

P10-2011-108

О. А. Афанасьев¹, П. В. Зрелов¹, В. В. Иванов¹,
Р. В. Полозов^{1,2}, В. С. Сивожелезов³, В. А. Степаненко¹,
Ю. Н. Чиргадзе⁴

КОМПЛЕКС ПРОГРАММ
ДЛЯ КАРТОГРАФИРОВАНИЯ
И ИССЛЕДОВАНИЯ БЕЛКОВ
И НУКЛЕИНОВЫХ КИСЛОТ

¹Объединенный институт ядерных исследований, Дубна

²Институт теоретической и экспериментальной биофизики РАН,
Пушино

³Институт биофизики клетки РАН, Пушино

⁴Институт белка РАН, Пушино

Афанасьев О. А. и др.

P10-2011-108

Комплекс программ для картографирования и исследования белков и нуклеиновых кислот

Разработан комплекс программ с удобным графическим интерфейсом для картографирования и исследования в среде ОС Windows поверхности глобулярных белков, спиральных цепочек белков, ДНК, РНК на основе PDB-файлов. Для белков возможно получение карт с функциональной раскраской атомов и рельефа в проекции Аитова–Хаммера (программа SURFACE-2008) или в оригинальной цилиндрической проекции (программа PROT-Z). Программа helix-DNA обеспечивает получение карт поверхности спиральных молекул ДНК/РНК в виде ортогональной проекции трехмерной структуры на плоскость, перпендикулярную длинной оси спирали ДНК. В пакет программ комплекса включена модификация программы SURFACE, разработанная ранее в среде ОС DOS для картографирования поверхности глобулярных белков в проекции Аитова–Хаммера. В состав комплекса входят компактные версии вышеуказанных программ: SURFACE-compact, PROT-Zcompact и helix-DNA-Zcompact. Так как эти программы предназначены для массового счета карт в распределенной вычислительной среде, из них удален графический интерфейс. Этим достигается существенная экономия как вычислительных ресурсов, так и времени счета при исследовании однотипных структур данных. Для получения графических изображений карт необходимо использовать полные версии соответствующих программ. Управление заданиями в среде распределенных вычислений осуществляется с помощью специальной программы-монитора NetdBBio, работающей под ОС Windows.

Работа выполнена в Лаборатории информационных технологий ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна, 2011

Afanasiev O. A. et al.

P10-2011-108

A Software Complex for Mapping and Research of Proteins and Nucleic Acids

A software complex with a suitable graphic interface has been designed for cartography and studies of the surface of globular proteins, helix chains of proteins, DNA, and RNA on the basis of PDB files within the Windows environment. For the proteins it is possible to devise maps with a functional atom and relief coloration in Aitov–Hammer projection (SURFACE-2008 program) or else in an original cylinder projection (PROT-Z program). The helix-DNA program provides maps of the surfaces of the DNA/RNA spiral molecules as an orthogonal projection of a 3D structure onto the plane perpendicular to the long axis of the DNA helix. This software complex includes a modified program SURFACE developed earlier in the OS DOS environment for mapping the surfaces of globular proteins in the Aitov–Hammer projection. The complex comprises the compact versions of the mentioned computer programs, namely SURFACE-compact, PROT-Zcompact and helix-DNA-Zcompact programs. The graphic interface has been excluded from these versions of programs as these are used only to organize a mass calculation of the maps in a distributed computing environment. Thus, one can reach significant savings of both computing resources and computing time when studying single-type data structures. In order to obtain graphic images of the maps, one should use the full versions of corresponding programs. To manage jobs within the distributed computing environment, a specialized monitoring program NetdBBio working in OS Windows has been created.

The investigation has been performed at the Laboratory of Information Technologies, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna, 2011

ВВЕДЕНИЕ

В настоящее время исследования структуры белков и нуклеиновых кислот тесно связаны с вопросами расширения возможностей современной медицины и здравоохранения. Любые исследования такого рода основаны на результатах рентгеноструктурного анализа кристаллов указанных биополимеров на атомарном уровне, и они требуют применения самой современной аппаратуры и методов [1]. Быстрорастущие объемы компьютерных банков данных [2] по структурам биологических макромолекул привели к созданию новой области науки — биоинформатики. В настоящее время структуры расшифрованных белков, а также ДНК и РНК хранятся в специальном PDB-формате [3], где для каждого атома указаны параметры: номер цепочки, код нуклеотида, код аминокислоты и пространственные координаты центра атома, измеренные с точностью 2–3 Å. В банке пространственных структур белков (Protein Data Bank) и нуклеиновых кислот (Nucleic Acid Database) в настоящее время находится около 70 000 и 10 000 структур соответственно.

В качестве примера на рис. 1 представлена атомно-шаровая модель белковой молекулы кристаллина. Из молекул этого типа формируется прозрач-

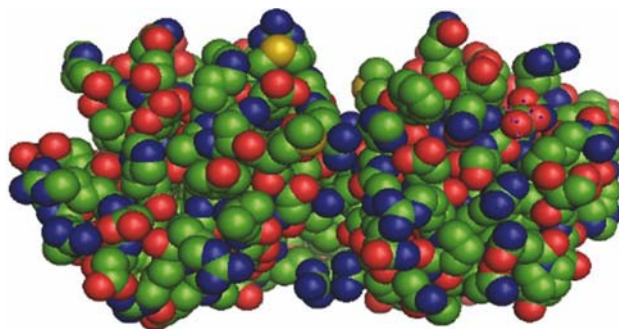


Рис. 1. Атомная модель глобулярного белка гамма-кристаллина из глазной линзы теленка. Молекула состоит из двух близких по структуре ядер, в каждом из которых около 130 аминокислотных остатков. Изображение построено с помощью программы PyMol [4]

ная плотная среда глазной линзы позвоночных. Интересной особенностью кристаллинов является их исключительно большое время существования — вплоть до 100 лет. После создания молекула продолжает существовать и функционировать, несмотря на то, что находится в органе, который непрерывно подвергается воздействию внешней среды: солнечной ультрафиолетовой радиации и различных агрессивных молекул, находящихся в воздушной среде, например, таких окислителей, как озон воздуха. Секрет в том, что часть остатков кристаллина содержит сульфгидрильные группы, которые под действием соответствующих ферментов осуществляют репарацию поврежденной разрывами молекулярной цепи.

В настоящее время имеется ряд программ, которые используются в повседневной практике при определении структуры и работе с биологическими молекулами. В данном сообщении рассмотрен комплекс программ, предназначенных для количественного исследования поверхностей белков и нуклеиновых кислот.

Следует отметить, что программы комплекса существенно отличаются от известных ранее. Они предоставляют возможность детального анализа как отдельных участков, так и всей поверхности молекулы на атомном уровне. С их помощью можно проводить следующие типы анализа поверхности белков и нуклеиновых кислот:

- анализ особенностей рельефа и формирующих его остатков или атомных групп;
- поиск кластеров, содержащих остатки или атомные группы заданного типа;
- анализ области связывания субстратов в активных центрах;
- сравнительный анализ поверхности мутантных и гомологичных форм;
- специальные анализы поверхности, например, локализацию функционально важных остатков, поиск возможных мест связывания данной молекулы с другими молекулами.

Не менее важной особенностью разработанного программного комплекса является совершенно новая возможность работать непосредственно с отобранными сериями определенных молекул в режиме потоковых расчетов и просмотров получаемых графических результатов.

1. МЕТОДЫ КАРТОГРАФИРОВАНИЯ БЕЛКОВЫХ СТРУКТУР

С точки зрения задачи картографирования PDB-файл фрагмента белковой структуры представляет упорядоченное множество N центров атомов, заданных в прямоугольной декартовой системе координат XYZ с метками, характеризующими свойство каждого атома фрагмента. Для построения карты

рельефа поверхности молекулы относительно некоторой плоскости или центра масс требуется развернуть фрагмент в некоторое начальное стандартное положение и произвести масштабирование изображения фрагмента на карте с учетом радиусов Ван-дер-Ваальса для атомов.

Так как общее количество атомов N может быть достаточно большим, для вычисления координат центра тяжести обычно используют некоторое подмножество одинаковых атомов n , характеризующих конфигурацию фрагмента с достаточной точностью. Тогда формулы расчета координат центра масс принимают следующий вид:

$$C_x = \frac{1}{n} \sum_{i=1}^n x_i; \quad C_y = \frac{1}{n} \sum_{i=1}^n y_i; \quad C_z = \frac{1}{n} \sum_{i=1}^n z_i, \quad (1)$$

где n — количество подмножества атомов; x, y, z — координаты атомов в начальной системе отсчета.

Для начальной ориентации фрагмента в пространстве требуется вычислить моменты инерции относительно каждой из осей новых декартовых прямоугольных координат, проходящих через центр масс. По определению, моментом инерции J_u тела относительно оси u называют выражение

$$J_u = \sum_{i=1}^n m_i r_i^2, \quad (2)$$

где m_i — масса точки тела, r_i — расстояние этой точки от оси u .

