоящему моменту было определено лишь 300 уникальных структур таких белков, несмотря на их важность во всех живых клетках. С изобретением фемтосекундного рентгеновского лазера возникла возможность разорвать связь между дозой облучения и пространственным разрешением. Фемтосекундные рентгеновские импульсы могут быть использованы, чтобы “обогнать” даже самые быстрые процессы повреждения. Эту возможность дает использование одиночных импульсов настолько коротких, что они заканчиваются перед проявлением повреждения образца [9]. Эксперименты на приборе FLASH, Германия, подтвердили возможность “дифракции до разрушения” при разрешении до 60 Å на тестовых образцах фиксированных на мембранах нитрида кремния [10].

Первое описание успешного применения фемтосекундной нанокристаллографии для изучения структуры биомакромолекул было представлено в работе [1] на примере суб-микрометровых кристаллов фотосистемы I из цианобактерий, которые доставлялись в пучок с помощью специальной системы впрыска через форсунку. Схема эксперимента приведена на рисунке 2.

В данном эксперименте регистрировались дифракционные картины от отдельных кристаллов из потока впрыскиваемого через 4 микрометровую форсунку в область пучка в вакуум со скоростью 10 мкл/мин. В отличие от криоэлектронной микроскопии или стандартной кристаллографии микрокристаллов, в которых требуется криогенное охлаждение, в данном методе экспериментальный образец находился при комнатной температуре в полностью гидратированном состоянии. Кристалл, пролетающий в момент вспышки лазера через область взаимодействия, приводит к возникновению дифракционной картины, регистрируемой двумя группами детекторов, состоящими из рентгеновских CCD матриц с высоким соотношением сигнал/шум. Сигнал успевают считывать с матрицы при частоте повторения вспышек в 30 Гц.

При энергии фотонного импульса в 1.8 кэВ, количестве фотонов порядка 10^{12} и при длительности импульса в 10, 70, 200 фс нанокристалл испаряется, однако анализируемая область быстро восстанавливается за счет движения жидкостного потока.

В настоящее время исследуются также возможности по изучению единичных крупных структур – вирусов. В работе было продемонстрировано использование ЛСЭ для реконструкции структуры частицы ми-ми вируса размером 40 мкм. Схема эксперимента аналогична таковой для нанокристаллов.

Получение и доставка образцов в пучок

Система впрыска раствора нанокристаллов изображена на рисунке 3 [11]. Форсунка состоит из центрального капилляра, в котором находится жидкий образец и внесшей окружающей трубки, через которую подается сжатый воздух. Раствор попадает из центрального капилляра в устье форсунки и далее ускоряется потоком сжатого воздуха. Вылетая из форсунки, струя, в силу поверхностных сил, превращается в набор капелек. Причем диаметр капелек может быть сильно меньше диаметра форсунки.

Технология такой газодинамической форсунки является на данный момент наиболее подходящей для доставки нанокристаллов и единичных биомолекул в пучок рентгеновского лазера. В отличие от технологии электрораспыления газодинамические форсунки позволяют получить незаряженные микроскопические капельки. В то время, как при электрораспылении капельки становятся заряженным, что может приводить к изменению конформаций биомакромолекул, а также другим нежелательным феноменам. При этом биомакромолекулы, доставляемые в капелях находятся в гидратированной форме, что оказывает несомненно положительное влияние на естественное состояние образца.

Дальнейшее улучшение качества форсунок и уменьшение размера капелек возможно с применением технологий струйной печати. По оценкам работы применение таких технологий позволит значительно (в тысячи раз) уменьшить необходимый объем образца.

Высказываются также предположения относительно использования лазерных ловушек и оптических пинцетов для позиционирования и удержания образцов.

Несмотря на то, что были проведены успешные испытания методики на примере нанокристаллов фотосистемы I, следует отметить, что данный макромолекулярный комплекс является особенным, поскольку кристаллизуется за несколько часов при низкой ионной силе. Для большинства других мембранных белков требуется использовать более высокие концентрации соли, а также полиэтиленгликоль, что в

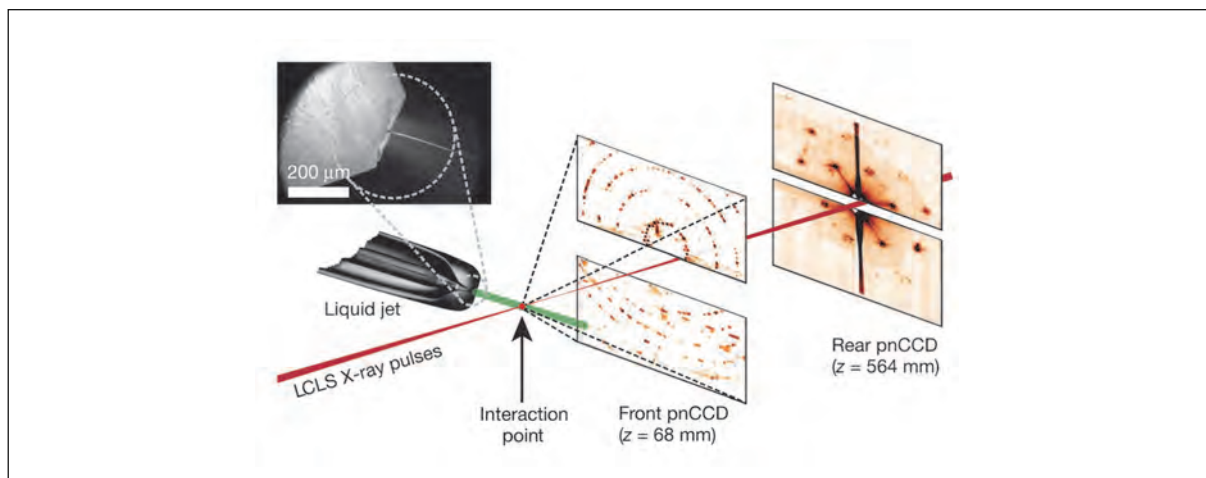


Рис. 2. Схема эксперимента по исследованию структуры нанокристаллов фотосистемы I с использованием лазера на свободных электронах [1]

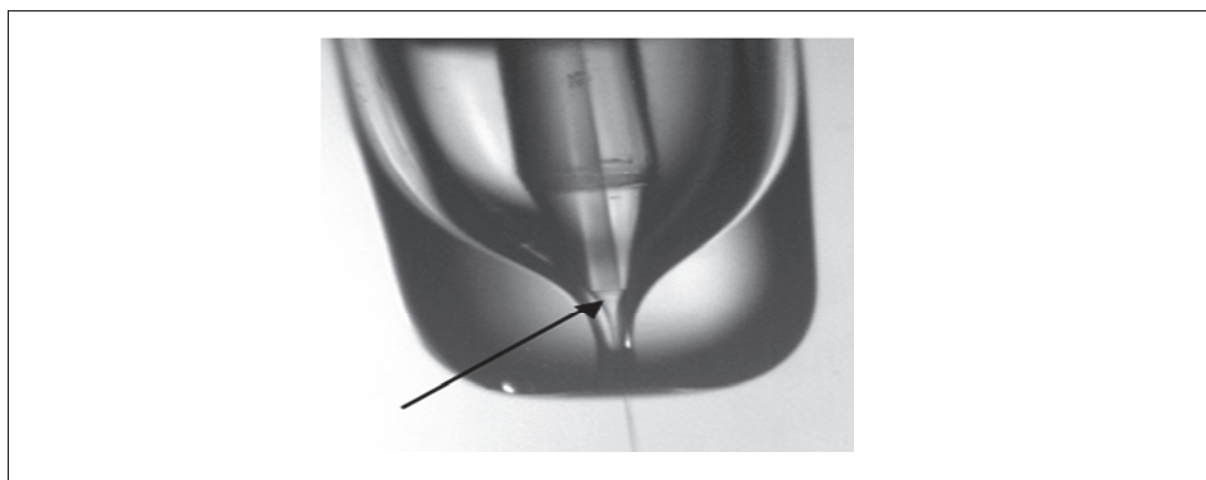


Рис. 3. Изображения форсунки для впрыскивания образца. Стрелка указывает на местоположение раствора с образцом [11]

свою очередь вызывает проблемы с использованием впрыска образцов через форсунку из-за большой вязкости и риска образования микрокристаллов соли, которые забьют выходное отверстие. Одним из наиболее широко используемых методов кристаллизации мембранных белков является метод кристаллизации в липидной фазе (lipidic cubic phase) [12]. Метод позволяет создать окружение белка похожее на липидный бислой и таким образом, увеличить стабильность системы во время кристаллизации. Однако из-за полутвердой природы липидной фазы образец при его впрыске через форсунку не создает поток нужной интенсивности. Недавно было показано, что эту проблему можно решить за счет использования методов кристаллизации в липидной губчатой фазе (lipidic sponge phase), которую можно сформировать путем смешивания моноолеина и воды с третьим агентом, или полиэтиленгликолем, который разбавляет систему до жидкой фазы.

Особенности проведения эксперимента и получаемых данных

Дифракционные картины получаемые при прохождении лазерного рентгеновского излучения через отдельные нанокристаллы содержат лишь “частичные” брэгговские рефлексы, в отличие от экспериментов на синхротроне, где вращение образца на

гониометре позволяет провести угловое усреднение по брэгговскому условию. Таким образом, проблема индексации и совмещения миллионов дифракционных паттернов порождает новые вычислительные сложности, связанные с обработкой многих терабайт данных [13].

Для нанокристаллов размером менее 10 повторений элементарной ячейки появляются направления заметного рассеяния в не-брэгговских направлениях, что затрудняет использование программного обеспечения по автоматическому индексированию рефлексов. В работе [14] для решения этой задачи был предложен алгоритм на основе метода Монте-Карло, основанные на предположении, что все ориентации нанокристалла равновероятны.

На рисунке 4 изображен пример дифракционной картины полученной от ми-ми вируса, характерной особенностью данного типа экспериментов является отсутствие данных в центре детектора, отвечающих за низкочастотные моды исследуемого образца.

Для реконструкции электронной плотности по данным дифракции используется итеративные алгоритмы получения фазы, такие как, RAAR [15] и Shrinkwrap [16].

Методы анализа картины рассеяния при изучении структуры белков

Эксперименты, выполненные с использованием лазеров на свободных электронах, можно условно разделить на два типа, которые принципиально отличаются с точки зрения анализа сигнала [17]. В экспериментах «первого типа» экспериментатору доступно большое количество идентичных копий исследуемого объекта, например, молекулы белка. В экспериментах «второго типа» исследуемый объект является уникальным по своей структуре (например, клетка), и его структура должна быть установлена на основании одного единственного дифракционного изображения. В обоих случаях конкретный исследуемый объект разрушается под действием рентгеновского импульса, но в первом случае, в отличие от второго, есть возможность собрать большое количество (например, $N \sim 100000$ [18]) дифракционных картин от объектов, по-разному ориентированных относительно направления импульса, и на их основании восстановить трехмерную структуру объекта с высоким пространственным разрешением.

При использовании XFEL для анализа структуры белков мы имеем дело с экспериментом «первого типа». Обсуждение методов анализа сигнала мы начнем с эксперимента, относящегося ко «второму типу» [19], т.к. это позволит продемонстрировать некоторые принципиальные особенности. Обсуждаемый эксперимент был выполнен на лазере FLASH (Гамбург, Германия). Длительность импульса составляла 35 фс, число фотонов $\sim 10^{12}$, длина волны 32 нм. Исследуемым объектом была мембрана из нитрида кремния толщиной 20 нм, на которой с помощью сфокусированного ионного пучка был вырезан узор (схема эксперимента представлена на рисунке 5). Были зарегистрированы дифракционные картины от двух последовательных импульсов, они представлены на рисунке 6.

Дифракционные картины, представленные на рисунке 6А и 6Б, были зарегистрированы с интервалом в 20 фс. Первая из них несет в себе информацию о структуре образца (рисунок 6, В) и эта информация может быть восстановлена (рисунок 6, Г). В то же время, изображение, полученное при второй вспышке (рисунок 6, Б) абсолютно неинформативно, т.к. образец был разрушен первой вспышкой.

Наибольший интерес для нас сейчас представляет математическая процедура, которая позволяет перейти от дифракционного изображения (рисунок 6, А) к структуре (рисунок 6, Г). Такой переход осуществляется с помощью обратного преобразования Фурье. Принципиально, что в обсуждаемом случае образец является двумерным и

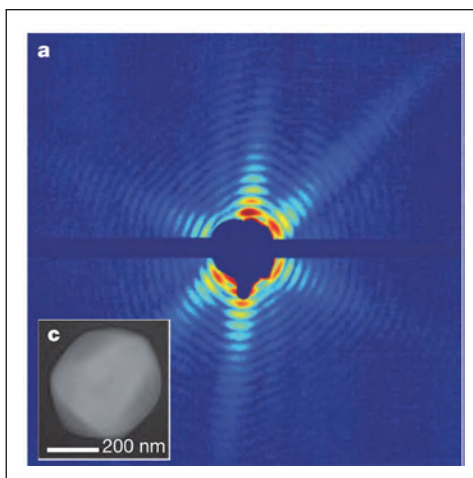


Рис. 4. Пример получаемой дифракционной картины от ми-ми вируса [6].

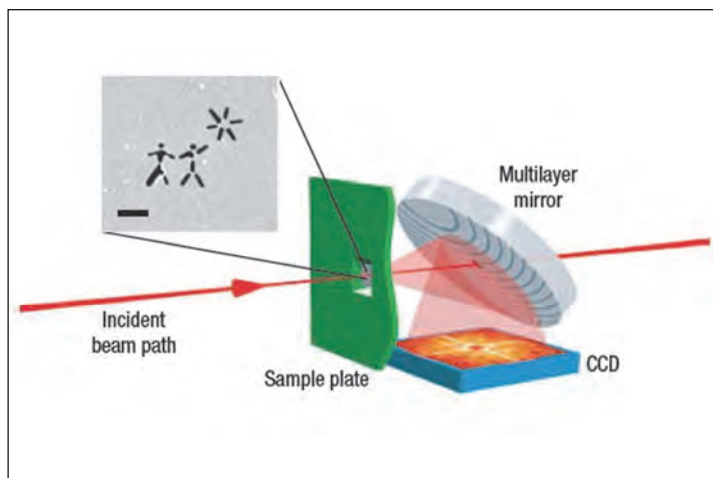


Рис. 5. Схема эксперимента по дифракции рентгеновских лучей на мембране из нитрида кремния

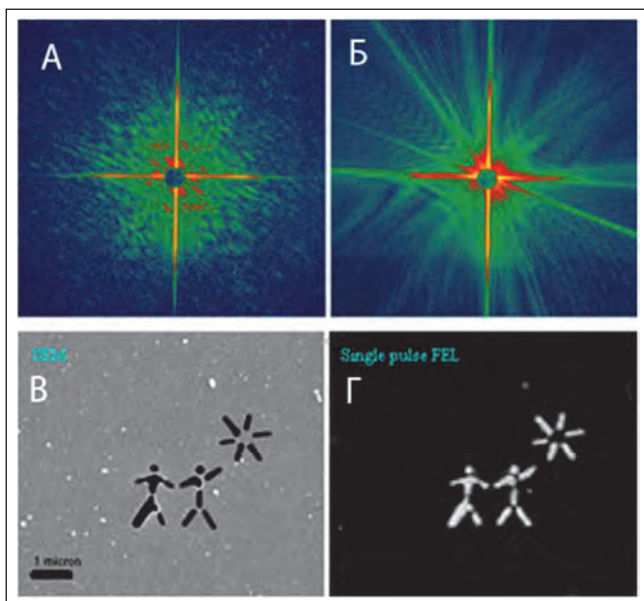


Рис. 6. А – дифракционная картина, полученная при первом импульсе. Б – дифракционная картина, полученная при втором импульсе, после того, как образец был разрушен. В – изображение образца, полученное СЭМ. Г – результат восстановления структуры по дифракционной картине А. Адаптировано из

вопрос о его трехмерной структуре не ставится. В случае с анализом сигнала, полученного от молекул белков, ситуация сложнее, и это будет рассмотрено ниже.

Необходимость использования преобразования Фурье связана с процессом дифракции рентгеновских лучей (рисунок 7). Амплитуда рассеянной волны $A(\mathbf{q})$ связана с электронной плотностью $\rho_e(\mathbf{r})$ соотношением

$$A(\mathbf{q}) = \int \rho_e(\mathbf{r}) \exp(-i\mathbf{q} \cdot \mathbf{r}) d\mathbf{r},$$

где $\mathbf{q}$ – вектор рассеяния. Детектор регистрирует интенсивность, которая пропорциональна квадрату модуля амплитуды

$$S(\mathbf{q}) = \left| \int \rho_e(\vec{r}) e^{-i\vec{q} \cdot \vec{r}} d\vec{r} \right|^2$$

Ясно, что при вычислении модуля теряется информация о фазе $A(\mathbf{q})$, и это порождает так называемую фазовую проблему. Напрямую восстановить распределение электронной плотности

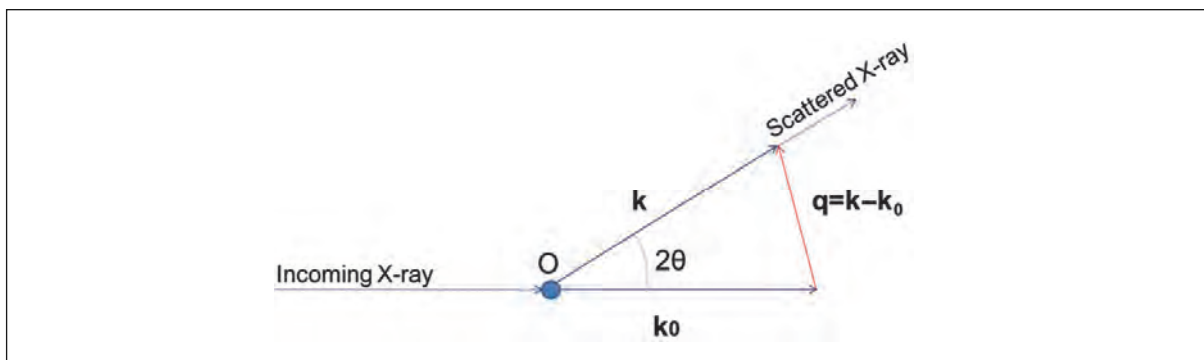


Рис. 7. Рассеяние рентгеновских лучей на объекте. q – вектор рассеяния

$$\rho_e(\mathbf{r}) = \frac{1}{(2\pi)^3} \int A(\mathbf{q}) \exp(i\mathbf{q} \cdot \mathbf{r}) d\mathbf{q}$$

с помощью обратного преобразования Фурье не удастся, т.к. экспериментально регистрируются только амплитуды (точнее, квадраты амплитуд), но не фазы $A(\mathbf{q})$. Решение фазовой проблемы – один из ключевых этапов анализа сигнала при обработке результатов, полученных с помощью PCA и XFEL.

Еще один важный аспект, связанный с преобразованием Фурье, состоит в потере информации о низких пространственных частотах [21]. В центральной части детектора, в которую попадает недифрагированный пучок, обычно делают отверстие, чтобы избежать избыточной засветки детектора. Информация о рентгеновских квантах, рассеянных на малые углы, неизбежно теряется, при этом теряется соответствующая информация о низких пространственных частотах, т.е. о размерах и форме объекта. Чтобы ее восстановить, необходимо использовать данные, полученные независимыми методами, например, ПЭМ, или, в том случае если исследуемым объектом является белок, с помощью моделирования по гомологии.

Важно подчеркнуть, что потеря информации о низких пространственных частотах является принципиальной и фундаментальной особенностью эксперимента, выполненного с помощью XFEL. Поэтому важным является создание алгоритмов интеграции данных, полученных с помощью XFEL с данными, полученными другими методами: электронной микроскопией и молекулярным моделированием.

Кроме обратного преобразования Фурье, при анализе сигнала, зарегистрированного от молекул белков, выполняют еще один принципиальный шаг. Выше мы отмечали, что изучение структуры белков с помощью XFEL проводят в том случае, когда в распоряжении экспериментаторов имеется достаточное количество идентичных между собой экземпляров молекулы белка. Молекулы вводят в пучок рентгеновских квантов в потоке газа [22] и каждый снимок, полученный XFEL, является результатом дифракции пучка на случайно ориентированной молекуле. При восстановлении трехмерной структуры возникает задача совмещения дифракционных изображений, полученных от различных экземпляров исследуемой молекулы, с учетом их возможных поворотов и шумов.

В общем случае два дифракционных изображения, полученных от двух идентичных экземпляров молекул исследуемого белка (рисунок 8), отличаются из-за двух факторов. Во-первых, ориентация молекул относительно падающего излучения не контролируется, и различные ориентации приводят к возникновению различных дифракционных картин. Во-вторых, на каждую дифракционную картину накладывается случайный (и поэтому уникальный) шум [23]. Шумы необходимо учитывать потому, что количество рассеянных рентгеновских квантов мало (в особенности для малых молекул), и каждый единичный квант, попавший на детектор, важен при анализе изображений.

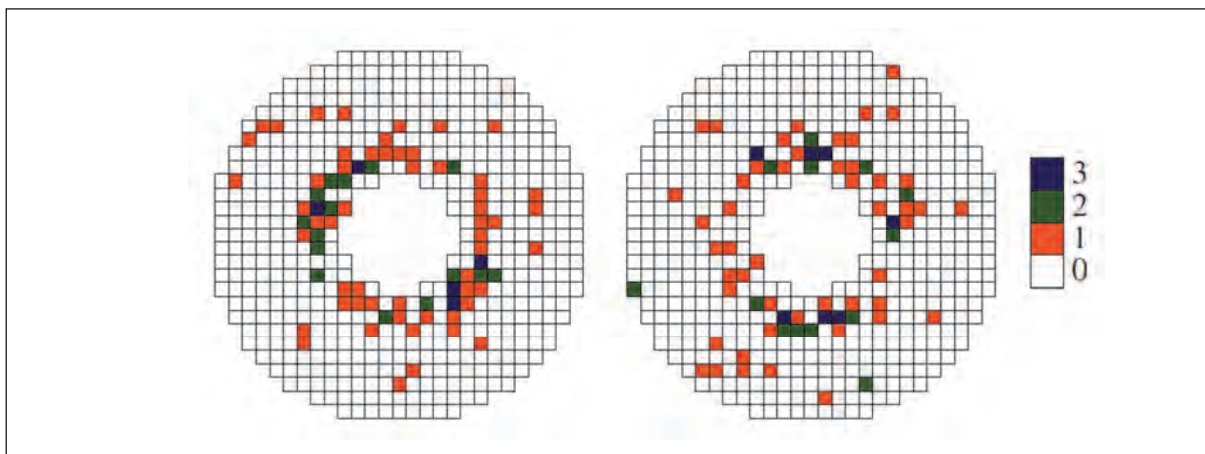


Рис. 8. Пара модельных дифракционных изображений [24]. В центре детектора отверстие для недифрагировавшего пучка. Цвет пикселия показывает количество зарегистрированных рентгеновских квантов (0-3).

Итак, при анализе сигнала от молекул белков возникает следующая задача. Имеется множество дифракционных изображений, которые были получены от молекул, имеющих различную ориентацию. Необходимо получить трехмерное распределение электронной плотности в белке. Это можно сделать двумя принципиально разными путями [25].

Первый путь состоит в том, чтобы к каждому из дифракционных изображений применить обратное преобразование Фурье и получить набор проекций электронной плотности. По этим проекциям далее можно восстановить трехмерное распределение электронной плотности аналогично тому, как это делается в просвечивающей электронной микроскопии.

Второй путь состоит в том, чтобы переход $2D \rightarrow 3D$ выполнить не в прямом пространстве, а в обратном, т.е. совместить дифракционные картины так, чтобы получить трехмерное распределение интенсивности рассеянного излучения. К этому трехмерному распределению интенсивности применяют обратное преобразование Фурье (трехмерное, а не двумерное), и получают трехмерное распределение электронной плотности.

Решение фазовой проблемы при обратном преобразовании Фурье обычно требует итерационного алгоритма. Аналогично, итерационные алгоритмы используются при переходе от двумерных изображений к трехмерным (как в прямом пространстве, так и в обратном). Время, необходимое для реализации любого итерационного алгоритма стремительно растет с увеличением количества изображений, подлежащих обработке. В связи с этим, важной характеристикой каждого используемого способа обработки сигнала является его эффективность с точки зрения затрат времени и вычислительной мощности.

Построение трехмерного распределения интенсивности из двумерных дифракционных изображений (рисунок 9) обычно основано на двух следующих утверждениях:

1. Дифракционные изображения, полученные от одного и того же объекта, с двух близких ракурсов похожи.

2. Два дифракционных изображения, полученных от одного и того же объекта с разных ракурсов, имеют линию, вдоль которой они одинаковы с точностью до шума.

Сопоставим процедуры обработки изображений, которые используются в ПЭМ для перехода $2D \rightarrow 3D$, с процедурами, которые используются для аналогичного перехода при построении трехмерного распределения интенсивности. Во-первых, в ПЭМ

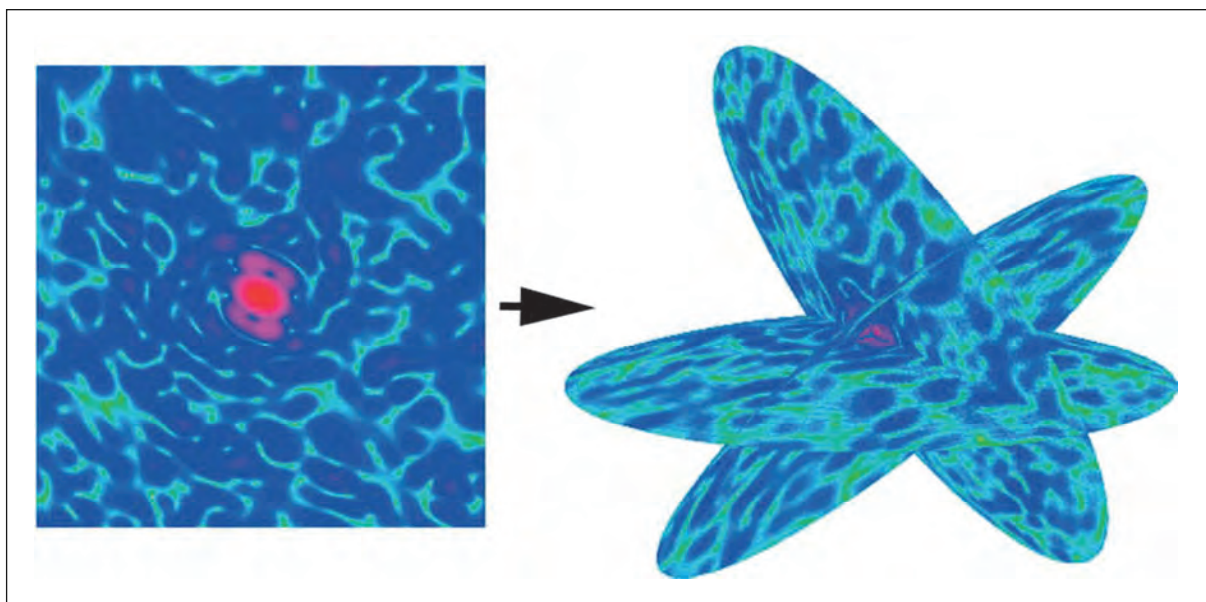


Рис. 9. Схема, иллюстрирующая, как двумерные дифракционные картины совмещаются между собой и формируют трехмерное распределение интенсивности

входными данными являются строго двумерные изображения, а в дифракционной задаче – сечения трехмерного распределения интенсивности фрагментами сферы (сферы Эвальда). Во-вторых, в дифракционной задаче все изображения центрированы, а в ПЭМ у каждого изображения необходимо определить центр. В-третьих, при анализе дифракционных изображений нет понятия фона: каждый пиксель детектора, зарегистрировавший хотя бы один фотон, должен учитываться при обработке. В то же время, при построении трехмерной реконструкции в ПЭМ частицы отделяют от фона, а шум фона влияет на шум сигнала. Наконец, при обработке ПЭМ-изображений используют коррекцию гаммы, а также настройки яркости и контрастности, а к дифракционным изображениям таких операций не применяют. Несмотря на перечисленные различия, логика решения этих двух задач имеет много общего. Поэтому актуально исследование существующих алгоритмов, их сопоставление и разработка новых алгоритмов визуализации и конвертации дифракционных изображений и карт электронной плотности биомакромолекул, в том числе для перехода от двумерных представлений к трехмерным.

Обсудим задачу сортировки дифракционных изображений. Аналитический подход к этой задаче предложен в работе. Каждое дифракционное изображение можно описать вектором $g = \{g_0, g_1, \dots, g_{n-1}\}$, где n – количество пикселей, а каждый элемент g_k это количество фотонов, попавших в соответствующий пиксель. В качестве меры корреляции изображений можно использовать скалярное произведение

$$(g, h) = \sum_{i=0}^{n-1} g_i \cdot h_i$$

Распределения этой величины можно аналитически оценить в двух крайних случаях: если разница между изображениями обусловлена тем, что они получены от разных ориентаций объекта, и если разница между изображениями обусловлена шумами (рисунок 10). Если эти распределения существенно перекрываются, то сортировка изображений невозможна. Если же они слабо перекрываются, то сортировка изображений возможна, и ее достоверность будет тем выше, чем меньше перекрывание распределений. Критерий перекрывания распределений позволяет связать между собой количество пикселей, достижимое разрешение и размер исследуемой частицы.

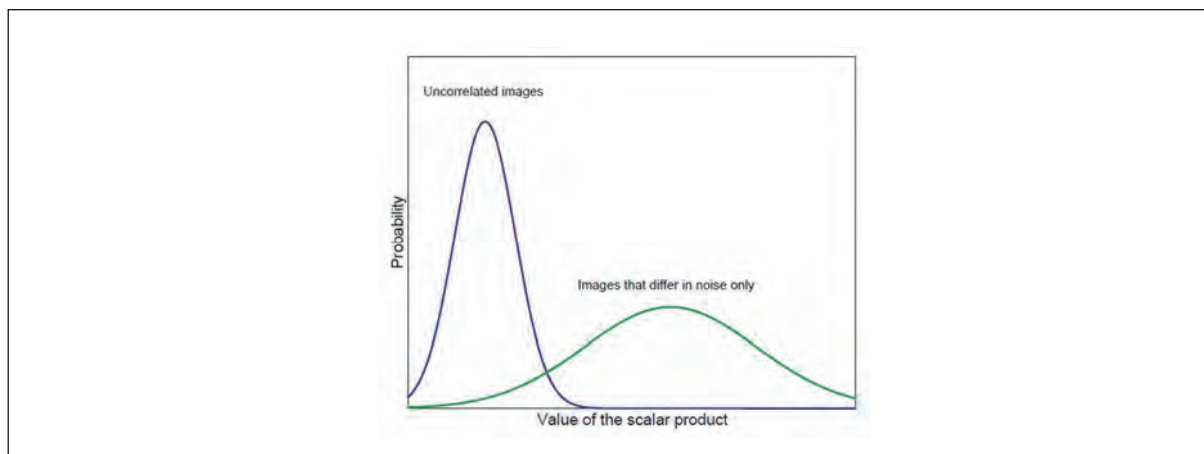


Рис. 10. Графики распределений величины скалярного произведения, характеризующего корреляцию изображений

Результаты, полученные в работе аналитически, были проверены численным моделированием [30]. В качестве меры корреляции было предложено использовать не просто скалярное произведение, а нормированное скалярное произведение

$$(g, h)_2 = \sum_{i=0}^{n-1} g_i \cdot h_i / \sqrt{\sum_{i=0}^{n-1} g_i^2 \cdot \sum_{i=0}^{n-1} h_i^2}$$

Использование такой величины в качестве меры корреляции предпочтительно, т.к. при освещении исследуемого объекта с разных ракурсов общая интенсивность рассеяния может быть различна.

Далее было показано, что в реальных экспериментах за счет того, что ориентации объекта относительно пучка случайны, количество пар изображений, которые будут различаться только из-за шума, существенно меньше, чем количество пар изображений, которые различаются из-за того, что сняты под разными углами. Это означает, что на рисунке 10 будет только один (первый) максимум. Поэтому был предложен новый критерий для сортировки изображений, основанный на предельном угловом разрешении δ_{lim} . Пусть мы хотим восстановить структуру молекулы с размером d_{mol} с разрешением $d_{\text{res}} \ll d_{\text{mol}}$. Тогда $\delta_{\text{lim}} = d_{\text{res}}/d_{\text{mol}}$ – такой угол между ориентациями объекта, что разница между этими дифракционными картинками пренебрежимо мала. Чтобы это определение было корректным, нужно упомянуть, что для любых двух ориентаций твердого тела можно перевести одну в другую путем поворота на некоторый угол. Можно ожидать, что скалярное произведение $(g, h)_2$ большие значения в тех случаях, когда угол между ориентациями объекта мал. Вопрос состоит в том, как выбрать пороговое значение $(g, h)_{\text{crit}}$, при котором скалярное произведение считается достаточно большим. На этот вопрос предложен нетривиальный ответ: выбирают не по критическому значению, а по количеству изображений (из общего числа n), которые имеют $(g, h)_2 > (g, h)_{\text{crit}}$. Это количество изображений, равно $N(\delta_{\text{lim}} - \sin \delta_{\text{lim}})/\pi$, где $N = n(n-1)/2$. Не вдаваясь в обоснование этого критерия (оно связано с ошибкой алгоритма, т.е. с количеством изображений, которые классифицируются неправильно), подчеркнем, что его достоинство в том, что он связан с необходимым в эксперименте разрешением.

Обсудим работу, в которой описан один из наиболее современных алгоритмов перехода от отдельных дифракционных изображений к трехмерному распределению интенсивности. Схема этого алгоритма представлена на рисунке 11. От массива изображений $\tilde{M}$ необходимо перейти к трехмерному распределению интенсивности I . Идея алгоритма состоит в том, что, если бы мы знали приблизительное распределение

I , то по нему можно было бы определить ориентацию каждой из дифракционных картин $\tilde{M}_m$ это позволило бы уточнить распределение I и затем повторить такой цикл несколько раз.

Для описании поворота исследуемой молекулы относительно рентгеновского пучка (его направление принимают за Z) используют углы Эйлера Φ , Θ и Ψ . В простейшем случае, если он не поляризован, изменение угла Ψ приводит к повороту дифракционной картины вокруг центра.

Из трехмерного распределения интенсивности I получают N_r сечений фрагментами сферы Эвальда I_r . Дифракционные изображения $\tilde{M}_m$ сопоставляют с сечениями I_r путем вычисления коэффициентов корреляции $c_{mrj'}$, где m и r – номера изображения и сечения (по r можно однозначно установить ориентацию сечения), а j' характеризует угол поворота. В данном случае коэффициенты корреляции вычисляют в полярной системе координат: вычисляют корреляцию между окружностями (угол рассеяния $\theta = \text{const}$) и затем усредняют по всем окружностям. Зная коэффициенты корреляции $c_{mrj'}$, можно определить r и j' , при которых они максимальны. По этим индексам r и j' дифракционные картины $\tilde{M}_m$ поворачивают в их наиболее вероятные положения и получают новое распределение I . После этого начинают новую итерацию с новым распределением I . Критерием сходимости является рост коэффициентов $c_{mrj'}$.

Сходимость алгоритма проверяли на примере периплазматической нитрат редуктазы (масса 109 кДа, диаметр молекулы ~ 91 А, PDB ID: 3ML1). По известной структуре были смоделированы $N_r = 100000$ дифракционных изображений; алгоритм запускали, используя случайное распределение I . Он сходил за 9-25 итераций, на каждую итерацию требовалось не более 25 минут, при использовании двухъядерного компьютера Intel Xeon X5680 с четырьмя видеокартами Nvidia GTX 480. На примере нитрат редуктазы по трехмерному распределению интенсивности с помощью обратного преобразования Фурье было восстановлено распределение электронной плотности – оно отлично совпало со структурой молекулы.

Еще раз подчеркнем, что общепризнанного алгоритма для построения трехмерного распределения по двумерным распределениям на настоящий момент не существует. Это одинаково применимо для прямого пространства (распределений электронной плотности) и для обратного пространства (распределений интенсивности дифракции). Разработка новых алгоритмов, обеспечивающих быструю сходимость и лучшее разрешение, является чрезвычайно актуальной задачей.

Перейдем к задаче определения фаз при обратном преобразовании Фурье. Обсуждение следует начать с алгоритма, предложенного еще в 1978 году [31], потому что он так или иначе является фундаментальной основой всех последующих алгоритмов вплоть до самых современных. Пусть имеется функция $f(x)$, про которую известно, что она вещественная, неотрицательная, и отлична от нуля в ограниченной области S . Фурье-образ этой функции $F(u)$, и нам необходимо определить $f(x)$ по известному модулю $|F(u)|$. Это можно сделать, например, с помощью итерационного алгоритма, схема которого представлена на рисунке 12 а.

Мы пытаемся аппроксимировать неизвестную функцию $f(x)$ сконструированной функцией $g(x)$, которая строится следующим способом. На первом шаге используют случайную (неотрицательную и отличную от нуля только в ограниченной области S) функцию $g_1(x)$. На k -ом шаге алгоритма от функции $g_k(x)$ берут Фурье-образ $G(u)$ и, сохраняя фазы, изменяют модуль на $|F(u)|$. С полученной функцией $G'(u)$ выполняют обратное преобразование Фурье, получают $g_k'(x)$. Новая итерация выполняется для функции $g_{k+1}(x)$, которая равна $g_k'(x)$ в области S и равна нулю вне этой области. Таким образом, в прямом пространстве функцию «загоняют» в известные пределы, в которых она отлична от нуля:

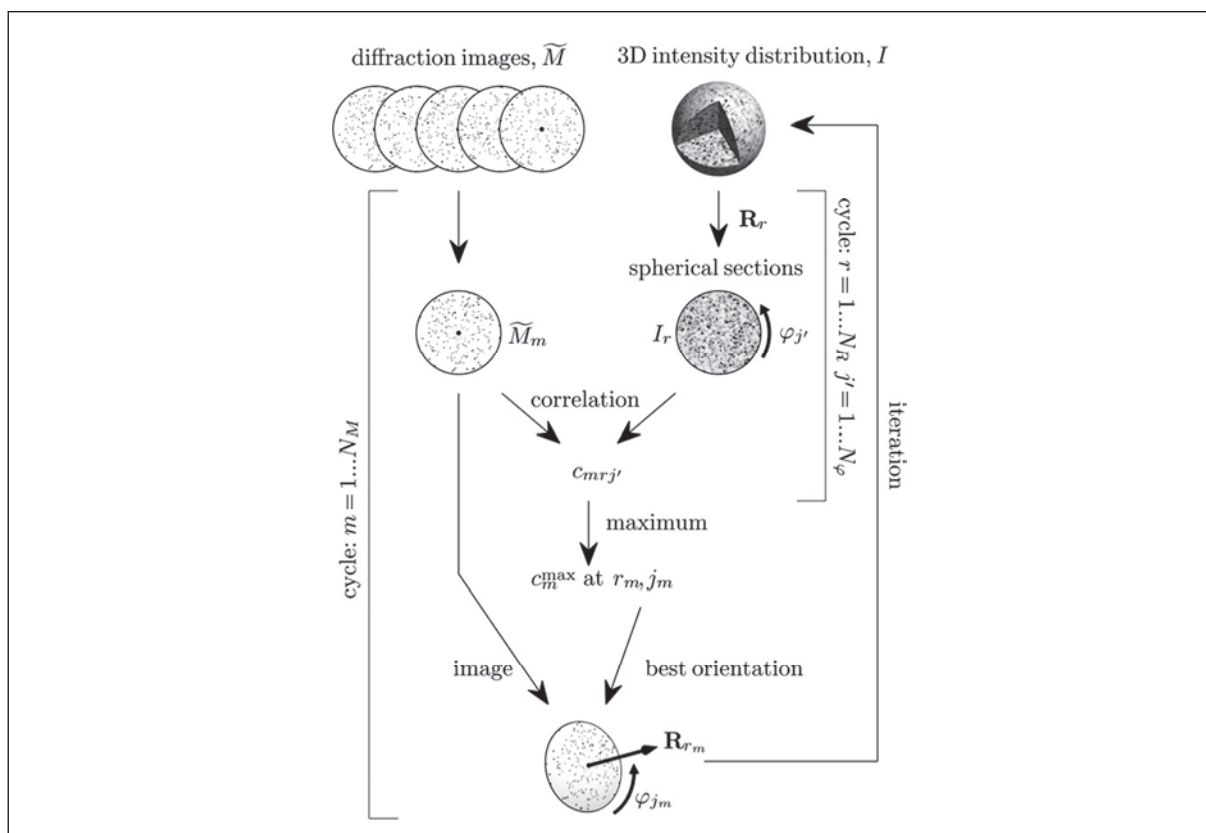


Рис. 11. Схема итерационного алгоритма для определения трехмерной интенсивности рассеяния I по набору дифракционных изображений Пара модельных дифракционных изображений [24]. В центре детектора отверстие для недифрагировавшего пучка. Цвет пикселия показывает количество зарегистрированных рентгеновских квантов (0-3).

$$g_{k+1}(x) = \begin{cases} g'_k(x), & x \in S \\ 0, & x \notin S \end{cases}$$

Мера ошибки такого алгоритма может быть вычислена как в прямом пространстве, так и в обратном. В прямом пространстве мерой ошибки является $|g(x)|$ вне области S , т.е. $E_{0k}^2 = \sum_{x \notin S} (g'_k(x))^2$

В обратном пространстве мерой ошибки является

$$E_{Fk}^2 = \sum_u (|F(u)| - |G_k(u)|)^2,$$

т.е. величина отклонения модуля Фурье-образа $|G(u)|$ построенной нами функции $g(x)$ от экспериментально измеренной величины $|F(u)|$. Описанный алгоритм называют алгоритмом минимизации ошибок, т.к. он обеспечивает монотонное уменьшение обеих ошибок с каждым шагом. Сходимость этого алгоритма доказана для любого начального приближения $g_1(x)$.