Момент инерции характеризует распределение массы тела относительно оси и требует довольно много вычислений. С направляющими косинусами $\cos \alpha, \cos \beta, \cos \gamma$ в новой системе координат выражение момента инерции представляется в виде

$$J_u = J_x \cos^2 \alpha + J_y \cos^2 \beta + J_z \cos^2 \gamma - 2J_{xy} \cos \alpha \cos \beta - \\ - 2J_{xz} \cos \alpha \cos \gamma - 2J_{yz} \cos \beta \cos \gamma, \quad (3)$$

где

$$\begin{aligned} J_x &= \sum m_i (y_i^2 + z_i^2), & J_{xy} &= \sum m_i x_i y_i, \\ J_y &= \sum m_i (x_i^2 + z_i^2), & J_{xz} &= \sum m_i x_i z_i, \\ J_z &= \sum m_i (x_i^2 + y_i^2), & J_{yz} &= \sum m_i y_i z_i \end{aligned}$$

— осевые и центробежные моменты инерции тела. Используя векторно-матричную форму записи для момента инерции, получим

$$J_u = \|\cos \alpha \cos \beta \cos \gamma\| \cdot \left\| \begin{array}{ccc} J_x & -J_{xy} & -J_{xz} \\ -J_{xy} & J_y & -J_{yz} \\ -J_{xz} & -J_{yz} & J_z \end{array} \right\| \cdot \left\| \begin{array}{c} \cos \alpha \\ \cos \beta \\ \cos \gamma \end{array} \right\| = \bar{e}^T J \bar{e}, \quad (4)$$

где $\bar{e}(\cos \alpha, \cos \beta, \cos \gamma)$ — орт оси u , заданный в указанной системе координат своими направляющими косинусами, матрица

$$J = \begin{vmatrix} J_x & -J_{xy} & -J_{xz} \\ -J_{xy} & J_y & -J_{yz} \\ -J_{xz} & -J_{yz} & J_z \end{vmatrix} \quad (5)$$

определяет тензор второго ранга, который еще называют тензором инерции тела.

Изменяя направление оси u и, следовательно, изменяя направляющие косинусы в выбранной системе координат, по известным компонентам тензора инерции можно найти момент инерции для любой оси, проходящей через точку O . Понятие момента инерции допускает геометрическую интерпретацию распределения моментов инерции относительно осей пучка с центром O .

Возьмем на оси u точку $M(x, y, z)$ на расстоянии, равном $K/\sqrt{J_u}$ от точки O (K — размерный коэффициент). Тогда

$$x = (K/\sqrt{J_u}) \cos \alpha, \quad y = (K/\sqrt{J_u}) \cos \beta, \quad z = (K/\sqrt{J_u}) \cos \gamma. \quad (6)$$

Подставляя полученные из этих равенств значения $\cos \alpha, \cos \beta, \cos \gamma$ в выражение (3), получим уравнение поверхности 2-го порядка

$$J_x x^2 + J_y y^2 + J_z z^2 - 2J_{xy}xy - 2J_{xz}xz - 2J_{yz}yz = K^2. \quad (7)$$

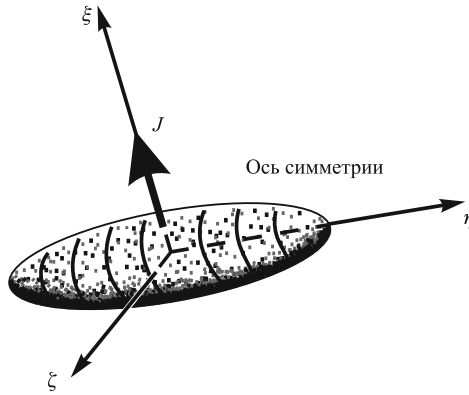


Рис. 2. Эллипсоид инерции

Эта поверхность является геометрическим местом точек M при всевозможных направлениях оси u . Поскольку момент инерции относительно произвольной оси — это величина, отличная от нуля, то поверхность (5) не имеет точек на бесконечности и, значит, является эллипсоидом. Этот эллипсоид называют эллипсоидом инерции. Каждой точке тела соответствует свой эллипсоид инерции.

Выбор осей симметрии эллипсоида инерции (рис. 2) в качестве координатных осей позволяет привести уравнение эллипсоида инерции (7) к канонической форме. Оси симметрии эллипсоида инерции, построенного в некоторой точке, называют главными осями инерции для этой точки. В осях $O\xi\eta\zeta$, главных для точки O , выражение момента инерции J_u приобретает вид

$$J_u = J_\xi \cos^2 \alpha + J_\eta \cos^2 \beta + J_\zeta \cos^2 \gamma. \quad (8)$$

Таким образом, распределение масс тела характеризуется связанными между собой понятиями тензора инерции тела и эллипсоида инерции. В главных осях тензор инерции имеет диагональный вид. Так как направлениям главных осей инерции соответствуют оси симметрии эллипсоида инерции, а значит, и стационарные значения моментов инерции, то задача нахождения направлений главных осей сводится к рассмотрению необходимого условия экстремума выражения (7) при условии

$$f(\cos \alpha, \cos \beta, \cos \gamma) = \cos^2 \alpha + \cos^2 \beta + \cos^2 \gamma - 1 = 0, \quad (9)$$

или, что то же, к рассмотрению необходимого условия экстремума функции

$$\phi(\cos \alpha, \cos \beta, \cos \gamma) = J_u(\cos \alpha, \cos \beta, \cos \gamma) - \lambda f(\cos \alpha, \cos \beta, \cos \gamma), \quad (10)$$

где λ — неопределенный множитель (метод Лагранжа). Условие стационарности функции $\phi(\cos \alpha, \cos \beta, \cos \gamma)$ дает

$$\frac{\partial \phi}{\partial X_i} = \frac{\partial J_u}{\partial X_i} - \lambda \frac{\partial f}{\partial X_i} = 0, \quad (i = \overline{1, 3}), \quad (11)$$

где $X_1 = \cos \alpha$, $X_2 = \cos \beta$, $X_3 = \cos \gamma$.

Или в развернутой форме

$$\begin{aligned} (J_x - \lambda) \cos \alpha - J_{xy} \cos \beta - J_{xz} \cos \gamma &= 0, \\ -J_{xy} \cos \alpha + (J_y - \lambda) \cos \beta - J_{yz} \cos \gamma &= 0, \\ -J_{xz} \cos \alpha - J_{yz} \cos \beta + (J_z - \lambda) \cos \gamma &= 0. \end{aligned} \quad (12)$$

Однородная система линейных уравнений относительно $\cos \alpha$, $\cos \beta$, $\cos \gamma$ имеет нетривиальное решение, если ее детерминант равен нулю:

$$\det |J - \lambda E| = 0. \quad (13)$$

Левая часть равенства (13) представляет собой полином третьей степени относительно λ . Этот полином имеет три действительных корня, так как тензор J симметрический с действительными компонентами. Кроме того, умножая каждое из равенств (12) соответственно на X_1 , X_2 , X_3 и складывая, получим

$$\sum_{i=1}^3 \frac{\partial J_u}{\partial X_i} X_i - 2\lambda = 0. \quad (14)$$

По теореме Эйлера об однородных функциях из (14) имеем

$$J_u - \lambda = 0, \quad (15)$$

т.е. значения множителя λ_i совпадают со значениями момента инерции $J_i (i = \overline{1,3})$ относительно главных осей эллипсоида инерции тела в выбранной точке.

Если корни уравнения (11) различны, то, подставляя значения λ в систему (9), найдем для каждого λ_i свой набор $\cos \alpha, \cos \beta, \cos \gamma$, который определяет соответствующее направление главной оси в выбранной системе координат XYZ . Известно, что если уравнение (11) имеет два равных корня, то два главных момента инерции равны между собой, а эллипсоид инерции является эллипсоидом вращения. В случае равенства всех корней уравнения (11) эллипсоид инерции превращается в сферу.

Если тело имеет ось материальной симметрии, то эта ось является главной центральной осью. Если тело имеет плоскость материальной симметрии, то для любой точки этой плоскости одна из главных осей совпадает с перпендикуляром, восстановленным к плоскости симметрии в данной точке.

Для вычисления направляющих применен прямой метод Якоби, который позволяет одновременно определять все собственные значения и собственные векторы. Исходя из этого можно сказать, что данный метод представляет собой численное преобразование к главным осям, потому что собственные векторы X_j являются непосредственно вектор-столбцами матрицы преобразований X , а собственные значения λ_j — это диагональные элементы диагональной матрицы D в преобразовании $X^T A X = D$.

Используются матрицы элементарных вращений Якоби:

$$T_{pq} = \begin{pmatrix} 1 & 0 & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & 1 & 0 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & \cos \varphi & \cdot & \cdot & \cdot & \sin \varphi & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & 1 & \cdot & \cdot & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 0 & \cdot & \cdot & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & -\sin \varphi & \cdot & \cdot & \cdot & \cos \varphi & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 0 & \cdot & \cdot & \cdot & 0 & \cdot & \cdot & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ -p\text{-я ячейка} \\ \\ \\ \\ -q\text{-я ячейка} \\ \\ \\ \\ \end{matrix}$$

$$\begin{matrix} | & | \\ p\text{-й} & q\text{-й} \\ \text{столбец} & \text{столбец} \end{matrix} \quad (16)$$

Идея метода состоит в вычислении матрицы преобразования X как произведения таких специальных матриц, вычисляемых методом итераций. Если к A' снова применяют элементарное вращение Якоби, то становятся нулями

два других элемента, а элементы, которые ранее обращались в нуль в A' , вновь становятся отличными от нуля.

Процедура вычислений с помощью метода Якоби состоит в следующем:

- 1) выбор элемента $a_{pq}^{(k)}$ из матрицы $A^{(k)}$,
- 2) вычисление $A^{(k+1)} = T_{pq}^T A^{(k)} T_{pq}$ ($k = 0, 1, \dots$),
- 3) конец вычислений при $\sum_{1, i \neq j}^n |a_{ij}^{(k+1)}|^2 < \varepsilon$, в противном случае — увеличение k на единицу и переход к п. 1.