Алгоритм минимизации ошибок, хотя и обеспечивает возможность определения объекта по дифракционной картине, но, как показывают модельные эксперименты, сходится чрезвычайно медленно и потому нуждается в оптимизации. Один из путей оптимизации состоит в том, чтобы рассматривать переход в обратное пространство и обратно, а также операции, выполняемые в обратном пространстве, как систему ввода-вывода. В обратном пространстве такой алгоритм аналогичен алгоритму минимизации ошибок (рисунки 12 а, б), а в реальном пространстве он существенно отличается.

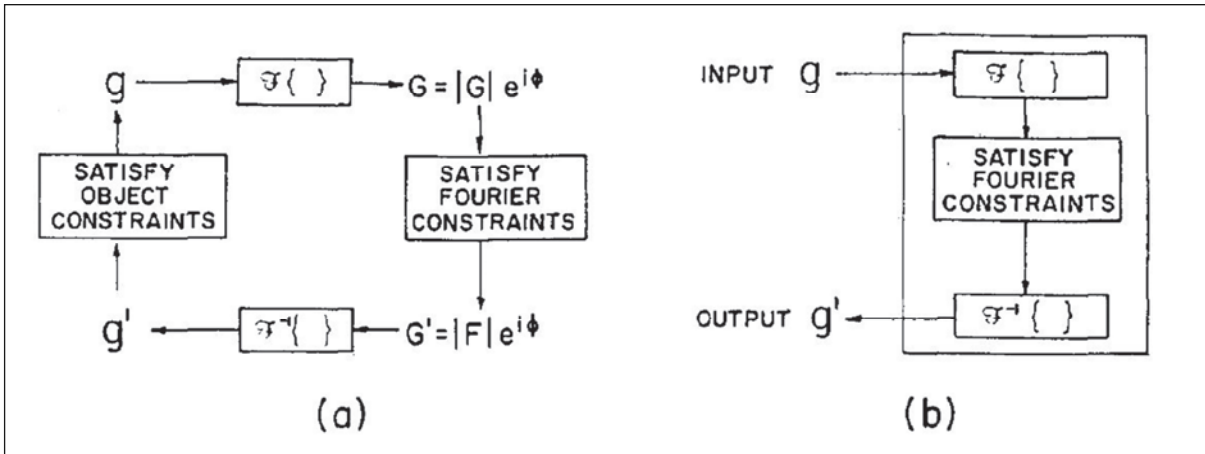


Рис. 12. Алгоритмы определения фаз по дифракционному изображению при обратном преобразовании Фурье. (а) – алгоритм минимизации ошибок, (б) – алгоритм ввода-вывода

Пусть на каждом шаге (на каждой итерации) новая функция $g_{k+1}(x)$ строится по следующей формуле (S – область, в которой функция $g(x)$ может быть отлична от нуля, за пределами S функция $g(x)$ должна обратиться в ноль):

$$g_{k+1}(x) = \begin{cases} g_k(x), & x \in S \\ g_k(x) - \beta g'_k(x), & x \notin S \end{cases}$$

В этом алгоритме β – число, которое является управляющим параметром. Алгоритм «постепенно подталкивает» функцию g к тому, чтобы вне области S она обратилась в ноль, причем это делается за счет изменения той функции, которая подается на вход преобразования Фурье. Это называется базовым алгоритмом ввода вывода (basic input-output). Однако наибольшее распространение получил гибридный алгоритм ввода-вывода НЮ (hybrid input-output), в котором в области S функция $g_{k+1}(x)$ отличается от $g_k(x)$:

$$g_{k+1}(x) = \begin{cases} g'_k(x), & x \in S \\ g_k(x) - \beta g'_k(x), & x \notin S \end{cases}$$

Для алгоритмов ввода-вывода подаваемая на вход функция $g_k(x)$ не является оценкой для функции $f(x)$, поэтому ошибку оценивают только в реальном пространстве и только по выходной функции $g'(x)$ – по указанной выше формуле

$$E_{0k}^2 = \sum_{x \notin S} (g'_k(x))^2$$

С точки зрения скорости сходимости, гибридный алгоритм ввода-вывода является одним из лидеров и поэтому активно используется. В то же время, идет активная работа по оптимизации существующих алгоритмов.

Направления оптимизации алгоритмов включают в себя, в том числе, способы выбора начального приближения $g_1(x)$ и способы сужения области-носителя S . Основные алгоритмы, используемые для решения фазовой проблемы при анализе дифракционных изображений, реализованы в программе HAWK.

Таким образом, перед исследователем стоит задача восстановления трехмерного распределения электронной плотности по множеству дифракционных изображений, полученных от отдельных молекул. При решении этой задачи необходимо сделать два принципиальных шага: выполнить обратное преобразование Фурье (при этом решить сопряженную с ним фазовую проблему), и перейти от двумерных представлений к трехмерным. Эти шаги могут быть выполнены в различном порядке. Для решения фазовой проблемы существуют известные алгоритмы, однако они нуждаются в

оптимизации и дальнейшей разработке – в первую очередь для повышения их скорости. Один из путей оптимизация состоит в интеграции данных рентгеновского рассеяния с данными, полученными другими методами, для получения информации о низких пространственных частотах и выбора начального приближения. Переход от двумерных представлений к трехмерным может выполняться как в прямом пространстве, так и в обратном. В обоих случаях используются итерационные алгоритмы, которые также нуждаются в дальнейшей разработке и оптимизации.

Методы анализа формы белковых структур

Полезной задачей является сравнение заданной формы с известным набором форм и выбор в некотором смысле наиболее близкой формы. Можно выделить пять основных методов решения данной задачи:

- сравнение простейших свойств формы: связность, протяжённость, компактность;
- метода, основанные на анализе коэффициентов некоторого преобразования (Фурье и т.п.);
- инварианты формы как точечного множества, то есть, представление рассматриваемого объекта в виде графа и исследование последнего;
- дифференциально-геометрические инварианты поверхности рассматриваемого объекта;
- инварианты моментов, то есть, специальные функции-моменты исследуемой формы.

В зависимости от решаемой задачи методы решения могут значительно отличаться.

Наиболее широко сегодня используется последний из перечисленных методов. Он заключается в построении инвариантов, основанных на т.н. функциях моментов, каждая из которых имеет вид интеграла

$$M^{(f)} = \iint_D p(x, y) f(x, y) dx dy \quad (1)$$

Здесь функция f является функцией-изображением изучаемой формы, а p – весовая функция, для которой вычисляется момент (в трёхмерном случае определение аналогичное с точностью числа переменных). Поскольку один момент малоинформативен, часто вместо одной функции p рассматривается целый класс $p_{i,j}(x, y)$ (базис в некотором пространстве функций), моменты от которых изучаются совместно. Обычно когда речь идёт о моментах, подразумевается, что функции $p_{i,j}(x, y)$ – многочлены от x и y . Ясно, что выбор такого набора не является однозначным. Более того, в зависимости от задачи этот базис варьируется и его вид играет важную роль в каждом частном случае. Если в качестве функций p взять $p_{ij} = x^i y^j$, то полученные инварианты

$$m_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^p y^q f(x, y) dx dy \quad (*)$$

называются геометрическими моментами, через которые обычно выражают инвариантные (как, впрочем, и остальные типы моментов).

Вспомним, что наша цель – различать объекты как минимум с точностью до сдвигов, поворотов и масштабирования. Другими словами, необходимо потребовать, чтобы выражение (*) сохраняло своё значение при аффинных преобразованиях пространства, имеющих вид

$$x' = kRx + d,$$

где $k > 0$ – константа растяжения, R – ортогональная матрица, d – вектор смещения начала координат. Моменты, удовлетворяющие таким свойством называются TRS-инвариантными (translation-rotation-scaling). Сформулированное определение момента пока, однако, не является инвариантным относительно введенных выше преобразований. Для смещения объекта и его гомотетии инварианты придумываются достаточно просто. Выражения для поворотов в двумерном случае впервые даны работе Ну. Инвариантность относительно поворота является достаточно важной чертой распознающего механизма, поэтому его работа вызвала достаточно большой резонанс.

Первые попытку обобщить идеи Ну на трёхмерный (n -мерный) случай были предприняты в []. Изложенная выше информация в более подробном виде изложена, например, в [].

Что касается количества моментов, которое необходимо вычислить для удовлетворительного определения формы, то здесь задача, по-видимому, сильно зависит выбираемого базиса в пространстве функций. В работе [36] используется базис из обычных сферических гармоник и показано, что необходимо около 600 моментов для грубого описания формы объекта, и около 16000 для достаточно подробного.

Применение указанных методов к задаче распознавания биологических структур может осуществляться несколькими способами. Отметим, к примеру, что метод моментов имеет то преимущество, что все вычисления можно провести достаточно быстро. Недостатком же этого метода является, в частности, глобальность всех вычисляемых параметров. В этом отношении, для определения, например, субъединиц или единичных аминокислот белка необходимо проводить некоторые дополнительные действия, поскольку описание с помощью моментов носит глобальный характер.

В работе приведен метод распознавания плоских групп (например, в макромолекулах) по карте электронной плотности. Этот шаг весьма важен для распознавания молекул. Так, описание составных частей объекта может оказаться плодотворным при распознавании.

Всё вышесказанное наводит на мысль о том, что для продуктивного распознавания макромолекулярных объектов имеет смысл сочетать различные (локальные и глобальные) алгоритмы, комбинируя и сравнивая полученные результаты.

Вскоре после создания стереохимической теории молекул было отмечено, что возникающие при этом геометрические образы хорошо описываются с помощью графов. В свою очередь симметричные свойства молекул соответствуют симметриям графов. Последние являются конечными группами, т.е. подгруппами группами перестановок – симметрических групп.

Метод Пойа в теории графов, сводит задачу перечисления графов к комбинаторной задаче перечисления при наличии групповой симметрии. Таким образом, решается и задача перечисления изомеров в стереохимии. Ответ формулируется в терминах производящей функции. Таким образом, комбинаторная задача перечисления с учётом групповой симметрии, в частности перечисления графов, оказывается сведённой к алгебраической задаче из теории конечных групп. Последняя, в принципе, допускает всегда эффективное решение, например перебором. На практике, для больших систем, поиск решения может быть затруднителен.

Метод Пойа, в его непосредственном виде, позволяет найти только число изомеров. Их стереохимический вид требует дополнительного исследования. Сама по себе идея симметричного анализа графа может быть полезна.

3. Разработка методов и алгоритмов обработки экспериментальной информации о распределении электронной плотности в биомакромолекулах и их комплексах

Результатом эксперимента на XFEL является дифракционное изображение, полученное от исследуемого трехмерного объекта в случайной ориентации относительно рентгеновского пучка, причем обычно экспериментатор работает с большим количеством ($N \sim 100000$ или более) таких изображений. Каждое изображение представляет собой матрицу чисел, количество элементов которой равно количеству пикселей детектора. Для представления этой матрицы используют графический файл, в котором цвет пикселя (или оттенок цвета) однозначно связан с количеством рентгеновских квантов, попавших в соответствующий пиксель детектора. Обсуждаемые дифракционные картины обладают следующими особенностями:

1. Центральная часть детектора либо закрыта непрозрачной для пучка пластиной (beamstop), либо содержит отверстие для недифрагировавшего пучка. В результате на дифракционном изображении в центральной части находится неинформативная область. Это равносильно тому, что дифракционное изображение на аппаратном уровне обработано фильтром низких частот (high-pass). Информация о низких пространственных частотах (о размерах и форме объекта) может быть получена, например, с использованием комплементарных аналитических методов, в первую очередь, просвечивающей электронной микроскопии.

2. Для представления дифракционных изображений часто используют логарифмический масштаб. Фактически вместо элементов матрицы a_{ij} отображают логарифмы $\lg(1+|a_{ij}|)$. Это позволяет отобразить на одном изображении существенно различные значения интенсивности дифракции.

3. Дифракционные изображения обычно являются зашумленными, и это значительно усложняет их сортировку. Обычно предполагается, что шум подчиняется распределению Пуассона, т.е. вероятность того, что шумовой сигнал в конкретном пикселе равен m , равна

$$P(m) = \frac{\chi^m}{m!} e^{-\chi}, \text{ где } \chi - \text{средний уровень шума на изображении.}$$

4. При проведении дифракционных экспериментов с отдельными молекулами необходимо добиться того, чтобы оцифровка дифракционной картины удовлетворяла условию дискретизации. Чтобы прояснить это условие, проследим за рассуждениями о решении фазовой проблемы, приведенными в работе [40].

Пусть функции $f(\mathbf{x})$ соответствует Фурье-образ $F(\mathbf{k})$:

$$F(\mathbf{k}) = \sum_{\mathbf{x}=0}^{N-1} f(\mathbf{x}) \exp(2\pi i \mathbf{k} \cdot \mathbf{x}/N)$$

В общем случае функция $f(\mathbf{x})$ является комплексной. Из эксперимента известен модуль Фурье-образа $|F(\mathbf{k})|$, причем только на ограниченном отрезке значений $\mathbf{k}$. Если мы пытаемся определить $f(\mathbf{x})$, то мы не сможем отличить ее от функций $f(\mathbf{x})$, $f(\mathbf{x}+\mathbf{x}_0)e^{i\theta}$ и $f^*(\mathbf{x}+\mathbf{x}_0)e^{i\theta}$, где $\mathbf{x}_0$ и θ вещественные константы, $*$ – означает комплексное сопряжение (для всех этих функций Фурье-образы $|F(\mathbf{k})|$ совпадают).

Пусть функции $f(\mathbf{x})$ и $F(\mathbf{k})$ определены в N точках. Расширим область определения функции $f(\mathbf{x})$ и потребуем, чтобы в крайних точках этой расширенной области определения функция была равна нулю $f(\mathbf{x})=0$. Например, будем искать функцию $\rho(\mathbf{x})$, которая удовлетворяет условию:

$$\rho(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & 0 \leq \mathbf{x} \leq N-1 \\ 0 & N \leq \mathbf{x} \leq 2N-1 \end{cases}$$

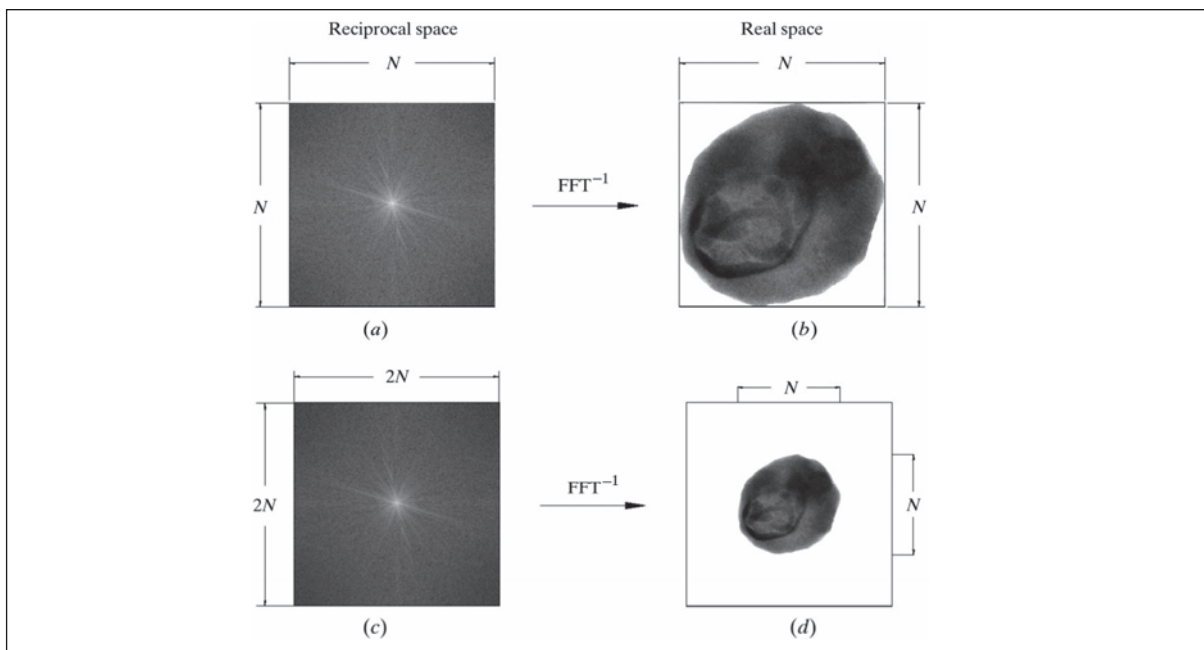


Рис. 13. Влияние дискретизации на изображения, получаемые Фурье-преобразованием.
 (a,b) – Фурье-образ размером $N \times N$ используется для восстановления объекта размером $N \times N$.
 (c,d) – Фурье-образ размером $2N \times 2N$ используется для восстановления объекта размером $N \times N$, вписанного в область $2N \times 2N$

Можно ввести величину σ , которая является мерой избыточности дискретизации (oversampling):

$$\sigma = \frac{\text{Количество точек в которых определена } \rho(x)}{\text{Количество точек в которых определена } \rho(x) \neq 0}$$

Восстановление фазы возможно в том случае, если $\sigma > 2$ по каждому измерению. Например, если функция $f(x)$ одномерная и определена на множестве из N точек, то требование $\sigma > 2$ означает, что величина $|F(\mathbf{k})|$ должна быть измерена минимум в $2N$ точках. Если функция $f(\mathbf{x})$ задана в трехмерном пространстве на N^3 точках, то $|F(\mathbf{k})|$ должна быть задана минимум в $8N^3$ точках. При разработке алгоритма мы будем предполагать, что обрабатываемые изображения удовлетворяют условию $\sigma > 2$. Это наглядно представлено на рисунке 13.