В результате вычислений получаем матрицу, в которой с точностью $\varepsilon > 0$ диагональные элементы являются приближениями собственных значений, а столбцы произведения всех переменных T_{pq} являются приближениями собственных векторов.

В окончательной системе пространственных прямоугольных декартовых координат мы можем выбрать ось Z совпадающей с максимальным значением момента инерции (для горизонтального расположения фрагмента белка) или совпадающей с минимальным значением момента инерции (для вертикального расположения фрагмента белка).

Второй вариант больше подходит для исследования протяженных спиралевидных структур, и при этом оси Z и Y как бы меняются местами, а ось X соответствует среднему значению момента инерции эллипсоида. Таким образом, мы получаем правую декартову систему координат; ось X направлена к наблюдателю, ось Y — влево, а ось Z — вверх. Масштабирование осуществляется по оси Z таким образом, чтобы фрагмент белковой молекулы был виден наилучшим образом.

Дальше вокруг центра тяжести фрагмента строится сфера радиусом $R = 1$. Поверхность сферы покрывается равномерной сеткой с заданным числом узлов в зодиакальном направлении и $2K$ — в азимутальном. Узлы сетки на поверхности сферы соответствуют сферическим координатам для заданного атома (точки) следующим образом:

- φ — угол между радиусом-вектором точки и осью X в плоскости XY ;
- ψ — угол между радиусом-вектором точки и полярной осью Z .

Этот подход позволяет сравнительно просто получать карту рельефа для глобулярных белков посредством проецирования луча из центра координат вдоль радиуса-вектора в заданную точку на сфере. Если на луч попадают несколько атомов, то выбирается наиболее удаленный из них. Таким образом получают таблицу значений для каждого узла сетки, которые в дальнейшем используются для построения изолиний рельефа и границ атомов на карте.

Для глобулярных белков первоначально была использована проекция поверхности сферы на плоскость по методу Аитова–Хаммера, который часто используется в астрономии. Координаты проекции связаны с углами φ и ψ

по формулам

$$X X = 2m \cos(\psi) \sin(\varphi/2), \quad (17)$$

$$Y Y = m \sin(\psi/2), \quad (18)$$

где $m = 1/v[1 + \cos(\psi) \sin(\varphi/2)]$.

В этих формулах углы меняются в пределах

$$-\pi \leq \varphi \leq \pi, \quad (19)$$

$$-\frac{\pi}{2} \leq \psi \leq \frac{\pi}{2}. \quad (20)$$

При этом предполагается, что наблюдатель в центральной части карты в виде круга видит изображение перед собой, а справа и слева от этого круга, соответственно, проецируется изображение задней полусферы. В результате получается равновеликая эллиптическая проекция, характеризующаяся значительными нелинейными искажениями (рис. 3).

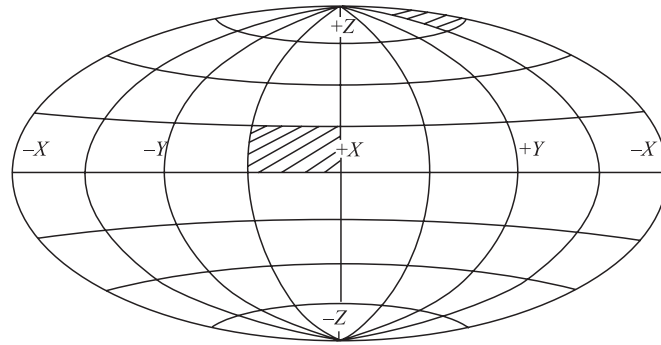


Рис. 3. Общий вид равновеликой эллиптической проекции Аитова–Хаммера (закрашены участки карты равной площади)

Искажения по краям центральной полосы карты можно частично компенсировать вращением карты вокруг оси Z , но вблизи полюсов значительные нелинейные искажения остаются в любом случае. Попытка вращения вокруг оси Y дает еще большие искажения (рис. 4).

В отличие от карты звездного неба в астрономии (где звезды представляются точечными объектами из-за огромных расстояний) атомы могут находиться вблизи центра тяжести и существенно экранировать своими размерами такие же, но более удаленные атомы белковой молекулы (рис. 5).

В силу указанных причин в такой проекции практически невозможно использовать шкалу расстояний. По меткам можно измерить расстояние в пространстве между выбранными центрами видимой пары атомов фрагмента.

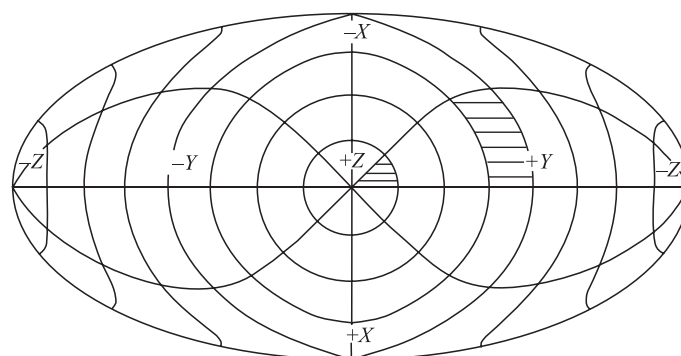


Рис. 4. Вид карты в проекции Аитова–Хаммера после вращения вокруг оси Y

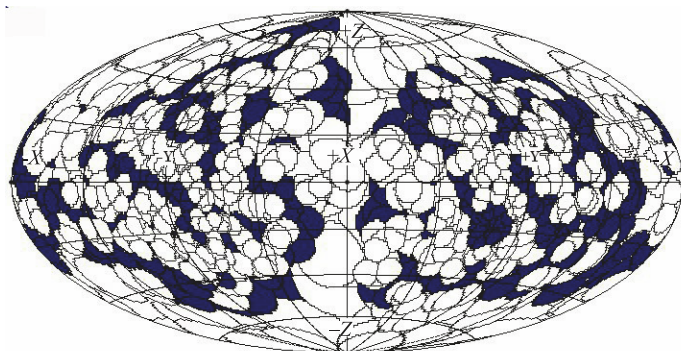


Рис. 5. Пример карты атомов глобулярного белка с функциональной раскраской атомов азота в проекции Аитова–Хаммера

Переход в сферическую систему координат производится по формулам в предположении, что радиус сферы $R = 1$:

$$\begin{aligned}
 R &= \sqrt{x^2 + y^2 + z^2}, \\
 \varphi &= \arctan(y/x), \\
 \phi &= \arctan(\sqrt{x^2 + y^2}/z).
 \end{aligned}
 \tag{21}$$

Проецирование точек поверхности атомов на сетку сферы единичного радиуса существенно упрощает вычислительные формулы при вырезании некоторого участка карты и вращениях.

Для исследования вытянутых спирализованных белковых структур проекция Аитова–Хаммера мало подходит. По этой причине авторами был разрабо-

тан метод построения срезов вдоль оси Z с проецированием на сферу ортогональных векторов в радиальной плоскости каждого среза. Посредством предложенного псевдоконформного преобразования SC удалось «развернуть» поверхность сферы единичного радиуса на прямоугольную боковую поверхность цилиндра по формулам

$$\begin{aligned}x &= x_0 + \text{round}(x_0 \sin(\varphi/4) \text{sign}(\sin(\psi))\sqrt{2}), \\y &= y_0 - \text{round}(y_0 \cos(\psi)/(1 + \text{abs}(\sin(\psi))/2)).\end{aligned}\tag{22}$$

Вводя поправку на расстояние до оси Z , в этом случае мы получаем карту атомов белка с существенно меньшими искажениями и со шкалой по оси Z в ангстремах. Соответственно, была введена поправка, зависящая от расстояния до оси цилиндра:

$$\begin{aligned}\cos \psi &= Z_i \frac{S}{R_S}, \\ \sin \phi &= \frac{y}{r},\end{aligned}\tag{23}$$

где i — номер слоя сетки, S — коэффициент масштабирования, R_S — радиус сферы вокруг фрагмента, r — радиальное расстояние атома до оси Z , $Z_i = Z$ -координата, приведенная к заданному слою.

Радиусы атомов в приближении Ван-дер-Ваальса были изменены аналогичным образом:

$$R_a = R_a \cdot r \cdot S/R_S.\tag{24}$$

1.1. Методика картографирования глобулярных белков. Глобулярные белки представляют собой сплетение цепочек атомов в виде клубка неправильной формы. Поэтому на первом этапе по координатам C-альфа атомов находится положение центра масс и вычисляются главные моменты инерции. Координаты всех неводородных атомов молекулы пересчитываются в собственную систему координат, оси которой совпадают с главными моментами инерции, как было указано ранее. После задания входного PDB-файла необходимо ввести управляющие параметры:

1) число разбиений области картографической проекции, которое определяет координатную сетку проекции в соотношении $2N \times N$ (по осям X и Y) и, соответственно, разрешение карты;

2) величину добавки R_{add} к радиусам атомов, которая необходима для расчета сглаженной молекулярной поверхности (по умолчанию $R_{\text{add}} = 0$).

На втором этапе в этой системе координат вычисляется функция молекулярной поверхности, на основе которой в дальнейшем строятся карты рельефа остатков или атомов. Функция молекулярной поверхности представляет собой функцию удаления точек поверхности от центра масс. Если

радиус-вектор пересекает поверхность в нескольких точках, то в этом случае берется максимальное удаление. Каждой такой точке приписывается адрес соответствующего атома. Удаление точки H определено в виде функции от сферических координат:

$$H = H(\varphi, \psi). \quad (25)$$

Для устранения неопределенностей, связанных с разрывом поверхности, состоящей из атомов конечного размера, при расчете вводится добавка R_{add} к радиусам всех атомов модели. Этот прием приводит к сглаживанию деталей поверхности и устраняет описанный дефект модели белка.