5. Обычно предполагается, что трехмерное распределение электронной плотности, от которого получают дифракционную картину, вещественно – при этом Фурье-образ будет центрально-симметричным.

Алгоритм обработки информации в эксперименте по определению распределения электронной плотности в биомакромолекулах и их комплексах должен включать два принципиальных шага – переход от двумерных представлений к трехмерным и обратное преобразование Фурье, которое, в свою очередь, связано с необходимостью решения фазовой проблемы. На основании проведенного анализа литературы можно предложить следующий принципиальный алгоритм (рисунок 14). Два обсуждаемых принципиальных шага могут выполняться в различном порядке.

Проанализируем некоторые принципиальные особенности каждого из основных шагов.

Начнем с обратного преобразования Фурье, выполняемого в 2D. В аналитическом обзоре обсуждались некоторые алгоритмы, которые могут использоваться для его выполнения. Смысл этих алгоритмов в 2D и в 3D аналогичен, поэтому можно обсудить их вместе. Наиболее широко применяются три из них: НПО (hybrid input-output, он

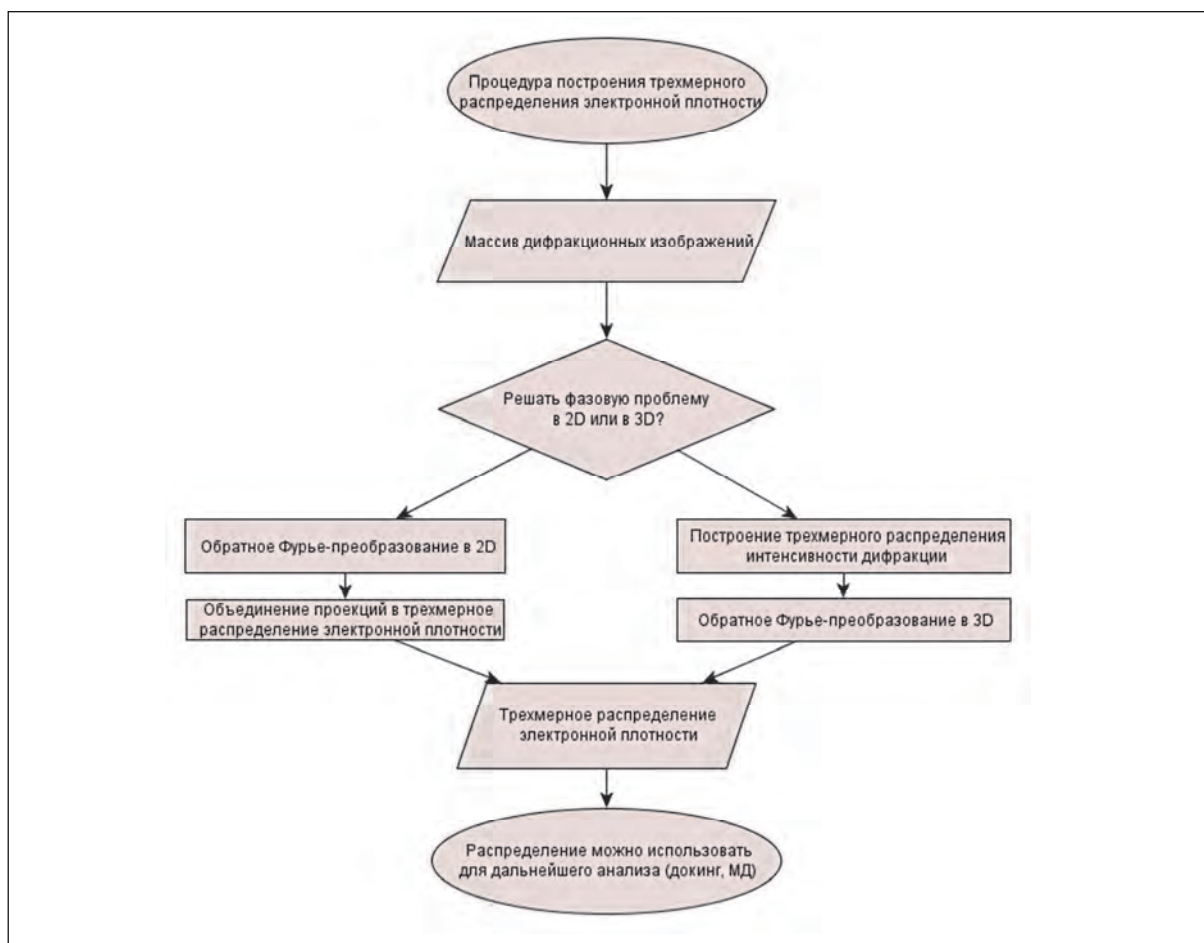


Рис. 14. Принципиальная схема алгоритма обработки дифракционных изображений

обсуждался ранее), RAAR (Relaxed averaged alternating reflections, он описан в работе) и difference map (). Применительно к задаче восстановления фазы в экспериментах на XFEL, смысл этих алгоритмов в том, чтобы построить функцию $f(\mathbf{r})$, которая удовлетворяет двум условиям:

1. Условие в прямом пространстве: функция $f(\mathbf{r})$ вещественна, положительна и отлична от нуля только внутри ограниченного носителя – области S , а за пределами S равна нулю.

2. Условие в обратном пространстве: модуль Фурье-образа функции $f(\mathbf{r})$ равен $|F(\mathbf{k})|$, известному из эксперимента.

В основе всех трех алгоритмов лежат итерационные процедуры: строится последовательность функций $g_k(\mathbf{r})$, $k = 0, 1, 2, \dots$, причем на каждом шаге получается новая оценка для функции $f(\mathbf{r})$. Приведем примеры работы этих алгоритмов.

Для их тестирования были выбраны два изображения, представленные на рисунке 15. В основе обоих изображений лежит один и тот же решетчатый узор размером 64×64 пикселя. Перед тем, как брать Фурье-образ, вокруг этого узора добавляли чистые поля, в результате получили два изображения размером 128×128 и 256×256 пикселей (рисунок 15, соответственно, А и Б). От этих изображений взяли Фурье-преобразование и вычислили спектральную плотность мощности, т.е. $|F(\mathbf{k})|^2$. Эта спектральная плотность мощности является простейшим случаем смоделированного дифракционного изображения. Ее обрабатывали в программе восстановления фаз HAWK с использованием различных алгоритмов. Во всех алгоритмах использовали требования вещественности и неотрицательности исходной функции. Количество итераций составляло $N=4000$.

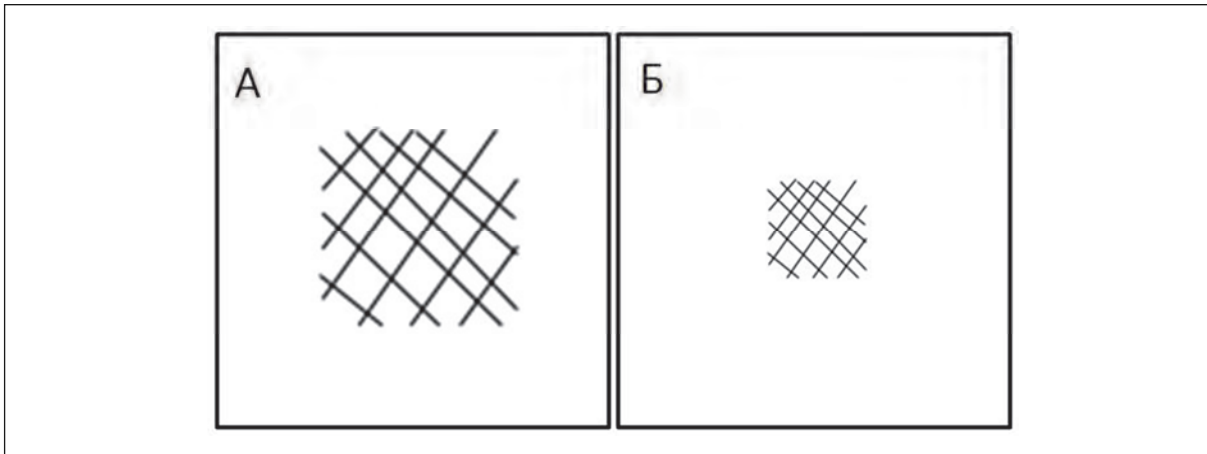


Рис. 15. Тестовое изображение для проверки алгоритмов восстановления фазы. А – линейный размер кадра в 2 раза больше содержательной части изображения. Б – линейный размер кадра в 4 раза больше содержательной части изображения

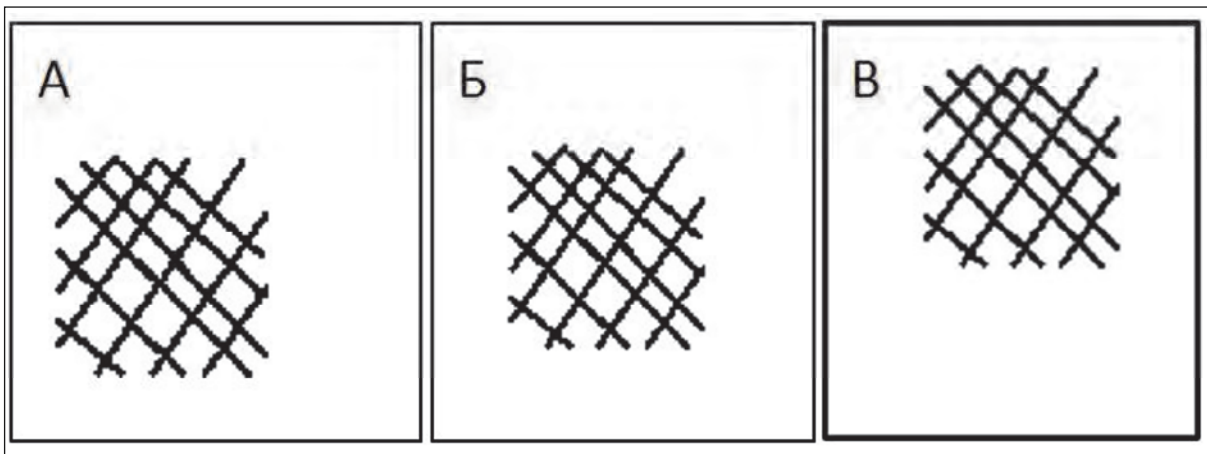


Рис. 16. Результаты восстановления тестового изображения размером 128x128 пикселей с использованием различных алгоритмов. А – НЮ, Б – RAAR, В – difference map

Рассмотрим результаты, полученные при восстановлении меньшего по размерам изображения (рисунок 16). Видно, что все три алгоритма хорошо справились с задачей восстановления изображения.

В отличие от исходного изображения, восстановленный вычислительной процедурой результат не центрирован. Действительно, циклический сдвиг вещественного сигнала приводит к повороту фазового спектра, но не изменяет амплитудного спектра, т.е., другими словами, исходное и сдвинутое изображение имеют одинаковые $|F(\mathbf{k})|^2$.

В некоторых случаях алгоритм сходил к изображению, которое является зеркальным отражением исходного. Это связано с центральной симметрией Фурье-образа.

Количественной мерой сходимости алгоритмов являются ошибки, вычисленные в прямом и обратном пространстве. В прямом пространстве это

$$E_F^2 = \sum_k (|F(\vec{k})| - |G_n(\vec{k})|)^2$$

величина функции $g(\mathbf{r})$ вне области носителя S , n – номер итерации. В обратном пространстве ошибка вычисляется как отклонение $G_n(\mathbf{k})$ – Фурье-образа пробной функции $g(\mathbf{r})$ – от известного $|F(\mathbf{k})|$. Проследим, как уменьшаются ошибки для каждого из обсуждаемых алгоритмов (рисунки 17-19).

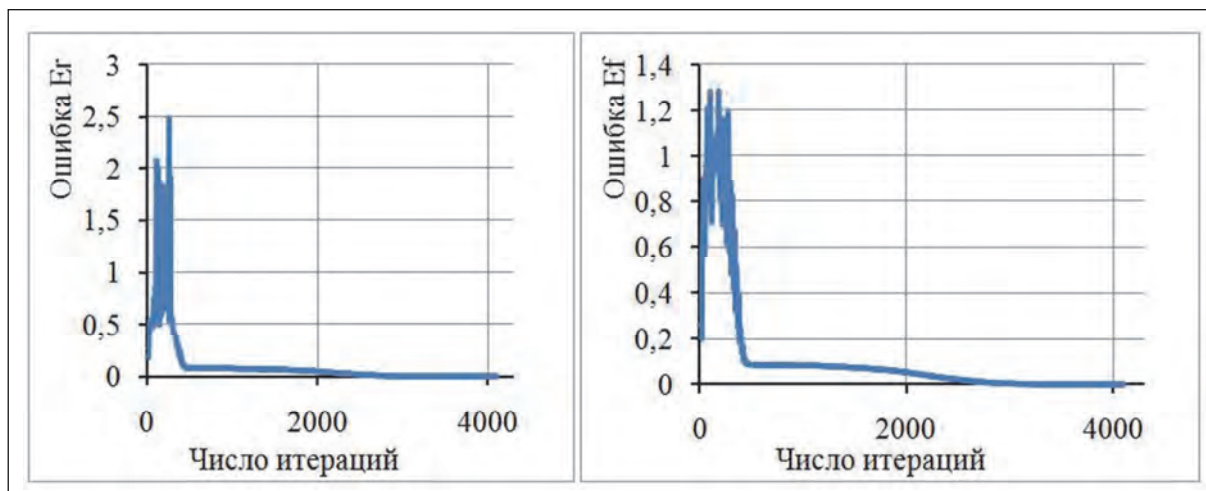


Рис. 17. Графики зависимости ошибок E_g и E_f от числа итераций, алгоритм НЮ. Размер изображения 128x128 пикселей

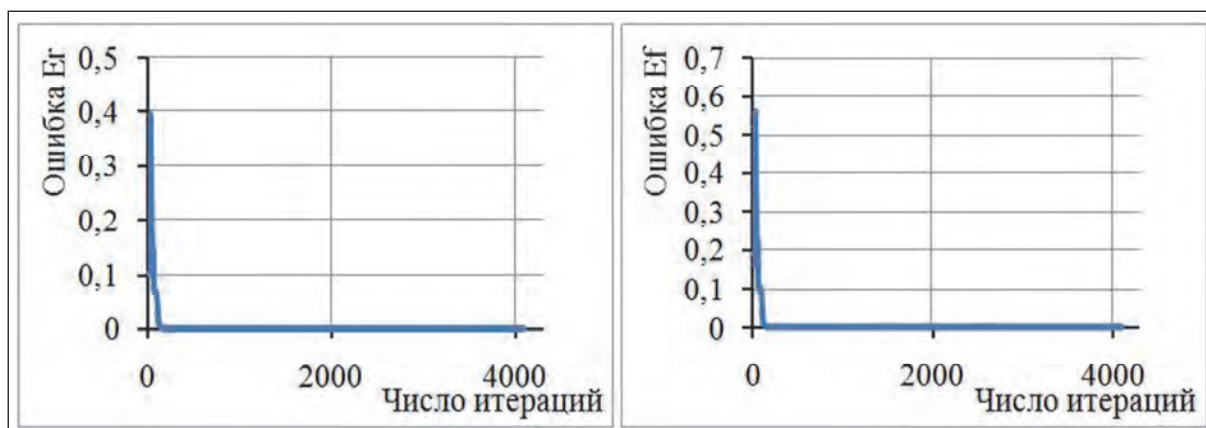


Рис. 18. График зависимости ошибки E_g и E_f от числа итераций, алгоритм RAAR. Размер изображения 128x128 пикселей

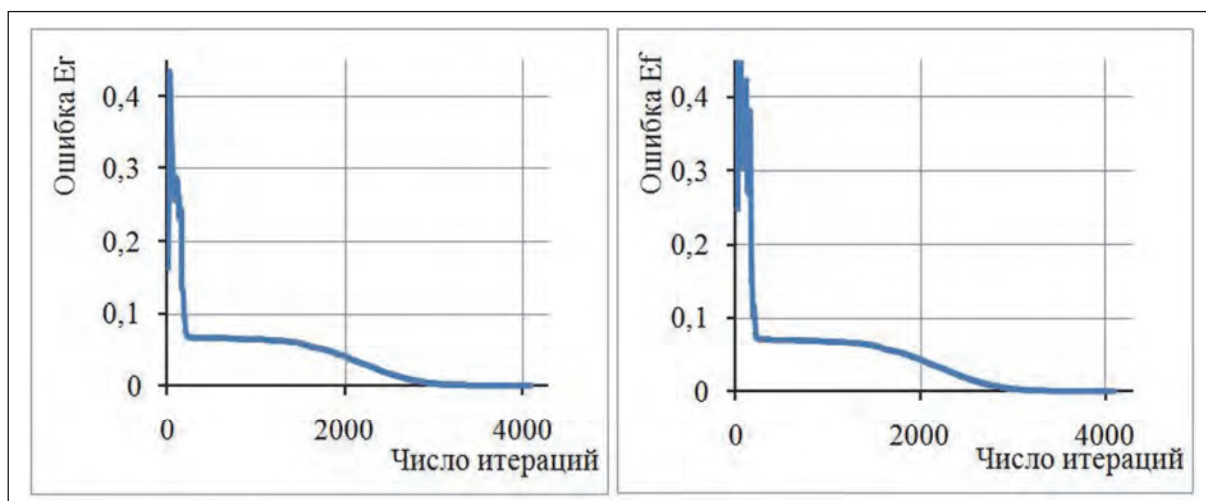


Рис. 19. График зависимости ошибки E_g и E_f от числа итераций, алгоритм difference map. Размер изображения 128x128 пикселей

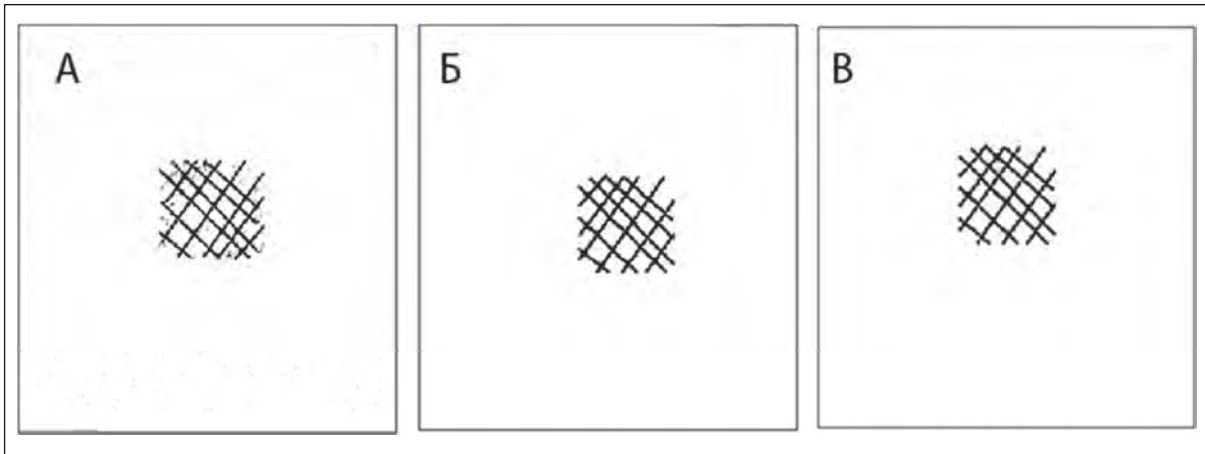


Рис. 20. Результаты восстановления тестового изображения размером 256x256 пикселей с использованием различных алгоритмов. А – НИО, Б – RAAR, В – difference map

Из сопоставления графиков видно, что алгоритм RAAR сходится существенно быстрее других: величина ошибки E_f приходит к малому значению 0,001 уже при $N_{RAAR}=150$, тогда как для других алгоритмов эта величина составляет $N_{HIO}=3110$ и $N_{diffmap}=3170$. Для алгоритмов НИО и difference map видно также, что величина ошибки E_f , вычисленная в обратном пространстве, на начальных шагах флуктуирует сильнее, чем величина E_r .

Проследим, что изменяется при увеличении размера кадра. На рисунке 20 представлены результаты восстановления изображения размером 256x256 пикселей с использованием различных алгоритмов. Видно, что в данном случае алгоритм НИО дает более зашумленный результат, чем другие алгоритмы.

Графики зависимости величины ошибок от номера итерации показывают, что при увеличении размера кадра с 128x128 до 256x256 алгоритмы сходятся существенно медленнее (рисунки 21-23). Алгоритм RAAR по-прежнему оказывается оптимальным.

В работе всех обсуждаемых алгоритмов восстановления фаз большое значение имеет информация о носителе S . Если независимыми методами (например, ПЭМ) были определены размер и форма объекта, то эта информация может быть использована для задания носителя S . Однако, с точки зрения алгоритмов, особый интерес представляет восстановление носителя S на основании дифракционных изображений. Это может быть сделано с использованием теоремы Винера-Хинчина (Хинчина-Колмогорова). Согласно этой теореме, спектральной плотностью мощности стационарного случайного процесса является Фурье-образ его автокорреляционной функции. Применим эту теорему к задаче восстановления реального объекта по его Фурье-образу.

Пусть, как и ранее, $f(\mathbf{r})$ описывает исследуемый объект, и в эксперименте зарегистрировано распределение интенсивностей $|F(\mathbf{k})|^2$. Введем автокорреляционную функцию:

$$P(\vec{R}) = \int f(\vec{r})f(\vec{r} + \vec{R})d\vec{r}.$$

Комплексное сопряжение в данном случае нет смысла ставить, т.к. мы предполагаем, что $f(\mathbf{r})$ вещественна. Согласно теореме Винера-Хинчина, выполняется соотношение

$$P(\mathbf{R}) = FT^{-1}(|F(\mathbf{k})|^2),$$

где FT^{-1} – обратное преобразование Фурье.

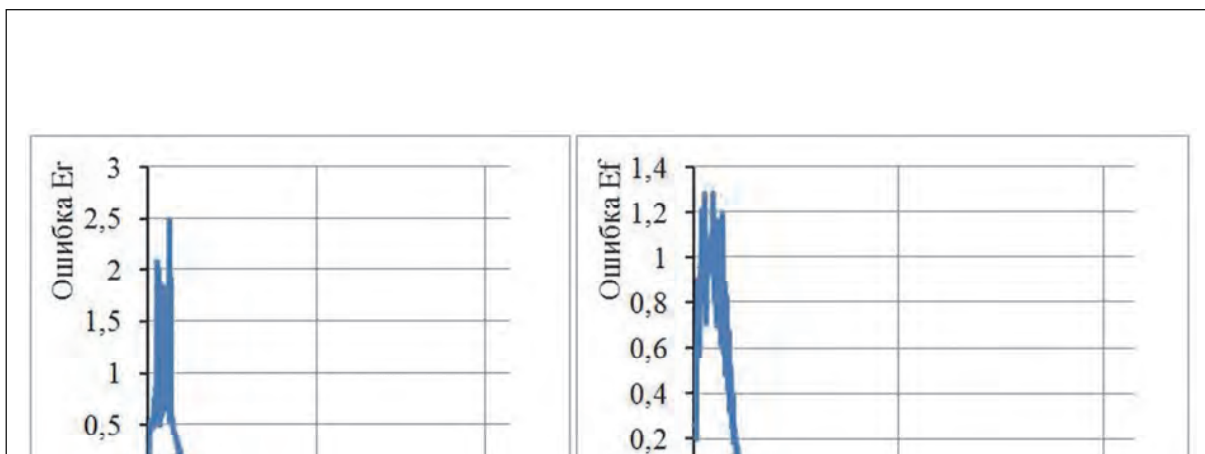


Рис. 21. Графики зависимости ошибок E_g и E_f от числа итераций, алгоритм НЮ. Размер изображения 256x256 пикселей

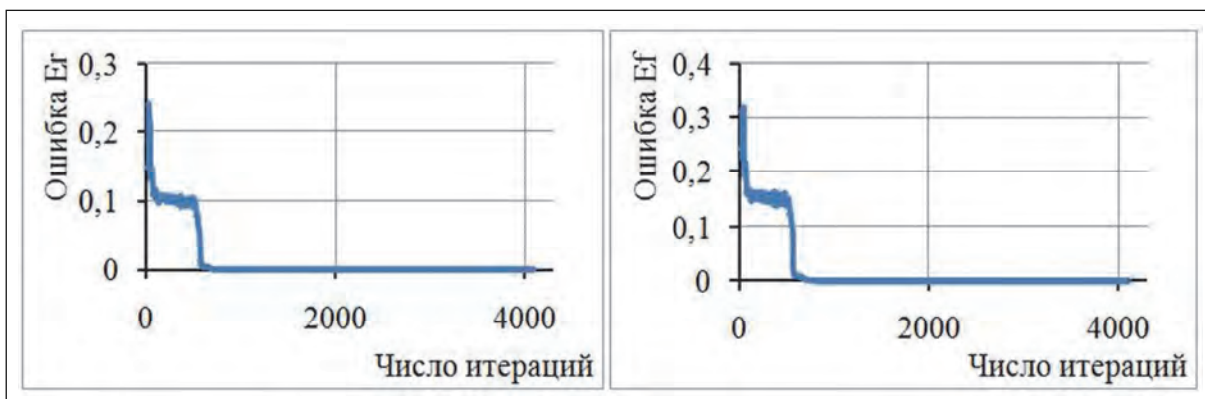


Рис. 22. Графики зависимости ошибок E_g и E_f от числа итераций, алгоритм Raar. Размер изображения 256x256 пикселей

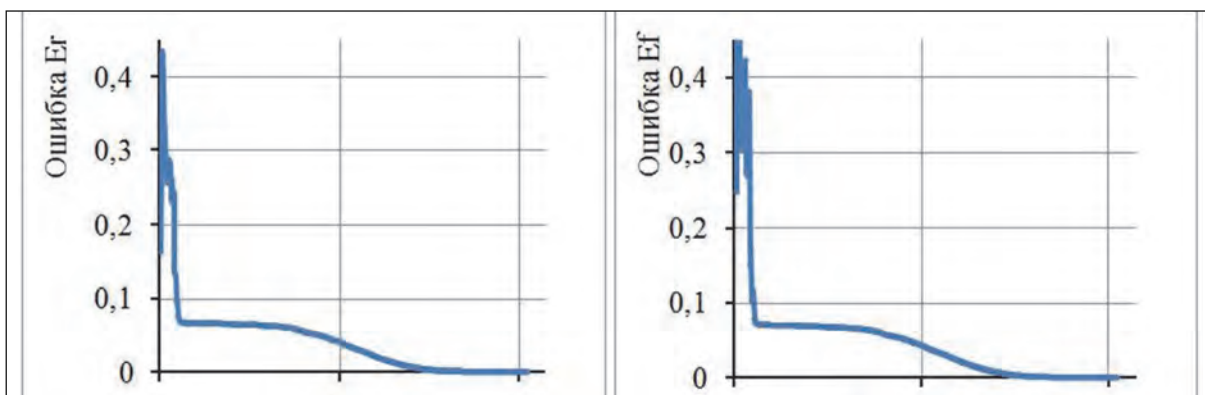


Рис. 23. Графики зависимости ошибок E_g и E_f от числа итераций, алгоритм difference map. Размер изображения 256x256 пикселей

Таким образом, экспериментально измеренное распределение $|F(\mathbf{k})|^2$ может использоваться для получения информации о форме носителя S . Действительно, носитель автокорреляционной функции $P(\mathbf{R})$ напрямую связан с носителем S функции $f(\mathbf{r})$. Эта связь проиллюстрирована на рисунке 24. Пусть функция $f(\mathbf{r})$ имеет носитель треугольной формы (рисунок 24, а). Если при вычислении автокорреляционной функции $P(\mathbf{R})$ точка $\mathbf{r}+\mathbf{R}$ оказывается за пределами области S , то она не дает вклада в интеграл $P(\mathbf{R})$. Чтобы построить носитель функции $P(\mathbf{R})$, необходимо рассмотреть все

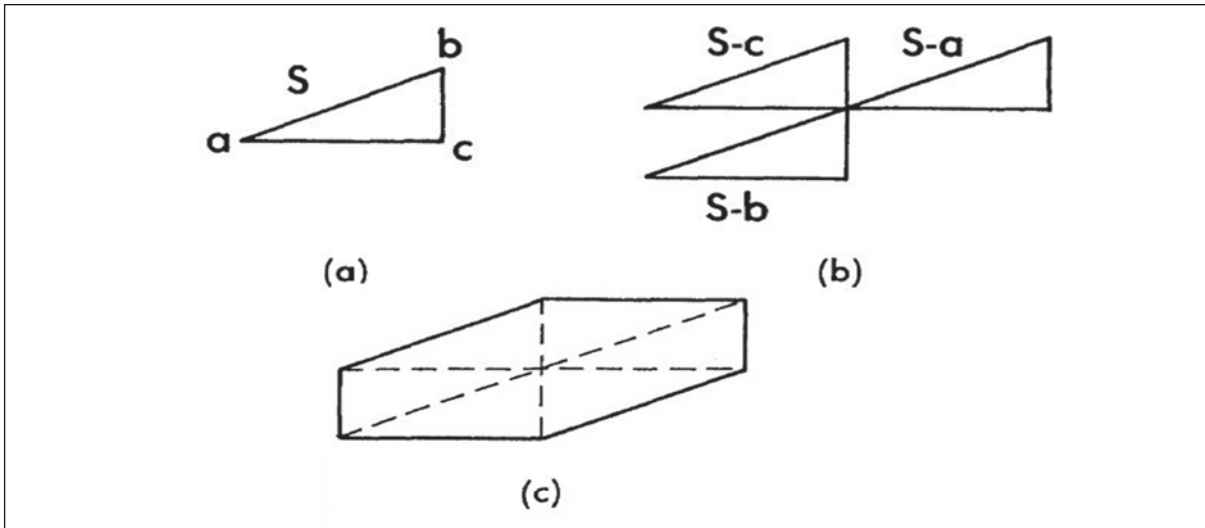


Рис. 24. Пример определения носителя автокорреляционной функции. (a) – область S , носитель функции $f(r)$. (b) – результаты трансляции области S . (c) – носитель автокорреляционной функции

возможные трансляции области S на такие вектора, у которых и начало, и конец лежат в области этой S . Три такие трансляции представлены на рисунке 24, b. Если же для рассматриваемой треугольной фигуры собрать все возможные трансляции, то мы получим фигуру, представленную на рисунке 24 c – это и есть носитель функции $P(\mathbf{R})$.

Приведенные рассуждения позволили нам построить носитель функции $P(\mathbf{R})$ по известной области S . В реальных экспериментах возникает обратная задача: требуется определить S по известному из эксперимента носителю автокорреляционной функции. Для этого можно использовать следующий сравнительно простой подход:

1. Автокорреляционную функцию вычисляют как $P(\mathbf{R}) = FT^{-1}(|F(\mathbf{k})|^2)$, исходя из имеющегося дифракционного изображения.
2. Находят максимальное значение $P_{\max}$.
3. Выбирают порог α , лежащий в диапазоне от 0 до 1.
4. Определяют область, в которой значение $P(\mathbf{R})$ составляет не менее, чем $\alpha \cdot P_{\max}$. Эту область используют как начальное приближение для оценки S .

После того, как построена начальная оценка носителя S , в процессе вычисления фаз важно улучшить эту оценку. Это делается путем постепенного сужения области от носителя $P(\mathbf{R})$ к носителю S распределения электронной плотности ρ . Анализ литературы показал, что для этого можно использовать алгоритм под названием SHRINKWRAP – под таким названием на него ссылаются, хотя в оригинальной работе такого названия не встречается. В основе алгоритма SHRINKWRAP лежит идея о том, что при работе итерационного алгоритма восстановления фаз каждые n итераций (обычно выбирают $n=20$) область S изменяется. Это схематично представлено на рисунке 25. Носитель S необходимо теперь рассматривать как S_i , где i – номер итерации.

Если внутренний цикл (восстановление фазы) может работать по алгоритму НЮ, RAAR, difference map (или какому-то еще), то во внешнем цикле, который отвечает за улучшение носителя S , можно использовать один из двух основных подходов.

Первый из них (Threshold support update) аналогичен описанной выше процедуре вычисления порога с той лишь разницей, что пороговой фильтрации подвергается не сама область S_i (носитель функции $g_i(\mathbf{r})$), а носитель *размытой* функции $g_i(\mathbf{r})$. Это означает, что выполняются следующие шаги:

1. На функции $g_i(\mathbf{r})$ выполняют размытие по Гауссу, получают функцию $g_{iG}(\mathbf{r})$.

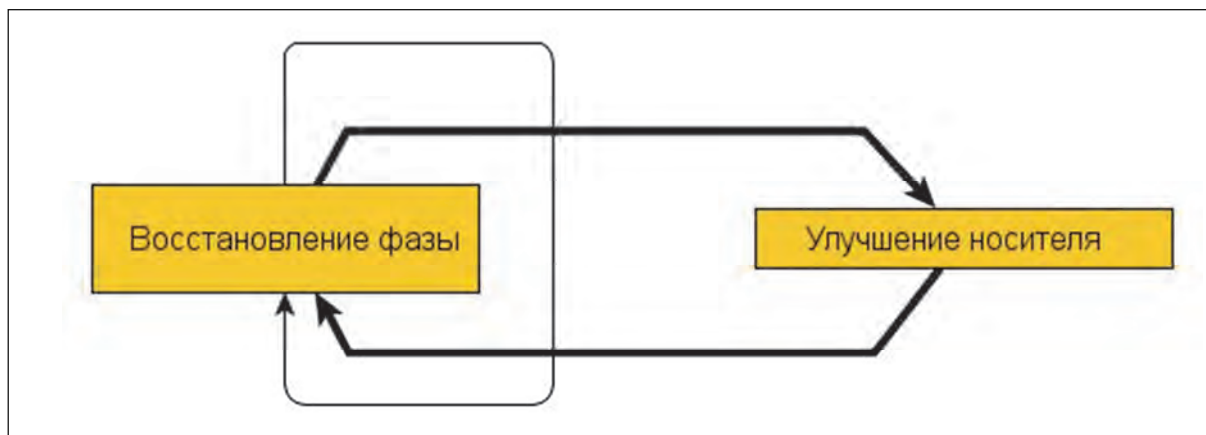


Рис. 25. Внутренний (восстановление фазы) и внешний (улучшение носителя) циклы алгоритма SHRINKWRAP

2. Определяют максимальное значение функции $g_{iG}(\mathbf{r})$ и применяют к нему порог α , т.е. определяют $c = \alpha \cdot \max(g_{iG}(\mathbf{r}))$.

3. В новую область S_{i+1} включают те и только те точки, в которых выполняется условие $g_{iG}(\mathbf{r}) > c$.

Второй подход (Area support update) к улучшению носителя аналогичен первому, но параметр α в нем используется не как порог, а как мера изменения площади (в трехмерном случае – объема) области S . Этот подход предполагает следующую последовательность шагов:

1. На функции $g_i(\mathbf{r})$ выполняют размытие по Гауссу, получают функцию $g_{iG}(\mathbf{r})$.

2. Все значения $g_{iG}(\mathbf{r})$ сортируют по возрастанию. Их количество принимают за N_{iG} .

3. Определяют значение c с номером αN_{iG} , его принимают за порог c .

4. В новую область S_{i+1} включают те и только те точки, в которых выполняется условие $g_{iG}(\mathbf{r}) > c$.

Использование такого подхода означает, например, что если $\alpha = 0,3$, то 30% точек, в которых значения функции $g_{iG}(\mathbf{r})$ наименьшие, будут отброшены.

Пример использования алгоритма SHRINKWRAP приведен на рисунке 26. В данном случае использовалось сочетание алгоритма НЮ с алгоритмом улучшения носителя по площади. Видно, что результат работы вычислительных алгоритмов отлично совпадает с изображением, полученным с помощью электронной микроскопии.

Таким образом, изложенный выше алгоритм анализа экспериментальных данных позволяет перейти от представлений в обратном пространстве к представлениям в реальном пространстве, причем как в 2D, так и 3D.

Вернемся к алгоритму, блок-схема которого представлена на рисунке 14. Обсудим два шага, которые связаны с переходом от двумерных представлений к трехмерным – в прямом или в обратном пространстве. В зависимости от того, в каком пространстве выполняется этот переход, процедура имеет ряд особенностей, которые собраны в таблице 2.

Построение трехмерных реконструкций в прямом пространстве сравнительно хорошо отработано в просвечивающей электронной микроскопии. В то же время, для решения аналогичной задачи в обратном пространстве общепризнанных алгоритмов пока вовсе не существует. Разберем идеи, которые могут лечь в их основу.

Прежде всего, это две идеи, обсуждавшиеся ранее:

- Если два дифракционных изображения были получены с близких ракурсов, то они похожи.

Таблица 2. Переход от двумерных представлений к трехмерным в прямом и обратном пространстве

Переход от двумерных представлений к трехмерным	
В прямом пространстве	В обратном пространстве
Исходный материал для обработки – проекции трехмерного объекта на случайно ориентированные плоскости	Исходный материал для обработки – сечения трехмерного объекта фрагментами сфер (сфер Эвальда)
Необходимо определить центр каждого изображения	Все изображения исходно центрированы
Изображение объекта необходимо отделить от фона	Понятия «фона» не существует, все пиксели каждого изображения учитываются при обработке
При оптимизации изображений используют операции коррекции яркости, контрастности и гаммы	Операции коррекции яркости, контрастности и гаммы не используют

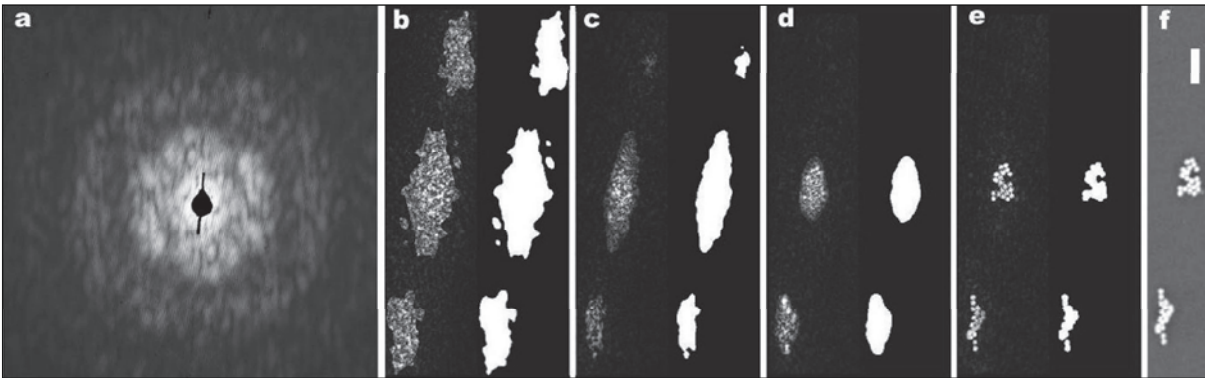


Рис. 25. Пример постепенного улучшения носителя S с помощью алгоритма SHRINKWRAP. А – дифракционное изображение, полученное от наночастиц золота, $\lambda=2$ нм, b-e – результат восстановления изображения и носителя при количестве итераций, равном 1 (b), 20 (c), 100 (d) и 1000 (e). f – изображение наночастиц, полученное сканирующим электронным микроскопом, величина масштабного отрезка 300 нм

- Два дифракционных изображения, полученные с разных ракурсов, имеют линию, вдоль которой они одинаковы с точностью до шума.