Функция рельефа поверхности описывается изолиниями, которые соответствуют точкам поверхности, находящимся на равном удалении от центра масс молекулы.

Данный алгоритм хорошо описывает рельеф поверхности молекулы, форма которой аппроксимируется шаром или эллипсоидом с небольшим соотношением осей, например, 1:2. В случае эллипсоида минимальные искажения рельефа могут наблюдаться только в направлениях, сильно отличающихся от направлений его полуосей. Для мультидоменных белков сложной формы рекомендуется исследовать поверхность каждого домена по отдельности. Заметные искажения поверхности на проекции наблюдаются вблизи оси Z при значениях широты от 70° до 90° (рис. 6).

Могут быть построены три основных типа карт:

- 1) карта рельефа,
- 2) карта остатков,
- 3) карта атомов.

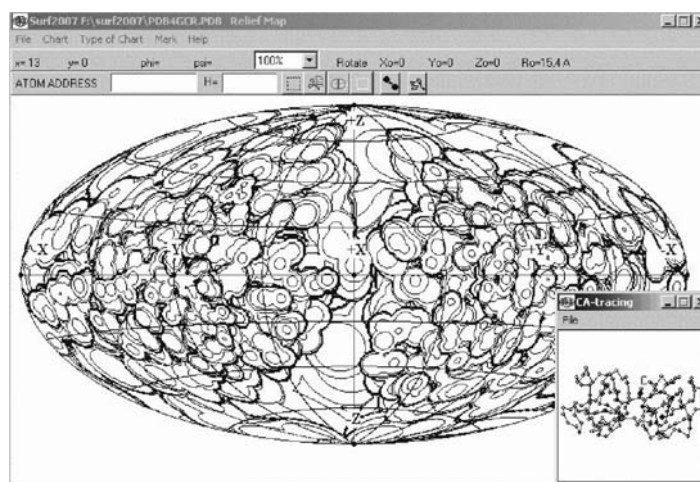


Рис. 6. Пример наложения карты схемы СА-цепочки атомов на карту рельефа поверхности того же глобулярного белка с двумя ядрами в проекции Аитова–Хаммера

На карте второго и третьего типа выводятся границы остатков или атомов, находящихся на поверхности молекулы. Соответственно, эти карты являются картами функциональной раскраски по желанию пользователя. Карты всех указанных типов можно наложить друг на друга.

Карты поверхности разных типов можно сохранить в jpg-формате в интерактивном режиме работы с программой. По окончании работы с PDB-файлом автоматически сохраняются два файла out.pdb и sa.pdb:

1) out.pdb — в системе главных осей молекулы записаны координаты всех атомов, формат PDB;

2) sa.pdb — в системе главных осей молекулы записаны координаты CA-атомов, формат PDB.

Дополнительно можно сохранить файл восстановления карты поверхности в файлах с расширениями cht и sav. Они имеют название основного файла PDB, но с другим расширением, и используются обычно для восстановления предыдущего сеанса работы с программой.

Недостатком данного подхода является зависимость размера атомов от расстояния до центра тяжести фрагмента белка.

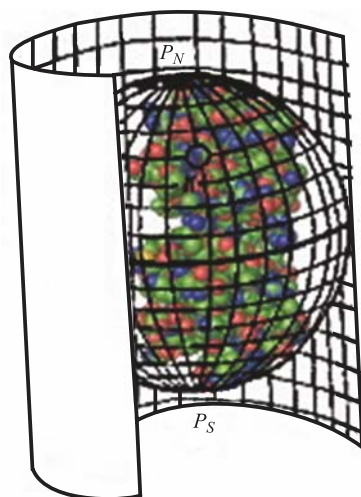


Рис. 7. Схема построения цилиндрической проекции

1.2. Методика картографирования

спирализованных белков. При исследовании структуры спирализованных фрагментов белков проекция Аитова–Хаммера малоприспособна, да и сам фрагмент удобно представить вдоль оси Z без искажений формы атомов. Для построения цилиндрической проекции был разработан алгоритм построения слоев вдоль оси Z с постоянным шагом в зависимости от длины фрагмента и параметра разбиения сетки построения карты проекции (рис. 7).

Переход в сферическую систему координат или систему координат на карте Аитова–Хаммера (с масштабom сторон 1:2) осуществляется с использованием ранее вычисленного массива коэффициентов в узлах сетки размером $N \cdot 2N$ на сфере.

В этом случае координаты точки $P(x, y)$ в проекции Аитова–Хаммера получаем по формулам (26) с учетом того, что $P(x_0, y_0)$ — центр карты на экране монитора и $0 \leq x \leq 720$; $0 \leq y \leq 360$ (в пикселях экрана):

$$x = x_0 + \text{round} \left(x_0 \cos(\pi/2 - \phi) \sin(\varphi/2) / \sqrt{(1 + \cos(\pi/2 - \phi) \cos(\varphi/2))} \right),$$

$$y = y_0 - \text{round}(y_0 \sin(\pi/2 - \phi)) / \sqrt{(1 + \cos(\pi/2 - \phi) \cos(\varphi/2))}. \quad (26)$$

Так как расстояния между атомами вычисляются в пространстве декартовых координат и не зависят от системы координат карты, мы применили псевдоконформное преобразование [5] первого типа для сферы на прямоугольник с соотношением сторон 1:2, используя формулы (22). Это следует из того, что наша область связная и непрерывная, а углы между линиями практически в любой точке сохраняют значение и направление отсчета (против часовой стрелки).

Для построения цилиндрической проекции сначала берется первая плоскость $z = 0$ и в ней находится МАХ — значение полярного вектора ρ , соответствующего поверхности максимально удаленного от оси Z атома. Таким образом, задача определения поверхности для одной точки полярного вектора с координатами $\rho, \varphi, z = 0$ решена. Точно такие же расчеты проводятся для всех векторов ρ в плоскости $z = 0$. А уже затем для всех других плоскостей с величиной $z = z_0 + \Delta z$. Данный алгоритм хорошо описывает рельеф поверхности вытянутого фрагмента белковой молекулы.

SC-преобразование сохраняет углы во внутренней области карты и имеет незначительные искажения по краям карты (рис. 8) в сравнении с известными в картографии цилиндрическими проекциями (например, проекция Меркатора) [6]. При этом на карте по шкале Z мы имеем расстояния в ангстремах, а по оси Y — угол поворота φ относительно начального положения в радиальной плоскости к оси Z .

На рис. 8,9 виден результат применения этих преобразований для рабочего поля карты и фрагмента белка.

Конформные отображения впервые применил Л. Эйлер в 1777 г. для картографирования части поверхности земного шара на плоскости (на карте) с сохранением величин всех углов. Таковыми же являются стереографическая проекция и проекция [6, 7].

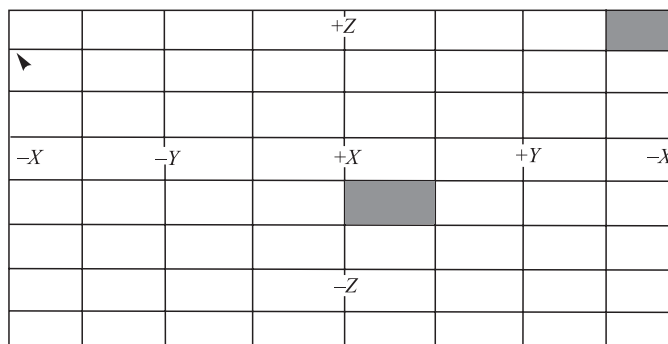


Рис. 8. Результат SC-преобразования системы координат сферы на боковую поверхность цилиндра (для сравнения закрашены равновеликие районы)

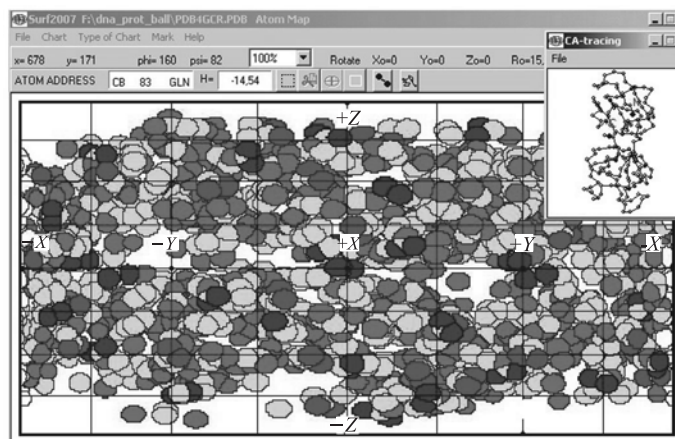


Рис. 9. Вид карты глобулярного белка с двумя ядрами в SC-проекции

Один из ведущих специалистов в области математической картографии Г. А. Мещеряков так оценил современное состояние этих исследований: «Одной из центральных проблем математической картографии является задача построения наивыгоднейших картографических проекций, т. е. проекций, в которых искажения в каком-либо смысле сведены к минимуму. Она полностью еще не решена даже для хорошо известных классов проекций, хотя частными случаями этой задачи занимались многие известные ученые (Л. Эйлер, К. Гаусс, П. Л. Чебышев и др.)» [7].

1.3. Методика картографирования ДНК. При создании программы для построения карт ДНК авторы столкнулись с самыми большими трудностями, так как требовалось не только по-другому выбирать остов, пункты меню, изменять функциональную раскраску нуклеотидов, но и производить проецирование атомов вдоль радиуса-вектора, ортогонального оси Z в обратном направлении, для получения карт внешней поверхности. Это требовалось для детального изучения распределения зарядов атомов на поверхности большой бороздки ДНК в районах промоторных участков [8], к которым обычно присоединяются белки.