Вторая из этих идей особенно интересна с учетом того, что распределение электронной плотности вещественно и, следовательно, трехмерное распределение амплитуды (или интенсивности) дифракции обладает центральной симметрией. Центральная симметрия приводит к тому, что каждая пара сечений будет иметь не одну, а две общие линии – это проиллюстрировано на рисунке 27. Это является радикальным отличием от просвечивающей электронной микроскопии.

Ясно, что с вычислительной точки зрения распределения интенсивности $I(\mathbf{k})$ и амплитуды $F(\mathbf{k})$ практически равноправны, т.к. тривиальным образом пересчитываются друг в друга. Тем не менее, обычно говорят о построении трехмерного распределения именно интенсивности дифракции (3D intensity).

Прокомментируем математический смысл 3D intensity с точки зрения построения Эвальда. Обычно его применяют при анализе дифракционных изображений, полученных от кристаллов, и в этом случае удобно использовать понятие обратной решетки. Если $\mathbf{R}$ – вектор, проведенный из начала координат в один из узлов прямой решетки, то вектор $\mathbf{K}$ обратной решетки должен удовлетворять соотношению $e^{i\mathbf{KR}}=1$. Сфера Эвальда это сфера с радиусом $1/\lambda$ и центром в начале координат прямой решетки.

Условие максимума на дифракционном изображении, полученном от кристалла, может быть сформулировано так: максимум возникает тогда и только тогда, когда и начало, и конец вектора рассеяния $\mathbf{q}$ лежат на сфере Эвальда, как показано на рисунке 28.

Чтобы определить, в каких точках возникают дифракционные максимумы от кристалла, поступают следующим образом. Начало отсчета обратной решетки

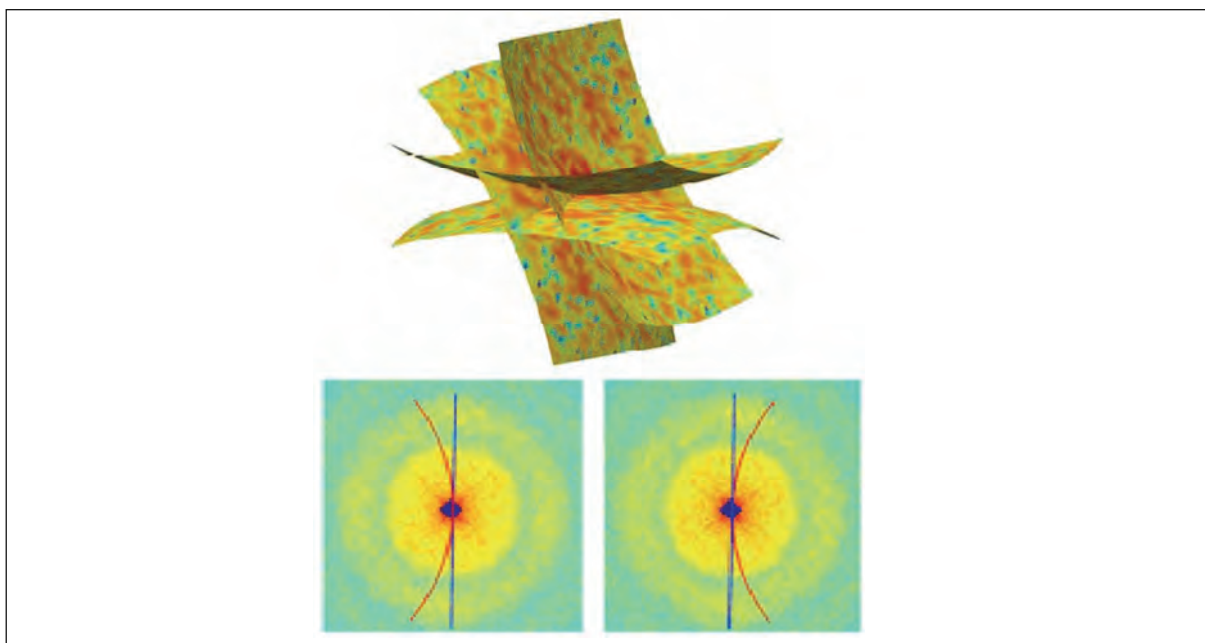


Рис. 27. Трехмерное распределение интенсивности обладает центральной симметрией (вверху). Это приводит к тому, что каждая пара сечений имеет не одну, а две общих линии. Внизу показаны соответствующие сечения, и на них выделены совпадающие линии

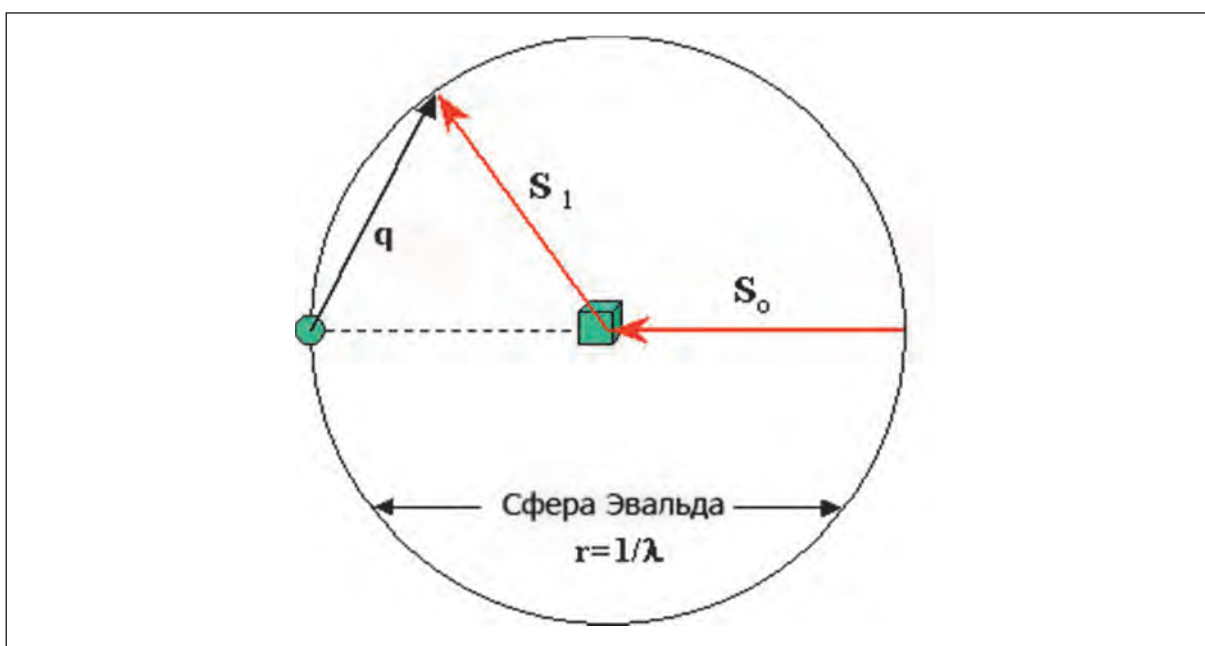


Рис. 28. Сфера Эвальда

помещают в точку, где луч, прошедший через образец, пересекает сферу Эвальда (рисунок 29). Чтобы на экране наблюдался дифракционный максимум от одного из узлов обратной решетки, должны быть выполнены три условия:

1. Этот узел должен попасть на сферу Эвальда.
2. Этот узел должен быть внутри сферы разрешения. Сфера разрешения это сфера с радиусом $1/d$, где d —необходимое разрешение. Иногда за d принимают минимальный размер элементарной ячейки, т.к. лучшего разрешения при дифракционном эксперименте с кристаллом получить невозможно.
3. Луч, прошедший через этот узел, должен попадать на экран.

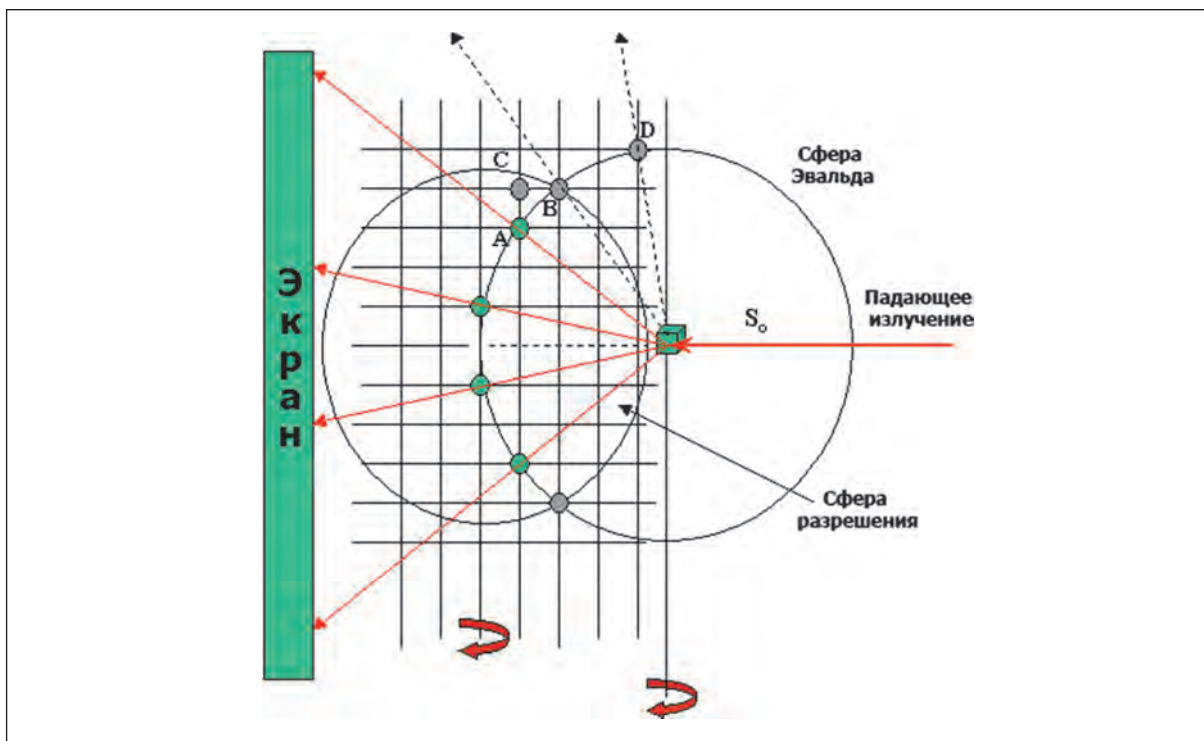


Рис. 29. Иллюстрация условий наблюдения дифракционного максимума

На рисунке 29 всем трем перечисленным условиям удовлетворяет узел А. В то же время, узел В не даст наблюдаемого рефлекса, т.к. прошедший через него луч не попадает на экран. Узел D не даст наблюдаемого рефлекса по той же причине и, кроме того, потому, что он лежит вне сферы разрешения. Узел С не приведет к возникновению наблюдаемого рефлекса, т.к. лежит вне сферы Эвальда.

Ясно, что в общем случае количество узлов, которые попадают на сферу Эвальда и удовлетворяют всем перечисленным условиям, мало. Поэтому в процессе эксперимента монокристалл поворачивают вокруг одной из осей, при этом получают множество сечений обратного пространства сферой Эвальда. Аналогичным образом в экспериментах с единичными молекулами строят множество сечений функции $I(\mathbf{k})$ сферой Эвальда.

Таким образом, алгоритм обработки экспериментальных данных о распределении электронной плотности в биомакромолекулах и их комплексах может быть реализован двумя способами. Эти способы отличаются последовательностью действий. Первый способ предусматривает выполнение обратного преобразования Фурье на каждом дифракционном изображении и последующее построение трехмерного распределения электронной плотности. Второй способ предусматривает построение трехмерного распределения интенсивности и последующее трехмерное обратное преобразование Фурье. На тестовых изображениях были сопоставлены три алгоритма для решения фазовой проблемы. Обнаружено, что алгоритм RAAR обычно оказывается быстрее, чем алгоритмы НЮ и difference map.

4. Методы анализа и интеграции экспериментальных данных и данных моделирования для исследования структуры и динамики биомакромолекул и их комплексов

Биомакромолекулярные комплексы состоят из нековалентно взаимодействующих макромолекулярных компонентов, таких как белки и нуклеиновые кислоты. Такие

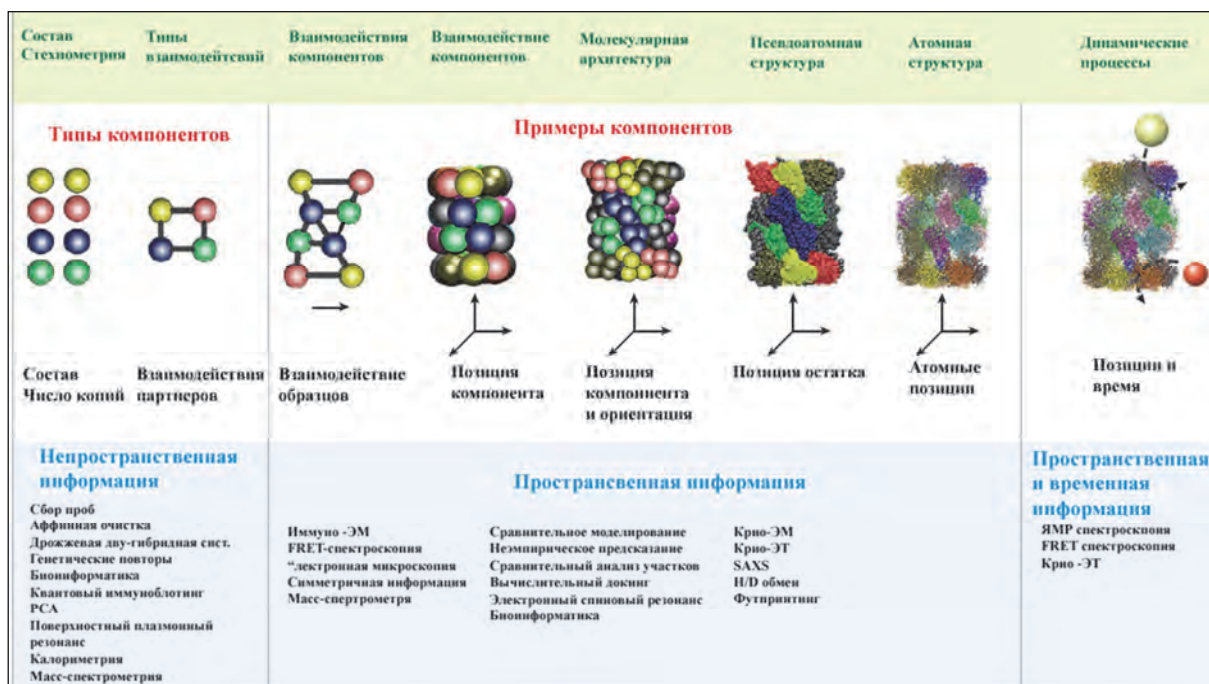


Рис. 30. Структурная информация об объектах

комплексы могут сильно различаются в размерах, но играют решающую роль в большинстве клеточных процессов. Многие комплексы состоят из 10 и даже 100 отдельных компонентов. Например, комплекс ядерной поры состоит примерно из 456 белков, которые регулируют макромолекулярный транспорт через ядерную мембрану; рибосомы состоят примерно из 80 белков и 15 молекул РНК и ответственны за белковый синтез.

Всесторонняя характеристика структур и динамики биологических механизмов необходима для понимания функционирования клетки. Даже грубое определение конфигурации высокомолекулярных компонентов в комплексе (рисунок 30) помогает выяснению принципов, которые лежат в основе клеточных процессов, а также являются необходимой отправной точкой для описания их в более высоком разрешении.

Сейчас становятся доступными полные списки макромолекулярных компонентов биологических систем. Однако, идентификация комплексов между этими компонентами является нетривиальной задачей. Эта трудность возникает частично из-за присутствия множества типов компонентов и различной продолжительности жизни комплексов. Наиболее всесторонняя информация о белковых взаимодействиях доступна для протеома *Saccharomyces cerevisiae*, содержащий примерно 6200 белков. Эти данные были созданы с помощью таких методов как дрожжевая двугибридная система и чистке на аффинной колонке в паре с масс-спектрометрией. Количество двойных белковых взаимодействий в дрожжах оценивается в 30000, соответствуя в среднем примерно 9 белковым партнерам на белок. Число комплексов более высокого порядка в дрожжах оценивается примерно в 800. Данные основаны на экспериментах по чистке на аффинной колонке. Протеом человека может иметь на порядок больше комплексов, чем в дрожжевых клетках, и число различных комплексов во всех соответствующих геномах может быть еще в несколько раз больше. Кроме того, в них могут быть тысячи биологически значимых макромолекулярных комплексов в нескольких сотнях ключевых клеточных процессах, которые стабилизируют структуры и еще характеризуют переходные взаимодействия.

По сравнению с определением структуры отдельных компонентов, структурная характеристика комплексов, как правило, сложнее и представляет собой серьезную проблему в структурной биологии. Например, рентгеновская кристаллография имеет ограничения в выращивании удобных кристаллов и построении молекулярных моделей с большой кристаллической ячейкой. Ядерно-магнитная резонансная спектроскопия лимитируется размером объектов. Электронная микроскопия, очищение на аффинной колонке, дрожжевые двух-гибридные системы, калориметрия, футпринтинг, химическое сшивание, малоугловое рентгеновское рассеивание (SAXS) и спектроскопия флуоресцентного резонансного переноса энергии (FRET) лимитируются низким разрешением соответствующей структурной информации; вычислительное моделирование белковой структуры и докинг лимитируются точностью.

Недостатки различных методов в определении структуры можно минимизировать путем одновременного рассмотрения всей доступной информации о данном комплексе. Эта информация может сильно различаться с точки зрения ее достоверности и точности и включает в себя данные из экспериментальных методов и теоретических соображений, перечисленных выше. Включение структурной информации о механизме из различных источников, может быть достигнуто только путем вычислительных средств.

Разнообразные экспериментальные методы дают разные типы структурной информации (рисунок). Эта информация различна по пространственным характеристикам, таким как разрешение, точность и количество. Стехиометрия и состав компонентов белков в комплексе может быть определен такими методами, как количественный иммуноблоттинг и масс-спектрометрия. Расположение компонентов может быть выяснено крио-ЭМ и маркирующими техниками.

Рентгеновская кристаллография самая успешная техника для анализа структур и белковых комплексов и все еще является золотым стандартом точности и разрешения. Рентгеновская кристаллография измеряет структурный фактор амплитуд и аппроксимирует фазы для кристаллического образца. Вместе с молекулярной механикой силовых полей эта информация используется в процессе оптимизации, которая может в результате дать атомную структуру макромолекулы.

ЯМР спектроскопия позволяет определять атомные структуры очень больших субъединиц и даже комплексов в растворителе с около нативными условиями. Данные от ЯМР спектроскопии включают верхнее расстояние границ между парами атомов и значения двугранных углов определенных групп атомов. В комбинации с молекулярной механикой силовых полей эта информация может в результате дать атомную структуру белка через оптимизационный процесс. ЯМР спектроскопия также часто используется для определения взаимодействия поверхностей белковых компонентов в комплексе от химического сдвига возмущений и остатков дипольной связи. Эта информация может быть скомбинирована с вычислительным докингом, чтобы получить приблизительные структуры белковых комплексов. Так же полезно, что методы ЯМР спектроскопии могут быть приложены к слабым и временным белковым комплексам, которые сложно изучать другими структурными методами. Например, могут быть визуализированы временные встречи комплексов в белок – белковых взаимодействиях. Более того, в клетке ЯМР спектроскопия предоставляет средства анализа структурных изменений, которые сопровождают белковые взаимодействия *in vivo* и в атомном разрешении.