Для ДНК и РНК имеется совсем небольшое различие в длине нуклеотидов, и вместо основания U (урацила) в ДНК мы имеем Т (тимин). По этой причине отдельная программа для картографирования спиральных РНК не создавалась. В качестве эталона взяли одновитковый фрагмент В-ДНК, в котором, по желанию, на уровне PDB-файла оставляли одну или две цепочки нуклеотидов.

Пришлось коренным образом переделать алгоритм построения радиусов-векторов относительно оси Z , так как попытки одновременного проецирова-

ния атомов из выбранной точки на оси Z в прямом (от оси Z) и обратном (извне) направлении для каждого радиального среза молекулы не увенчались успехом. Во-первых, алгоритм упорядочения атомов на карте в программе SURFACE был запрограммирован на работу только с положительными расстояниями до оси Z , а во-вторых, при малых расстояниях атомов относительно оси симметрии цилиндрической проекции наблюдалось удвоение контуров отдельных атомов по причине потери точности вычислений, а эту область и требовалось исследовать. Использование аппарата двойной точности было неприемлемо, так как это еще в 2–3 раза сокращало допустимое разрешение карт.

Выход был найден введением некоторого большого значения в качестве фона карты, которая строилась как ортогональная проекция изображений всех атомов молекулы на плоскость ZOY с учетом масштабирования и заданных углов поворота относительно каждой из пространственных осей координат. Такой подход потребовал расширить начальное меню программы введением опций для задания углов вращения фрагмента ДНК (рис. 10).

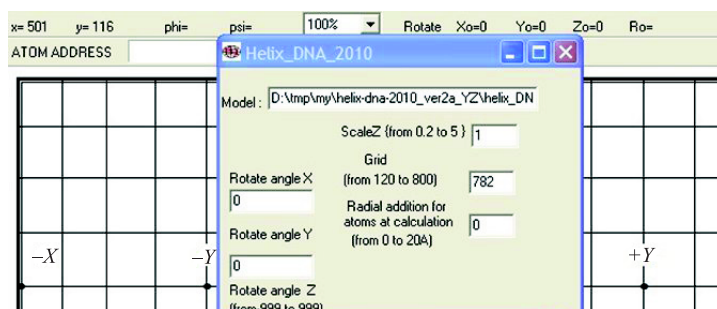


Рис. 10. Пример меню задания начальных параметров в программе helix-DNA

Насколько удачно это было реализовано, можно судить по рис. 11, где для сравнения выведены изображения независимого 3D-моделирования (справа) и остов фрагмента этой молекулы (слева).

На этом рисунке хорошо видны большая и малая бороздки ДНК. При этом практически полностью сохранились все функциональные возможности программы картографирования. С ее помощью можно проводить следующие типы анализа поверхности белков:

- 1) анализ особенностей рельефа и определение формирующих его атомных групп;
- 2) поиск кластеров, содержащих атомные группы определенного типа, например, заряженные;
- 3) анализ областей связывания спиральных молекул ДНК и РНК с лигандами и белками;

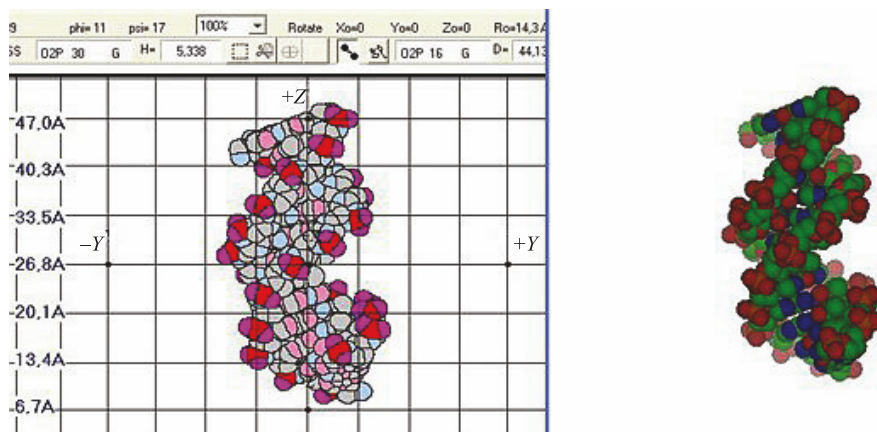


Рис. 11. Сравнение результатов картографирования фрагмента ДНК с результатами 3D-моделирования в программе РуMol [4]

4) сравнительный анализ структуры поверхности мутантных по нуклеотидному составу спиральных молекул ДНК и РНК;

5) анализ поверхности, например, локализацию функционально важных участков последовательности, поиск возможных мест связывания данной молекулы с другими молекулами и многое другое.

В результате мы получили некоторое подобие «супермикроскопа» с возможностью построения рельефа на фоне функциональной раскраски атомов с равномерной шкалой в горизонтальном и вертикальном направлениях карты (в ангстремах). Важно, что при этом есть возможность задать начальные углы поворота фрагмента макромолекулы относительно каждой из осей координат. Таким образом можно последовательно детально исследовать район промоторного участка ДНК вдоль большой бороздки.

2. ОПИСАНИЕ КОМПЛЕКСА ПРОГРАММ КАРТОГРАФИРОВАНИЯ БЕЛКОВЫХ СТРУКТУР

2.1. Программа картографирования глобулярных белков. При переводе программы SURFACE [9] в современный вариант Delphi [10] в среде ОС Windows XP [11] ее название было изменено на SURFACE-2008.

Графическая и диалоговая часть программы состоят из следующих модулей:

- Pr.dpr — главный управляющий модуль.
- Pr_U1.pas* — модуль главного окна, содержит подпрограммы чтения и записи данных с файлов, построение карт рельефа, наложение карт, а также

обработку «событий мыши». Все остальные модули вызываются из этого модуля, поэтому основные изменения были сделаны именно здесь.

- surf* — расчетная часть программы, где определяется центр масс молекулы и приводятся координаты атомов молекулы к главным осям согласно ее главным моментам инерции. Этот модуль был переписан с языка Fortran77 [12] на Delphi. Он включает в себя основные процедуры и функции, которые подверглись наибольшим изменениям:

- center* — определение центра масс молекулы по атомам Р-скелета молекулы;

- centroid* — приведение координат атомов молекулы к главным осям;

- tablcalc* — определение уровня рельефа поверхности в каждой точке карты.

- Pr_U2.pas* — выводит окно открытия PDB-файла. Из этого модуля запускается расчетный модуль surf.pas.

Из главного окна программы в режиме диалога вызываются процедуры для обмена данными между модулями и для вывода справочной информации:

- RadF.pas — окно Molecule Parameters;

- GirdF.pas — окно Map Coordinates;

- Pr_About.pas* — окно About;

- Pr_ReadMe.pas — окно ReadMe;

- Pr_CA.pas* — окно CA-tracing — расчет скелетной модели по атомам углерода и ее поворот;

- Pr_Level.pas — окно записи поверхностных атомов в формате PDB;

- Pr_Interval.pas — окно Interval изолиний рельефа;

- ChangeLines.pas — расчет уровней удаления точек поверхности от плоскости ZOY;

- Pr_Type.pas* — определение используемых типов данных, чтение и запись этих данных; параметры, ограничения общего характера для программы;

- Pr_ColMap.pas* — рисование карт раскраски остатков и атомов;

- Pr_Atom.pas — окно Color Style of Atoms;

- Pr_Residue.pas — окно Color Style of Residues;

- Pr_Style.pas — окно Color Style of Map;

- RotateF.pas — окно Rotate Map.

Символом «*» отмечены модули и процедуры программы, в которые были внесены изменения. Полностью переписан модуль surf.

Как и в первоначальной версии программы SURFACE, в данной программе установлены следующие ограничения:

- число разбиений площади карты 1000;

- число атомов около 20000;

- число остатков около 1700;

- число уровней изолиний 100;

- число адресов атомов для раскраски 100;

- число адресов остатков для раскраски 100;
- общее число надписей и меток 100.

Функциональная раскраска осуществляется с помощью указания маркера в меню типов атомов (см. приложение, табл. 5 и рис. 5):

- положительно заряженные;
- отрицательно заряженные;
- все заряженные;
- полярные;
- неполярные;
- атомы главной цепи;
- нанесение контуров всех поверхностных атомов.

Возможна также раскраска атомов по их адресу:

- если задано имя атома, то выделяются атомы данного типа для всех остатков;
- если задано имя атома и номер остатка (через пробел), то выделяется атом данного типа только для этого остатка.

При задании раскраски по адресу атома можно добавить за один раз не более одного атома. Раскраска по адресу делается в порядке задания адресов атомов. Например, заданы все СА-атомы, а затем СА 101. Тогда атом СА 101 будет выделен своим цветом. В противном случае этот атом будет окрашен так же, как и все СА-атомы. Другими словами, вначале раскрашиваются атомы по функциональному типу, затем — по адресу.

На этом этапе удалось избавиться от жесткой привязки программы картографирования к конкретной директории на заданном диске, исправить функциональную расцветку атомов и оснований, с помощью применения формата JPG [13] вместо BMP удалось в 5–7 раз сократить объем файлов карт на дисках, улучшить наглядность системы координат и остова молекулы введением специальных меток.

Следует отметить, что основными затратными ресурсами памяти являются разрешение карты и число атомов молекулы, которые в программе реализуются совместно. Это значит, что уменьшение разрешения позволяет увеличить число атомов.

2.2. Программа картографирования спирализованных белков. Наиболее существенным недостатком имеющихся программ молекулярной графики является отсутствие возможности наблюдать одновременно молекулярную поверхность со всех сторон — нужны по крайней мере шесть проекций с разных направлений. Данная программа в значительной степени решает этот вопрос для молекул белков, имеющих вытянутую форму вдоль оси Z . Она позволяет проводить работу с полной картой молекулярной поверхности белков в виде цилиндрической проекции (программа PROT отличается только отсутствием разворота фрагмента молекулы вдоль оси Z). Для обеих программ сохранен интерфейс задания параметров и функциональная раскраска групп атомов.

Функциональные возможности и интерфейс программы в данном случае практически изменились мало. По большому счету пришлось предусмотреть задание масштабирующего множителя для корректировки размеров атомов в соответствии со шкалой в ангстремах в вертикальном направлении карты. Это делается один раз, в начале работы с файлом данных для белка.

Наличие дополнительных вычислений приблизительно на 50 % замедлило работу программы PROT по сравнению с программой SURFACE. Это особенно заметно для фрагментов белка с большим числом атомов или заданием детализированной сетки картографирования.

С помощью программы можно проводить разные типы анализа поверхности вытянутых по форме белковых молекул:

- 1) анализ рельефа и определение формирующих его остатков или атомных групп;
- 2) поиск кластеров, содержащих остатки или атомные группы данного типа;
- 3) анализ области связывания субстратов в активных центрах ферментов;
- 4) сравнительный анализ поверхности мутантных и гомологичных белков;
- 5) специальные анализы поверхности белка, например, локализацию функционально важных остатков, поиск возможных мест связывания данной молекулы с другими молекулами и многое другое.

Эта версия программы явилась своего рода моделью для создания программы картографирования и исследования двойной спирали ДНК/РНК.

Важным результатом при разработке этой программы стало то, что удалось сохранить все функции картографирования программы SURFACE-2008 и впервые получить карту атомов в приближении Ван-дер-Ваальса без искажений в цилиндрической проекции со шкалой в ангстремах вдоль оси Z (рис. 12).

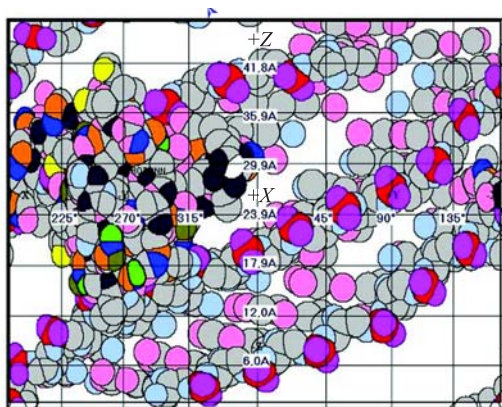


Рис. 12. Вид карты белкового комплекса в SC-проекции (со шкалой) с функциональной раскраской атомов

2.3. Программа картографирования ДНК/РНК. В программе helix_DNA, как и в других программах комплекса, визуализируются только неводородные атомы фрагмента молекулы. Так как для случая ДНК/РНК имеется другая классификация нуклеотидов (см. приложение, табл.5 и 6), то пришлось существенно изменить опции меню и функциональную раскраску атомов. Для ортогонального проектирования атомов на плоскость карты с учетом выбранного фрагмента и масштаба коренным образом была переписана процедура tabcalc. Ввиду существенного изменения функциональности объем программы увеличился и были установлены следующие ограничения:

- число разбиений площади карты 900;
- число атомов около 2000;
- число пар нуклеотидных оснований около 13;
- число уровней изолиний 100;
- число адресов атомов для раскраски 100;
- число адресов остатков для раскраски 100;
- общее число надписей и меток 100.

Важно отметить, что быстродействие программы картографирования при этом немного увеличилось из-за уменьшения вычислений по сложным формулам.

Отказавшись от режима вращения карты, для данной программы ввели режим вращения фрагмента ДНК/РНК относительно каждой из осей координат с точностью до градуса. Благодаря процедуре внешнего проецирования

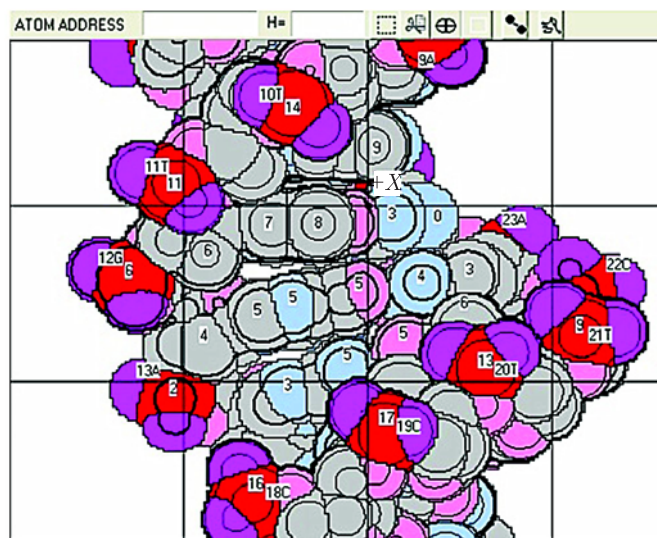


Рис. 13. Карта рельефа и функциональной раскраски молекулы фрагмента ДНК в увеличенном масштабе с метками

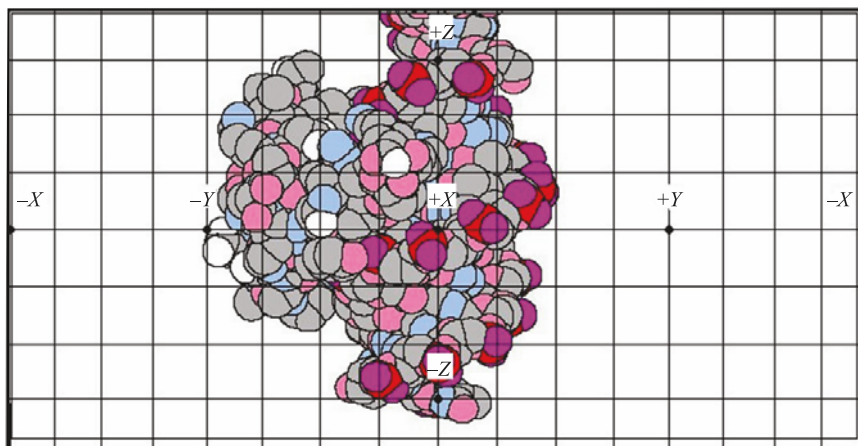


Рис. 14. Пример картографирования сложного комплекса ДНК + белок (атомы слева)

атомов на плоскость карты удалось обеспечить равномерную шкалу в ангстремах в двух направлениях и обеспечить все функциональные возможности программы в смысле масштабирования фрагмента, раскраски атомов на карте.

В качестве рекомендации для удобства работы с картой ДНК/РНК следует брать фрагменты из не более 60 нуклеотидов с выровненными цепочками. Это соответствует примерно двум полным виткам спирали В-ДНК.

2.4. Программы для организации массового счета карт на ЦИВК ОИЯИ. Решение использования сети распределенных вычислений для ускорения процесса картографирования белковых и ДНК-структур стало необходимым по мере накопления данных для исследований. Процесс запуска программ картографирования в сети распределенных вычислений подразумевает запуск программ и копирование файлов на вычислительных фермах центрального информационного вычислительного комплекса (ЦИВК) ОИЯИ [14]. Сейчас их всего три: ферма общего назначения; ферма для экспериментов на большом адронном коллайдере (ЛНС); ферма параллельных вычислений. Вычислительные фермы ЦИВК, образующие Linux-кластер [15], базируются на Unix-подобных операционных системах с распределенной файловой системой AFS (Andrew File System) [16], которая внедрена в ОИЯИ с целью создания единого файлового пространства пользователей при работе на различных фермах или рабочих станциях ОИЯИ.

Основной принцип AFS — это разбиение дискового пространства пользователя на три основных каталога: home, scratch, tmp (src). Home — домашняя директория для длительного хранения данных, это место хранения является наиболее безопасным с точки зрения защищенности от несанкционирован-

ного доступа и различных сбоев. Scratch — рабочие каталоги, предназначенные для хранения больших объемов данных. Tmp (src) — временные каталоги для временного хранения информации, например, на время счета программы; такие каталоги постоянно очищаются системой.

Для возможности реализации выполнения пакетных задач в AFS используется система пакетной обработки заданий PBS (Portable Batch System) [17]. Эта система обеспечивает управление выполнением заданий на широком наборе конфигураций вычислительных узлов: на рабочих станциях с разделением времени между задачами; на многопроцессорных системах с поддержкой как режима разделения времени на процессорах, так и эксклюзивного обслуживания каждым процессором отдельной задачи; на кластерных системах с одним или несколькими процессорами на вычислительных узлах; на произвольных комбинациях перечисленных выше систем.