Число структур макромолекулярных комплексов полученных кристаллографией или ЯМР спектроскопией все еще достаточно мало. В PDB находится примерно 5000 бинарных интерфейсов с менее 30% идентифицированными последовательностями в каждом. Вероятно, пройдет много времени до сбора полного репертуара высоко-разрешенных структур сотен бинарных и высокоорганизованных комплексов, нахо-

дящихся в типичной клетке. Это несоответствие объясняется, в основном, трудностью производства достаточного количества образцов и их кристаллизации.

Весьма полезным инструментом является электронная кристаллография (ЭК). Для ЭК необходимы макромолекулы, которые ограничены двумерными кристаллами (типичные для мембранных белков) или спиральные фибриллы (белки, вовлеченные в филаменты). Разрешение, получаемое в ЭК, часто достаточно, чтобы проследить белковый остов ($<4,5 \text{ \AA}$) или, по крайней мере, чтобы получить псевдоатомные модели путем помещения компонента атомной структуры в карту ($\sim 5\text{-}10 \text{ \AA}$). Однако этот метод не часто используется, из-за трудностей в получении периодических массивов и высоких технических требований.

ЭМ единичной частицы может быть использована для высушивания и напыления тяжелыми металлами (негативное окрашивание ЭМ) или для гидратирования и заморозки образца (крио-ЭМ). Хотя негативное контрастирование ЭМ может определить только размеры комплекса, крио-ЭМ также определяет все распределение электронно-оптической плотности. Для того чтобы получить изображения в ЭМ единичной частицы необходимо, либо большое количество образца, либо обладать кристаллом. ЭМ единичной частицы является мощным инструментом для исследований структур макромолекулярных комплексов, которые существуют в разных конформационных состояниях или для тех, где кристаллография бессильна. Типичный вес объектов должен быть примерно 250 кДа. Типичное получаемое разрешение лежит в пределах от 5 до 15 ангстрем.

Крио электронно-микроскопические карты плотности в особенности полезны, когда скомбинированы с атомно-разрешенной структурой компонентов. Например, установка атомных структур и моделей белков и нуклеиновых кислот в Крио-ЭМ картах дало результат в квази-атомных моделях комплексов вириальных субъединиц рибосом и рибосом взаимодействующих белков. Число single-particle реконструкций, находящиеся в ЭМ базе данных составляет около тысячи и показывает, что ЭМ единичной частицы становится стандартным методом в структурной биологии. Крио-электронная томография. Крио-ЭТ может использоваться для получения трехмерных структур плеоморфных объектов, например целых клеток. Огромный потенциал крио-ЭТ заключается в визуализации комплексов в невозмущенном клеточном контексте. Прокариотическая и тонкая эукариотическая клетки могут быть визуализированы в целом, и последние достижения в секционировании образцов позволяют проникать в толщу ткани. Текущее достижимое разрешение лежит в диапазоне 5-10 нм. Для некоторых макромолекул, высокое разрешение может быть получено путем усреднения предположительно тождественных частиц. Крио-ЭТ особенно привлекательна для изучения мембранно-связанных комплексов. Образы ретровирусных оболочек белковых комплексов были получены примерно с $30 \text{ \AA}$ разрешением *in situ*, где применяется твердотельная сборка, хотя и с низкой точностью. С более низким разрешением, были получены такие важные комплексы, как NPC. Крио-ЭТ потенциально может характеризовать временные процессы взаимодействия, получая их изображения в невозмущенной среде. Предпосылкой к анализу приближений макромолекул являются методы систематического определения стабильных состояний. Проблемы не инвазивного внедрения электронно-плотных маркеров способствует идентификации комплексов на основе известных структурных подписей. Исследования с использованием фантомных клеток (т.е., липосомы с известным содержанием) показали, что этот подход для размещения больших ансамблей ($MW > 1 \text{ MDa}$) возможен; в последнее время, был определен первый рибосомальный атлас целой клетки *Spiroplasma melliferum*.

Экспериментальные данные, о структуре описанные выше должны быть конвертированы в подробную структурную модель посредством вычислений. Атомные структуры белковых комплексов являются объектами сравнительного молекулярного моделирования. Они также используются для получения статистических потенциалов средней силы, которые могут быть полезными для сбора и оценки белковых комплексов, вычислительного докинга.

Моделирование, основанное на шаблоне опирается, на известные структуры гомологичных комплексов. Такие методы включают сравнительное моделирование и нарезание (собираание из кусочков).

Вполне возможно смоделировать белок, используя стандартную технику сравнения (рисунок 31). Требованием является наличие структурных шаблонов, которые могут быть связаны с последовательностью целевой сборки. Такие шаблоны могут охватывать всю сборку или достаточно перекрытия множества ее фрагментов. Сравнительная модель комплекса может быть оценена рядом разных энергий и набором функций, включая эмпирические статистические потенциалы, которые предназначены, для того чтобы собрать компоненты взаимодействий. Выводятся они из бинарных взаимодействий известных структур. Сравнительное моделирование предполагает, что гомологичные субъединицы составляют цель, и шаблоны комплексов образуют эквивалентное взаимодействие. Конечно, типы взаимодействия между белками одних и тех же фолдов, как правило структурно подобны [среднеквадратичное отклонение взаимодействия ($i\text{RMSD}$) $\leq 10\text{\AA}$], когда структурная идентичность составляет примерно 30%. Ниже этого порога, структуры белковых интерфейсов могут быть разными.

Структура двойного белкового комплекса может быть предсказана вычислительным белковым докингом, если доступны атомные структуры его компонентов из экспериментов или моделирования. В отличие от моделирования, основанного на шаблонах, белковый докинг также может использоваться, когда гомологичные структуры не известны и, кроме того, может предсказать новые механизмы связывания. Белковый докинг полагается на глобальные поиски большого набора возможных конфигураций комплексов, максимизированных геометрически и физико-химических комплементарностей между парами составляющих компонентов. Хотя подавляющее большинство методов докинга используется для двух компонентных белковых взаимодействий, недавно был предложен многокомпонентный докинг. Методы белкового докинга широко различаются в представленных компонентах, обчете конфигураций и протоколах оптимизации.

Большинство методов белкового докинга рассматривают компоненты как жесткие тела. Имеются и методы, в которых рассматриваются боковые цепи и гибкие остовы остатков этих компонентов. Однако, эти методы обычно вычислительно дороги, потому что поиск возможных конфигураций комплекса значительно увеличивается. Низкая точность вычислительного докинга белков обычно, получается, из-за: (a) конформационных различий в связанных и несвязанных местах в субъединицах сборки, (b) ограничений по выбору соответствующих конфигураций, или (c) сложности в различии конфигураций, подобных нативным от большого числа ненативных альтернатив. Как результат, типичный метод докинга производит ансамбль решений-кандидатов, и часто сложно выбрать какой-либо из них, который походил бы на нативный. Разнообразная экспериментальная информация о взаимодействиях между компонентами в комплексе может быть использована для повышения точности белкового докинга. Многие методы белкового докинга используют экспериментальные данные на этапе оценки для исключения некоторых решений-кандидатов, которые несовместимы с данными. Некоторые программы, включая Hex, GRAMM, и RosettaDOCK используют в комбинации с данными мутагенеза; DOT используется

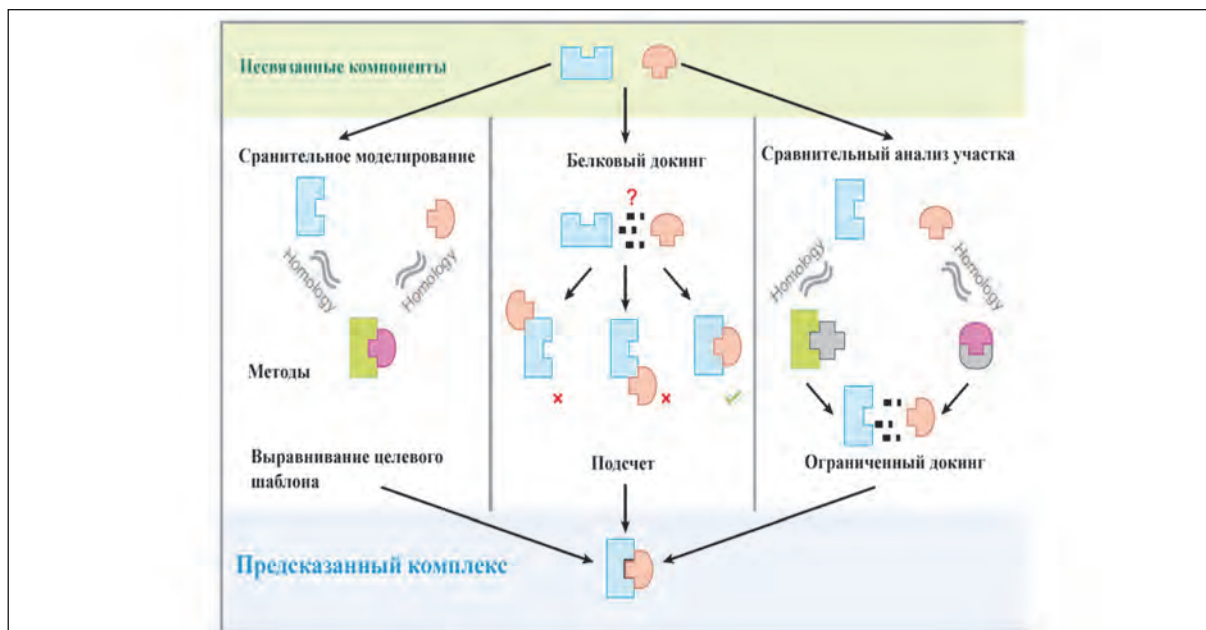


Рис. 31. Три вычислительных метода для моделирования структур белковых комплексов

вместе с водород/дейтериевым обменом данными, BIGGER, AUTODOCK, и FTDOCK, используют ЯМР возмущения химического сдвига. Другая группа методов включает в себя экспериментальные данные непосредственно в оценочной функции, которая должна быть оптимизирована для оптимизации протокола. К примеру, программа HADDOCK может использовать неоднозначные ограничения взаимодействий предполагаемые ЯМР экспериментами, полученными от возмущений химического сдвига или данными мутагенеза; взвешенный геометрический докинг использует экспериментальные данные в пользу определенных областей субъединиц поверхностей при вращении-перемещении сканирования. Другой метод, TREEDOCK, ограничивает конфигурационный поиск, используя пары якорных атомов, которые нуждаются в контакте между ними.

Местоположения мест связывания белка на белках часто сохраняется в эволюции, независимо от того какие фолды у их партнеров. Эта функция используется в сравнительном анализе участка, которая представляет из себя гибрид белкового докинга и сравнительного моделирования. Метод ограничивает вычислительный докинг для связанных сайтов, которые сохраняются в семьях гомологичный доменов. Для того, чтобы определить те сайты связывания, которые сохранились в эволюции, была применима следующая стратегия. Во-первых, для каждой субъединицы в бинарном комплексе определяется набор белок-связывающих мест и их гомологов, представленных в PIBASE базе данных структурно определенных интерфейсов. Во-вторых, эти сайты связывания отображаются на поверхности партнерской субъединицы, используя выравнивания по структуре между субъединицей и каждой из его гомологов. В-третьих, все пары отображенных мест связывания используются как стартовые точки для ограниченного (restrained) докинга для получения моделей-кандидатов бинарного комплекса. Затем, этот ансамбль моделей упорядочивается, используя меру геометрической взаимодополняемости и счета статистического потенциала. Сравнительный анализ участка имеет большую возможность для приложения, чем сравнительное моделирование и большую точность, чем белковый докинг. Сравнительное моделирование, белковый докинг и сравнительный анализ участка могут быть полезны, включая не только местную информацию о белок-белковых

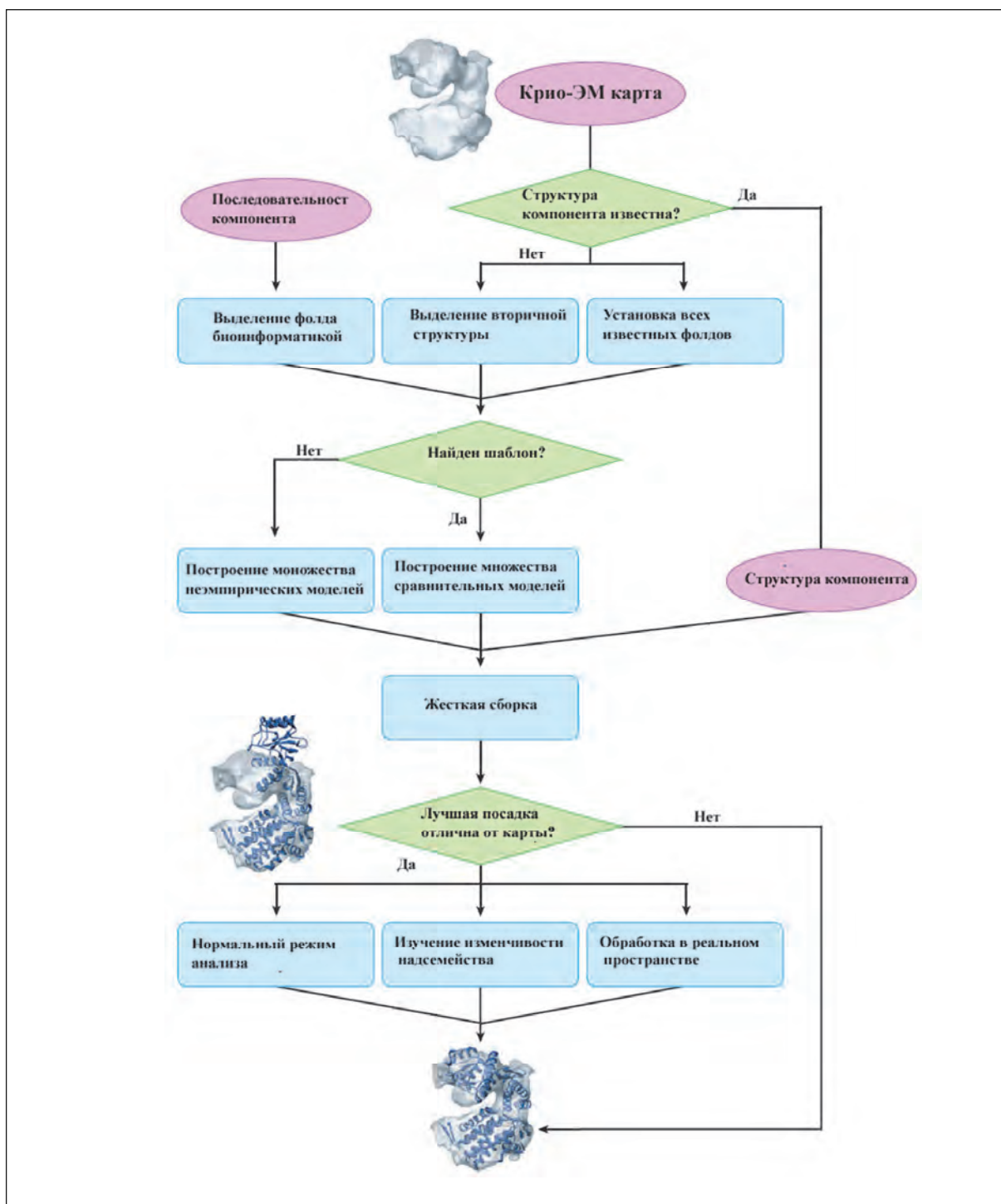


Рис. 32. Анализ карт плотности, полученных от single-particle EM

взаимодействиях, но глобальную информацию об общей форме сборки, о чем говорилось выше.

Структуры молекулярных ансамблей, органелл и даже целых клеток могут быть охарактеризованы картами плотности, получаемыми от single-particle EM, электронной кристаллографии и крио-ЭТ. Если структурные модели компонентов доступны на большем разрешении, то полезно соотносить эти модели с картой (рисунок 32). Здесь, мы обратим внимание на соотношение и деление методов, первые – для single-particle EM и электронной кристаллографией и затем крио-ЭТ.

Интерпретация ЕМ карт обычно начинается с идентификации разных структурных единиц (например, элементы вторичной структуры, домены, нуклеиновые кислоты, белки) в картах плотности, с помощью методов сегментации. Размер сегментированных единиц зависит от разрешения карт. Например, на разрешении 5-12 Å, можно увидеть сегменты вторичной структуры. Разбитие на сегменты обычно выполняются вручную, с помощью инструментов по визуализации, такие как Amira и Chimera. Кроме того, методы автоматической сегментации были разработаны как для крио-ЭМ, так и для крио-ЭТ карт. Методы для назначения структурных единиц в заданной карте плотности включают скелетирование, а также определение α -спиралей и β -листов. Если фолды компонентов известны, идентифицированные единицы могут быть полезны в обнаружении позиций компонентов на карте. В противном случае, когда фолды неизвестны, выявленные единицы могут помочь в прогнозировании фолдов компонентов.

Во многих случаях, модели с атомным разрешением компонентов часто доступны из эксперимента или предсказания. Поместив эти модели в соответствующие карты плотности, где разрешение лучше, чем примерно 20Å, можно получить псевдоатомную интерпретацию карт. Помещение компонент структур в крио-ЭМ карту было продемонстрировано с детальным псевдоатомным разрешением на примере 80S субъединицы рибосомы млекопитающего с разрешением 8,7 Å (рисунок 33).

В некоторых случаях, фиттинг компонента в карту может быть выполнена вручную с использованием интерактивных средств визуализации, такие как Chimera. Однако автоматические системы подсчета уменьшают меру субъективности, увеличивают точность и продуктивность. Фиттинг структур компонента как правило, предназначен для оптимизации сходств между компонентом и картой плотности (например, кросс-корреляционный коэффициент) как функцию от его перемещения и вращения по отношению к карте (жесткая установка). Были разработаны много методов для установки плотности в крио-ЭМ, включая SITUS, COAN, EMFIT, DOCKEM, FOLDHUNTER, URO, CHARMM, ModEM и ADP-EM. Для улучшения позиционирования компонентов, можно использовать дополнительные экспериментальные (например, маркировка золотом) и вычислительные данные, например, статистические потенциалы и геометрическую комплементарность между доменами.

Обычная проблема при фиттинге – структура изолированного компонента может быть в разном конформационном состоянии, по сравнению с неповрежденной структурой сборки. Эти конформационные различия могут происходить из разнообразных условий, под которыми изолированные компоненты объекта были определены, так же как от ошибок в экспериментальных методах (такие как появление кристаллов и шума). Обычные конформационные различия – движения сдвигов и петель доменов и элементов вторичной структуры, а так же искажения петель и движения. Кроме того, когда эксперимент недоступен, использование методов предсказания для получения моделей компонентов может включать дополнительные ошибки, такие как неправильное назначение элементов вторичной структуры неправильной последовательности участков и их изменений в пространстве, вызванной перекосом целевого шаблона в сравнительном моделировании.