Из-за такой организации сети распределенных вычислений необходимостью стала разработка ряда программ, работающих под Unix-подобными операционными системами, и отдельно создание программы-посредника между платформами Unix/Linux и Windows для осуществления операций над Linux-программами в ОС Windows, в которой работают SURFACE-2008, PROT-Z, helix-DNA-Z и программы редактирования PDB-файлов. Поэтому в эту подгруппу входят программы SURFACE-2008-compact, PROT-Zcompact и helix-DNA-Zcompact; скрипт-программа; программа-монитор NetdBBio. Программы SURFACE-2008-compact, PROT-Zcompact и helix-DNA-Zcompact — это консольные приложения, работающие под ОС Linux, аналоги графических программ SURFACE-2008, PROT-Z и helix-DNA-Z, перенявшие от них исходный код основных расчетных алгоритмов с исключением графического интерфейса. Скрипт-программа — это исполняемый текстовый файл, содержащий команды системы пакетной обработки заданий PBS для реализации и контроля процессов картографирования в сети распределенных вычислений. NetdBBio — программа с графическим интерфейсом, работающая под ОС Windows, осуществляющая мониторинг результатов счета в системе AFS. Программа находится на стадии разработки и будет содержать возможности управления счетом в среде распределенных вычислений с помощью скрипт-программы, позволит редактировать содержимое файлов с входными параметрами, производить действия над содержимым каталогов как на локальном компьютере пользователя, так и в директориях пользователя на удаленной машине системы AFS.

2.4.1. Консольные программы для вычисления карт глобулярных белков, спиральных белков и ДНК. Программы SURFACE-2008-compact, PROT-Zcompact и helix-DNA-Zcompact предназначены для вычисления карт глобулярных, спиральных белков и ДНК/РНК соответственно. Полученные карты с расширением СНТ — это текстовые файлы, содержащие информацию о вводимых параметрах закраски одной, двух и всех типов накладываемых карт:

карты атомов, нуклеотидов/остатков, карты рельефа. Эти файлы предназначены для загрузки в графические программы с целью дальнейшего визуального изучения.

Программы SURFACE-2008, PROT-Z и helix-DNA-Z написаны в объектно-ориентированной среде программирования Delphi версии 6 под ОС Windows с использованием языка OBJECT PASCAL.

Чтобы осуществить разработку консольных версий данных программ под ОС Linux [18], в которой функционирует сеть распределенных вычислений, был сделан выбор в пользу аналога данного пакета Lazarus [19]. Изначально для этой задачи мы планировали использовать Kylix — аналог Delphi под ОС Linux, разрабатываемый той же компанией Borland, но так и не завершённый проект из-за финансовых трудностей. Его разработка была заморожена в 2002 г., поэтому завершенность и свободная распространяемость бесплатного пакета Lazarus определили наш выбор в его пользу.

Lazarus — система с открытым исходным кодом, которая построена на компиляторе Free Pascal Compiler с добавлением интегрированной среды разработки (IDE), совместимой с библиотекой визуальных компонентов Delphi (VCL). Кроссплатформенность пакета позволяет установить его на большинство известных на сегодняшний день операционных систем, в частности, на ОС Scientific Linux 5 на ядре версии 2.6.

Реализация консольных версий SURFACE-2008-compact, PROT-Zcompact и helix-DNA-Zcompact подразумевает под собой процесс исключения графического интерфейса из исходного кода соответствующих программ SURFACE-2008, PROT-Z и helix-DNA-Z с помощью среды программирования Lazarus под ОС Linux. В реализованном варианте составляющими модулями каждой консольной версии являются

1) файлы-проекты: SURFACE-2008-compact.lpr, PROT-Zcompact.lpr (для белковых версий), helix-DNA-Zcompact.lpr (для версии ДНК/РНК);

2) Pr_U1.pas — модуль, содержащий основные процедуры и функции обработки входных PDB-данных;

3) Pr_LoadPar.pas — модуль обработки текстового файла, эмулирующего настройки интерфейса (размеры окна, масштаб, выбор типа карты, высоту координатной оси OZ ; раскраску атомов, остатков, рельефа; выбор цвета фона; координаты меток);

4) Pr_SaveChart.pas — модуль, сохраняющий результаты вычислений всех алгоритмов предыдущих модулей в СНТ-файл и копию PDB-файла в SAV-файл, необходимый при загрузке карты.

Скомпилированные в исполняемые файлы консольные программы нужно запускать из командной строки с четырьмя параметрами: имя программы, имя исследуемого PDB-файла (целиком или вырезанного участка атомов), имя файла эмуляции настроек интерфейса и имя карты фрагмента макромолекулы на выходе программы.

Примеры командных строк:

```
./ SURFACE-2008-compact protein1_1h8a-A.pdb loadpar_surf.txt protein1_1h8a-A
./ PROT-Zcompact prot7_helix_1trr-A_80-91.pdb loadpar_prot.txt prot7_helix_
1trr-A_80-91
./ helix-DNA-Zcompact dna6_3cro-L_A14-20_B2-8.pdb loadpar_dna.txt dna6_
3cro-L_A14-20_B2-8
```

Третий параметр — текстовый файл эмуляции настроек интерфейса — может носить любое имя, и для каждой программы он отличен по содержанию на незначительное число параметров интерфейса. Можно задать его имя, одинаковое для всех трех программ, с условием, что исполняемые программы будут находиться в разных директориях с соответствующими им входными файлами, но для сохранения целостности с учетом отличия содержимого данных параметров рекомендуется задавать осмысленные имена третьим параметрам и хранить их в той же директории, что и соответствующая программа.

Главным отличием этих параметров будет то, что в файл интерфейса для спиральных белков добавлен параметр масштабирования по Z , а для ДНК перед загрузкой входного PDB-файла к этому параметру добавляются еще три параметра поворота вокруг осей координат.

После введения всех параметров в командной строке и нажатия клавиши «ввод» результатом вычислений будут три файла с соответствующими именами:

- 1) СНТ-файл — сохраненная карта белкового комплекса или ДНК/РНК;
- 2) SAV-файл, содержащий в себе точную копию PDB-файла (необходим при визуализации карты);
- 3) новый файл INFO.txt, который содержит краткую информацию о результате счета (тип карты, время счета, дата счета, имя директории с результатами).

2.4.2. Управляющая массовым счетом скрипт-программа. Отправить выполняться задачу на ферму кластера можно вручную с помощью командной строки, и это действие сделать несложно для единичных расчетов, однако когда количество таких расчетов достигает порядка сотни — встает вопрос об упрощении данного процесса. Эти действия можно автоматизировать в скрипт-программе. По правилам организации системы AFS запуск задач на массовый счет должен осуществляться из временных scratch-каталогов, доступных на всех машинах Linux-кластера с помощью скрипт-программы, где в качестве инструмента мы использовали команды PBS:

- qdel — удаление задания;
- qhold — поставить запрет на исполнение задания;
- qmove — переместить задание;
- qmsg — послать сообщение заданию;
- qrls — убрать запрет на исполнение, установленный командой qhold;
- qselect — выборка заданий;

- `qsig` — посылка сигнала (в смысле ОС Unix) заданию;
- `qstat` — выдача состояния очередей (наиболее полезны команды `qstat -a` и `qstat -q`);
- `qsub` — постановка задания в очередь;
- `pestat` — выдача состояния всех вычислительных узлов.

Эта система осуществляет определение конкретной фермы Linux-кластера и машины в ней, управление выполнением задачи, проверку статуса задачи и отправку результатов пользователю в указанный каталог, рассылку оповещений пользователю об удачном выполнении задач или невыполнении вследствие ошибки на почтовый ящик в интернете. Все эти возможности можно использовать в скрипт-программе.

Так как входные данные с белковыми и ДНК-структурами разделены на три типа, то следует в рабочей директории создать три каталога с именами, соответствующими типу данных, и поместить в каждый из них PDB-файлы, файл эмуляции интерфейса и соответствующую консольную программу.

Разработка программы-скрипта осуществлялась в распределенной файловой системе AFS на языке shell (bash) версии 3.0. Для осуществления корректного выполнения поставленной задачи скрипт-программой мы разделили ее на два подскрипта (скрипт-модули).

Функции первого скрипт-модуля.

1. Инициализирует все переменные.
2. Осуществляет поиск входных PDB-файлов, поиск файла эмуляции интерфейса `loadpar.txt` и саму консольную программу.
3. Создает соответствующую названиям PDB-файлов структуру каталогов во временной директории:

`[temp dir]/[users]/[username]/[protein2008]` (для глобулярных белков);

`[temp dir]/[users]/[username]/[helix_protein]` (для спирализованных белков);

`[temp dir]/[users]/[username]/[helix_dnarna]` (для ДНК/РНК).

4. Копирует в эти разделы соответственно по одному входному файлу белкового образца, `loadpar.txt`, исполняемую программу и второй скрипт-модуль.

5. Реализует запуск второго скрипт-модуля с передачей ему параметров по каждому из PDB-файлов из каждой временной директории с помощью команды `qsub` системы управления заданиями PBS.

6. Создает иерархию каталогов для результатов в домашнем каталоге пользователя HOME.

Функции второго скрипт-модуля.

1. Реализует копирование на ферму кластера исполняемой программы, файла белковых данных и модуля эмуляции интерфейса из временной директории.

2. Запускает на счет консольную программу, вычисляет время счета и сохраняет, добавляя в `INFO.txt`, информацию о дате создания и времени вычислений входного PDB-фрагмента.

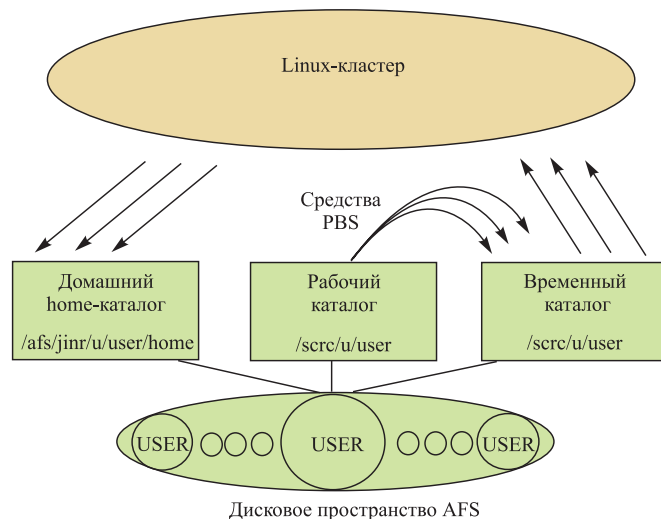


Рис. 15. Схема работы управляющей скрипт-программы на кластере ЦИВК ОИЯИ

3. Копирует результаты в указанный каталог домашней директории пользователя на удаленном сервере.