Самый простой подход к пониманию конформационных вариаций структур компонента в комплексе – сгенерировать набор разных конформаций, поместить их в карту плотности и выбрать, которая имеет лучшие показатели. Этот подход отрицает высокую корреляцию между точностью модели и кросс-корреляционным счетом, между моделью и соответствующей картой плотности. Есть несколько разных приближений для генерирования таких моделей. Первое, конформации – кандидаты могут быть вычислены путем изучения структурных вариаций в лоне надсемейства

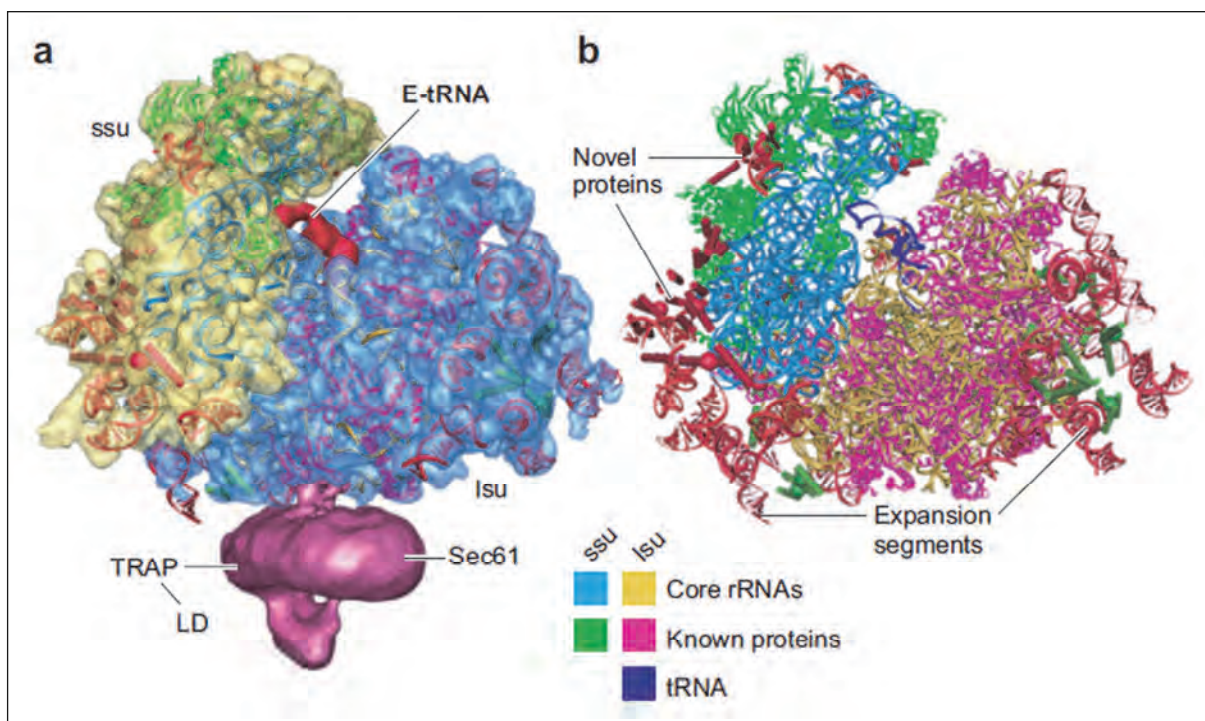


Рис. 33. Модель цитоплазматической 80S рибосомы, полученная на основе комплексного протокола сравнительного моделирования структуры белка, неэмпирического моделирования РНК, установки плотности, а также анализ карт плотности.

компонента; например, несколько альтернативных сравнительных моделей, которые основаны на различных шаблонах, могут быть помещены в карту, а затем будут выбраны наиболее совпадающие (MODELLER/Mod-EM). Второе, если фолд компонента неизвестен, модели-кандидаты могут соответствовать представлениям о всех известных фолдах (Spi-EM). Третье, разнообразные модели могут быть созданы через жесткотельные трансформации сегментов вторичной структуры руководствующие анализом главных компонент структурных соответствий белковых доменов в целевом надсемействе (S-flexfit). В итоге, для большого числа моделей может производиться неэмпирический прогноз, основанный на одной последовательности (ROSETTA/FOLDHUNTER).

Эффективность конформационного поиска возрастает при рассмотрении соответствия между компонентом и картой во время выборки. Эта цель может быть достигнута за счет оптимизации конформации компонента одновременно с его позиции и ориентации в карте крио-ЭМ при идеальном поддержании правильной стереохимии. Такие гибкие методы подгонки похожи на улучшение кристаллографических программ, кроме того, что они, как правило, улучшают твердые тела, состоящих из нескольких атомов (например, сегментов вторичной структуры и доменов) вместо индивидуальных атомов.

Были разработаны некоторые методы гибкой подгонки, используя различную выборку и схем подсчета. Например, улучшение программ в реальном пространстве RSRef и Flex-EM полагаются на методы стандартной оптимизации, включая сопряженных градиентов, молекулярной динамики с имитацией отжига, выборка Монте-Карло. Другой подход для гибкой подгонки попытаться повысить эффективность конформационного отбора проб нормальным режимом анализа работы компонента структуры.

Подобные большие конформационные изменения могут быть эффективно подсчитаны методом Moulder-EM, генетический алгоритмический протокол, который

генерирует модели сравнения через повторение выравнивание последовательностей, конструирования моделей, подгонки моделей, оценки моделей. Конформационные выборки возникают в результате итерационного изменения в расстановке, на котором основана модель. Пригодность этой функции в том, что этот генетический алгоритм совмещает подсчет кросс-корреляции между моделью и картой с атомным расстоянием – зависимость статистического потенциала для оценки модели.

В сравнении с картами плотности, полученными электронной кристаллографией и single-particle EM, карты крио-ЭТ имеют меньшее разрешение. Однако Крио-ЭТ может использоваться для реконструкции карт плотности больших клеточных объемов, даже целых клеток. Эти томограммы могут использоваться для идентификации местоположения объектов в клетке. Эта задача также известна, как сравнение шаблонов и является сложной задачей из-за низкого отношения сигнал-шум, различной контрастности всей томограммы и потери структурных факторов в связи с ограниченным диапазоном наклона («отсутствие эффекта клина»- missing-wedge effect). Детекция значительно облегчается благодаря будущим инструментальным достижениям в крио-ЭТ, которые позволят улучшить разрешение томограммы ожидаемого предела примерно до 2 нм.

Возможная схема интеграции данных

На основании вышеизложенного представляется актуальным изложение схемы интеграции различных методов (рисунок 34). Предлагаемый вариант выгодит следующим образом.

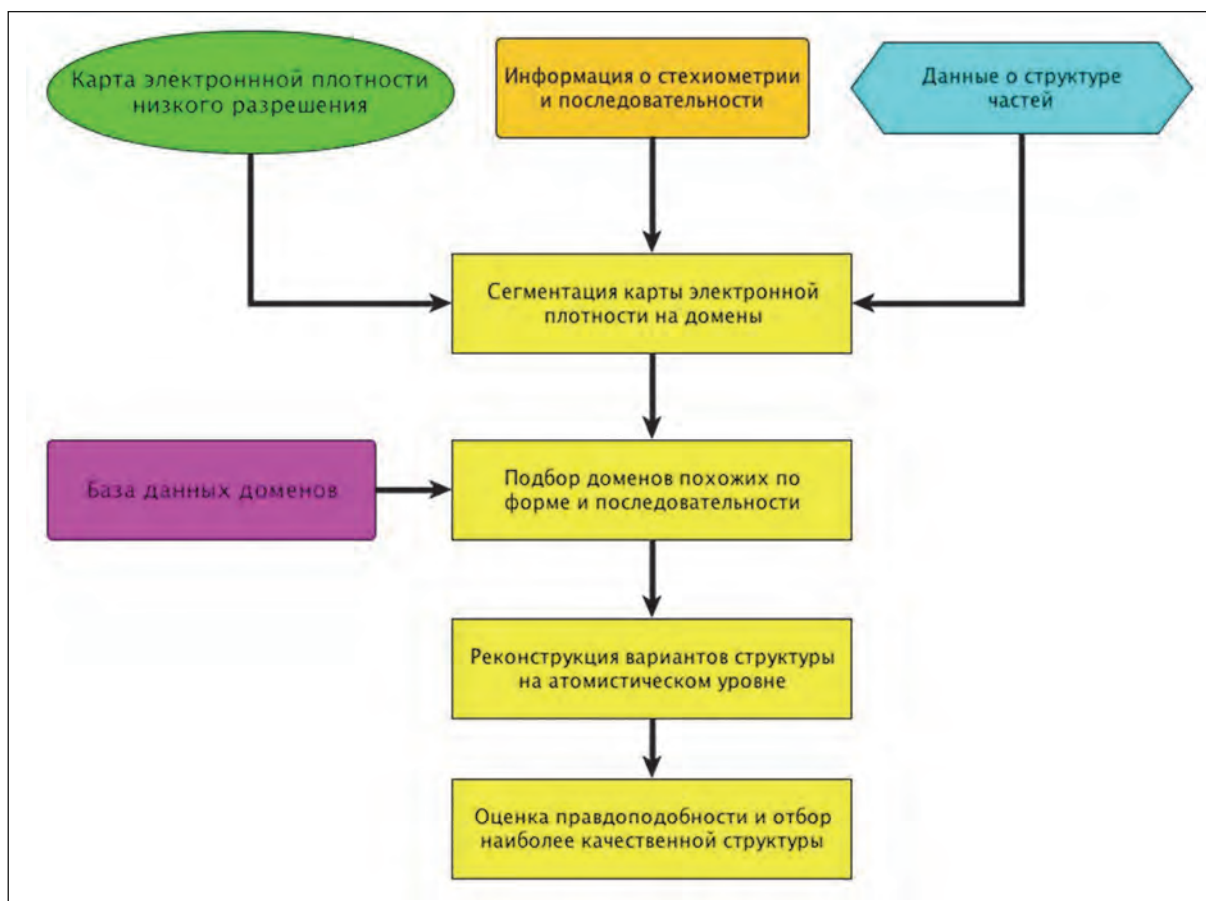


Рис. 34. Предлагаемая общая схема алгоритма интеграции данных рентгеновского рассеяния, электронной микроскопии и методов молекулярного моделирования для исследования структуры и динамики биомакромолекул и их комплексов

- На вход программному комплексу поступает карта электронной плотности низкого разрешения, полученная методами рентгеновского рассеяния или электронной микроскопии. Опционально на вход подается также информация об аминокислотной последовательности белковых компонент, а также структурная информация о входящих в комплекс частях (в виде структуры атомного разрешения или карт электронной плотности).
- На этапе предобработки проводится сегментация карты электронной плотности на возможные домены, с учетом представленной информации о последовательности, стехиометрии и структуре отдельных частей.
- Для каждого сегментированного домена из базы данных доменов подбирается похожие по форме домены, удовлетворяющие условиям сходства последовательностей.
- На заключительном этапе создается модель и вычисляется параметр ее правдоподобия. На основании этого параметра производится ранжирование моделей.

Суть схемы состоит в использовании данных о последовательности и стехиометрии молекул, входящих в комплекс, данных об известных структурах доменов и обработки этих данных методами компьютерного моделирования для интерпретации карт электронной плотности низкого разрешения, получаемых методами электронной микроскопии и рентгеновского рассеяния.

Заключение

В связи с развитием рентгеновских лазеров на свободных электронах нового поколения перед структурной биологией возникают новые амбициозные задачи. Установление пространственной атомной структуры белков и их комплексов, которые не поддаются кристаллизации, определение динамики функциональных временных переходов с временным разрешением вплоть до десятков фемтосекунд открывают новую эру в биологии. Эти методы позволяют с невиданной до настоящего времени точностью изучать физические характеристики и визуализировать процессы функционирования живой материи на наиболее фундаментальном уровне. Это открывает абсолютно новые перспективы перед фармакологией и биомедициной. Однако на этом пути предстоит решить ряд задач, связанных с обработкой сигнала, приготовлением образцов, интеграцией данных различных структурных методов для получения наиболее полной структурной и динамической картины. Значительное место здесь отводится новым математическим идеям, методам и алгоритмам.

Литература

1. *Chapman H.N. et al.* Femtosecond X-ray protein nanocrystallography // *Nature*, 2011. **470** (7332), p.73.
2. *Berman H.M. et al.* The Protein Data Bank // *Nucleic Acids Res*, 2000. **28** (1), p. 235.
3. *Raman P, Cherezov V. and Caffrey M.* The Membrane Protein Data Bank // *Cell Mol Life Sci*, 2006. **63** (1). p.36.
4. *Hunter M.S. and Fromme P.* Toward structure determination using membrane-protein nanocrystals and microcrystals // *Methods*, 2011. **55** (4), p. 387-404.
5. *Primož Rebernik R. and Margaritondo G.* Status and prospects of x-ray free-electron lasers (X-FELs): a simple presentation // *Journal of Physics D: Applied Physics*, 2012. **45** (21). p. 213001.
6. *Seibert M.M. et al.* Single mimivirus particles intercepted and imaged with an X-ray laser // *Nature*, 2011. **470** (7332). p.78-81.
7. *Johansson L.C. et al.* Lipidic phase membrane protein serial femtosecond crystallography // *Nat Methods*, 2012. **9**(3), p.263.
8. *Owen R.L., Rudino-Pinera E. and Garman E.F.* Experimental determination of the radiation dose limit for cryocooled protein crystals // *Proc Natl Acad Sci U S A*, 2006. **103** (13), p. 4912.
9. *Neutze R. et al.* Potential for biomolecular imaging with femtosecond X-ray pulses // *Nature*, 2000. **406** (6797). p. 752.
10. *Chapman H.N. et al.* Femtosecond time-delay X-ray holography // *Nature*, 2007. **448** (7154), p. 676.

11. *DePonte D.P. et al.* Gas dynamic virtual nozzle for generation of microscopic droplet streams // Journal of Physics D: Applied Physics, 2008. **41** (19). p. 195505.
12. *Johansson L.C. et al.* Membrane protein crystallization from lipidic phases // Curr Opin Struct Biol, 2009. **19** (4). p. 372.
13. *Fromme P. and Spence J.C.* Femtosecond nanocrystallography using X-ray lasers for membrane protein structure determination // Curr Opin Struct Biol, 2011. **21** (4), p.509.
14. *Kirian R.A. et al.* Femtosecond protein nanocrystallography-data analysis methods // Opt Express, 2010. **18** (6), p. 5713-5723.
15. *Luke, R.* Relaxed averaged alternating reflections for diffraction imaging // Inverse Problems, 2005. **21** (1), p. 37-50.
16. *Marchesini S. et al.* X-ray image reconstruction from a diffraction pattern alone // Physical Review B, 2003. **68** (14). p. 140101.
17. *Miao J. et al.* TAKING X-RAY DIFFRACTION TO THE LIMIT: Macromolecular Structures from Femtosecond X-Ray Pulses and Diffraction Microscopy of Cells with Synchrotron Radiation // Annual Review of Biophysics and Biomolecular Structure, 2004. **33** (1), p.157-176.
18. *Tegze M. and Bortel G.* Atomic structure of a single large biomolecule from diffraction patterns of random orientations // Journal of Structural Biology, 2012. **179** (1), p.41-45.
19. *Chapman H.N. et al.* Femtosecond diffractive imaging with a soft-X-ray free-electron laser // Nature Physics, 2006. **2** (12), p. 839-843.
20. *Ihee H. et al.* Ultrafast X-ray scattering: structural dynamics from diatomic to protein molecules // International Reviews in Physical Chemistry, 2010. **29** (3), p. 453-520.
21. *Marchesini S. et al.* X-ray image reconstruction from a diffraction pattern alone // Physical Review B, 2003. **68** (14), p.140101.
22. *Chapman H.N. et al.* Femtosecond X-ray protein nanocrystallography // Nature, 2011. **470** (7332), p.73-77.
23. *Fung R. et al.* Structure from fleeting illumination of faint spinning objects in flight // Nature Physics, 2008. **5** (1), p.64-67.
24. *Loh N.-T.D. and Elser V.* Reconstruction algorithm for single-particle diffraction imaging experiments // Physical Review E, 2009. **80** (2), p.026705 .
25. *Tokuhiya, A. et al.* Classifying and assembling two-dimensional X-ray laser diffraction patterns of a single particle to reconstruct the three-dimensional diffraction intensity function: resolution limit due to the quantum noise // Acta Crystallographica Section A Foundations of Crystallography, 2012. **68** (3), p.366-381.
26. *Ikeda S. and Kono H.* Phase retrieval from single biomolecule diffraction pattern // Opt Express, 2012. **20** (4), p. 3375-3387.
27. *Seibert M.M. et al.* Single mimivirus particles intercepted and imaged with an X-ray laser // Nature, 2011. **470** (7332), p.78-81.
28. *Huldt G., Szöke A. and Hajdu J.* Diffraction imaging of single particles and biomolecules // Journal of Structural Biology, 2003. **144** (1-2), p.219-227.
29. *Shneerson V.L., Ourmazd A. and Saldin D.K.* Crystallography without crystals. I. The common-line method for assembling a three-dimensional diffraction volume from single-particle scattering // Acta Crystallographica Section A Foundations of Crystallography, 2008. **64** (2), p.303-315.
30. *Bortel G. and Faigel G.* Classification of continuous diffraction patterns: a numerical study // Journal of Structural Biology, 2007. **158** (1), p.10-18.
31. *Fienup J.R.* Reconstruction of an object from the modulus of its Fourier transform // Optics letters, 1978. **3** (1), p.27-29.
32. *Fienup J.R.* Phase retrieval algorithms: a comparison // Applied optics, 1982. **21** (15), p. 2758-2769.
33. *Maia F.R.N.C. et al.* Hawk: the image reconstruction package for coherent X-ray diffractive imaging // Journal of Applied Crystallography, 2010. **43** (6), p.1535-1539.
34. *Sadjadi F.A. and Hall E.L.* Three-dimensional moment invariants // IEEE transactions on pattern analysis and machine intelligence, 1980. **2** (2), p.127-36.
35. *Flusser J., Suk T. and Zitov B.* Moments and moment invariants in pattern recognition // Chichester, West Sussex, U.K. ; Hoboken, N.J.: J. Wiley, 2009. xiv, 296 p.
36. *Saupe D., Vranić D.V.* 3D Model Retrieval with Spherical Harmonics and Moments // Pattern Recognition, Lecture Notes in Computer Science, 2001. **2191**, p.392-397.
37. *Hattné J. and Lamzin V.S.* Pattern-recognition-based detection of planar objects in three-dimensional electron-density maps // Acta crystallographica. Section D, Biological crystallography, 2008. **D64** (8), p.834-42.
38. *Pólya G. and Read R.C.* Combinatorial enumeration of groups, graphs, and chemical compounds // New York: Springer-Verlag, 1987, vi, 148 p.
39. *Harary F.* Graph theory // Addison-Wesley series in mathematics. 1969, Reading, Mass.,: Addison-Wesley Pub. Co. ix, 274 p.

40. *Miao J., Sayre D. and Chapman H.N.* Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects // *Journal of the Optical Society of America a-Optics Image Science and Vision*, 1998. **15** (6), p.1662-1669.
41. *Miao J. and Sayre D.* On possible extensions of X-ray crystallography through diffraction-pattern oversampling // *Acta Crystallogr A*, 2000. **56** (6), p.596-605.
42. *Luke D.R.* Relaxed averaged alternating reflections for diffraction imaging // *Inverse Problems*, 2005. **21** (1), p.37-50.
43. *Elser V.* Phase retrieval by iterated projections // *Journal of the Optical Society of America a-Optics Image Science and Vision*, 2003. **20**(1), p.40-55.
44. *Fienup JR C.T., Holsztynski W.* Reconstruction of the support of an object from the support of its autocorrelation // *J. Opt. Soc. Am.*, 1982. **72** (5), p.15.
45. *Alber F. et al.* Integrating diverse data for structure determination of macromolecular assemblies // *Annu Rev Biochem*, 2008. **77**, p.443-477.

PROTEIN MOLECULES THREE-DIMENSIONAL ATOMISTIC MODELS RECONSTRUCTION BY XFEL DATA ANALYSIS. ALGORITHMS AND METHODS

Shaitan K.V.^{1,2}, Shaitan A.K.¹, Bagrov D.V.¹, Blinov V.N.^{1,3},
Novoseletsky V.N.¹, Sokolova O.S.¹, Shurov D.L.^{1,2},
Turchenkov D.A.¹, Kirpichnikov M.P.¹

¹*Lomonosov Moscow State University, Moscow*

²*Semenov Institute of Chemical Physics, Moscow*

³*MIEM at National Research University HSE, Moscow*

shaytan49@yandex.ru

Received 15.08.2013

The methods and algorithms of XFEL data analysis for protein molecules are discussed. Experimental data on the structure and spatial distribution of the electron density in biomacromolecules and their complexes, algorithms, data analysis and integration of X-ray scattering, electron microscopy and molecular modeling techniques to study the structure and dynamics of biological macromolecules and their complexes are discussed as well.