На рис. 15 представлена схема всего вычислительного процесса, реализуемого с помощью обоих скрипт-модулей.

Соблюдая разделение по типу данных в рабочей директории пользователя, мы скопировали в каталоги `protein2008` (для глобулярных белков), `protein_z` (для спирализованных белков), `dnarna_z` (для ДНК/РНК) модули скрипт-программы. Скрипт-модулям дали названия в соответствии с их функциональностью и типом данных:

- `surf_prot2008_orgdir.sh` и `surf_prot2008_submitrun.sh` — для глобулярных белков;
- `surf_helixPROTZ_orgdir.sh` и `surf_helixPROTZ_submitrun.sh` — для спирализованных белков;
- `surf_helixDNAZ_orgdir.sh` и `surf_helixDNAZ_submitrun.sh` — для ДНК/РНК.

Согласно основному принципу разбиения дискового пространства AFS скрипт-программа сохраняет результаты в домашнюю директорию, создавая в ней отдельный каталог с заданной структурой:

- [каталог_результаты];
- [каталог_тип_данных (глобулярные белки, спирализованные белки, ДНК/РНК)];
- [каталог_дата_создания];
- [каталог_название_pdb-файла (содержит непосредственно результаты)];
- СHT-файл, SAV-файл, INFO.txt.

Структура полученных данных (карт) задается с помощью управляющей скрипт-программы и может быть изменена при необходимости. Для графического представления посчитанных карт и их дальнейшего анализа используются соответствующие полные версии программ SURFACE-2008, PROT-Z и helix-DNA-Z.

2.4.3. Программа-монитор результатов счета в сети распределенных вычислений. Разрабатываемая нами программа NetdBio предназначена для быстрого доступа к результатам массового счета и удобства их хранения. Ее функциональность позволяет установить SSH-зашифрованное соединение с удаленным сервером, где хранятся данные. Мы решили осуществить разработку приложения с помощью возможностей свободно доступного языка Python 2.5 [20], который известен весьма мощными средствами для работы с сетевыми протоколами, а также библиотеки виджетов PyQt4 [21], специализированной библиотеки Qt4 для данного языка. Чтобы осуществить SSH-соединение, было решено использовать бесплатный пакет Paramiko — универсальный модуль для работы с SSH-каналом [22].

Программа NetdBio содержит три функциональные вкладки: Connects, Surface dB, Local/Remote. На вкладке Connects расположены поля ввода для аутентификации данных:

- Name (имя соединения);
- Server address (адрес сервера);
- Server Port (порт сервера);
- Remote work path (рабочая директория на удаленной машине);
- Remote home path (домашняя директория на удаленной машине);
- Local path (директория на локальной машине);
- Username (имя пользователя);
- Password (пароль).

Эти данные можно сохранять на жесткий диск персонального компьютера нажатием кнопки Save. Каждая такая операция сохранения создает отдельный файл с именем соединения, в котором хранятся данные об аутентификации пользователя, кроме пароля. Количество соединений может быть достаточно велико, поэтому для удобства их представления на вкладке Connects расположен список, содержащий названия сохраненных соединений. Кнопка Delete, расположенная рядом с кнопкой Save, позволяет удалять сохраненные соединения. После заполнения всех полей можно осуществить соединение по кнопке Connect, нажатие на которую вызывает события проверки на корректность ввода аутентификационных данных, создает экземпляр класса SSHClient из Paramiko, при помощи которого и происходит непосредственное соединение с сервером. По кнопке Disconnect реализуется разрыв соединения.

Вкладка Surface dB содержит более подробную информацию о результатах массового счета, находящихся в домашнем каталоге пользователя на удаленном сервере. Этот модуль находится на стадии разработки.



Рис. 16. NetdBBio: вкладка Local/Remote

Вкладка Local/Remote визуально разделена на три основных окна, каждое из которых представляет собой список каталогов и файлов просматриваемой директории. Первый список отображает данные из локальной директории пользователя, два других — из рабочего и домашнего каталогов на удаленной машине соответственно. Эти нововведения нужны для более удобного доступа к основным директориям и осуществления операций копирования с помощью функциональных кнопок по принципу Total Commander [23], использующего FTP-протокол (рис. 16).

- ->work (копирование данных в рабочий каталог на сервере);
- ->home (копирование данных в домашний каталог на сервере);
- <-work (копирование данных в локальную директорию из рабочего каталога на сервере);
- <-home (копирование данных в локальную директорию из домашнего каталога на сервере);
- delete (удаление данных).

Операции редактирования/создания новых файлов и каталогов находятся на стадии отладки. Планируется добавление функции отправки на массовый счет программ с возможностью выбора любого количества PDB-файлов и изменения файла эмуляции настроек интерфейса. Данная программа позволит максимально упростить диалог персональной машины с удаленным сервером.

2.5. Программы редактирования PDB-файлов данных. Входные данные берутся из файла атомных координат в формате банка белковых данных PDB (по позициям):

- (1–6) — АТОМ — идентификатор строки (с первой позиции);
- (7–11) — порядковый номер атома;

- (13–14) — два пробела;
- (14–17) — идентификатор атома (с первой позиции);
- (18–21) — идентификатор остатка (с первой позиции);
- (22–26) — номер остатка;
- (27–30) — четыре пробела;
- (31–38) — *X*-координата, формат F8.3;
- (39–46) — *Y*-координата, формат F8.3;
- (47–54) — *Z*-координата, формат F8.3.

Остальные позиции значения не имеют.

Все строки, которые не начинаются словом АТОМ, игнорируются. Поэтому в начале файла может присутствовать любое число строк описаний типа HEADER, TITLE и др. Пустые строки тоже допускаются в любом месте файла. Окончание списка атомов: строка, начинающаяся с TER (terminate). Если идентификатор атома содержит символы А и В (два варианта одного и того же атома), то считывается только первый вариант А.

Соответствующий фрагмент может быть вырезан, например, из более сложного файла, который включает несколько комплексов белок–ДНК/РНК, полученных в результате кристаллографического анализа. Выбор фрагмента можно делать путем визуального анализа структуры файла в любом текстовом редакторе, но это утомительно, так как в PDB-файлах масса комментариев и много данных. Для облегчения работы по редактированию PDB-файлов созданы небольшие программы ED-PDB2, ED-DNA, ED-DNA-C.

```

-- (c) JINR, Dubna Stepanenko V. A. 2010 --
Программа сборки НУКЛЕОТИДОВ фрагмента ДНК(РНК)
из разных цепей стандартного PDB-файла в одну цепочку.
ИГНОРИРУЮТСЯ карты кроме: HEADER, TITLE, AUTHOR,
АТОМ (для ДНК (РНК)), TER (последней),
РЕДАКТИРУЮТСЯ метки: А С G T U вместо DA DC DG DT DU !
Данные записываются в новый файл с добавкой Z к имени.
Длина имени L<8 символов. Пример: a7d.pdb >>> A7DZ.PDB

Введи имя PDB-файла (или просто кл. ENTER - завершение работы.)1
Цепь1 (метка В) 12 нуклеотид(а/ов)
Цепь2 (метка С) 12 нуклеотид(а/ов)
Цепь 1 пропустить?(Y/N)n
Введите N первого НУКЛЕОТИДА 16>=N>=5
5
Введите требуемое к-во НУКЛЕОТИДОВ 12>=L>0
12
Цепь 2 пропустить?(Y/N)n
Введите N первого НУКЛЕОТИДА 29>=N>=18
18
Введите требуемое к-во НУКЛЕОТИДОВ 12>=L>0
12
File #Z.PDB has been created.
Введи имя PDB-файла (или просто кл. ENTER - завершение работы.)_

```

Рис. 17. Пример работы с программой ED-DNA-C

Первые две программы убирают часть комментариев и оставляют только карты АТОМ с одной картой TERMINATE в конце данных.

Программа ED-DNA-C позволяет в режиме диалога последовательно выбрать на одной или более цепочках ДНК/РНК требуемые атомы по начальному номеру нуклеотида и заданному количеству их в каждой цепочке ДНК или РНК в диалоговом режиме (рис. 17). Для работы программы рекомендуется набор координат атомов нуклеотидного фрагмента, содержащий две цепи с общим числом не более 13–15 и не менее 5–6 нуклеотидных пар.

Эти небольшие программы работают очень быстро и позволяют обрабатывать в режиме диалога несколько файлов под ОС Windows, включая новые версии этой ОС: Win7 и Vista. При этом выводится стандартное окно программы с меню на черном фоне. Как видно на рис. 17, с помощью выбора параметров окна программы в пункте меню PROPERTIES можно изменить фон, шрифт и увеличить это окно.

Признаком завершения программы перекодировки является нажатие клавиши ВВОД вместо имени файла. В конце работы программы окно автоматически закрывается.

3. СТРУКТУРА КОМПЛЕКСА ПРОГРАММ КАРТОГРАФИРОВАНИЯ БЕЛКОВ И ДНК

Программы картографирования белковых структур [24] можно использовать на ПК достаточной мощности и без использования среды распределенных вычислений (рис. 18). Для работы с Linux-кластером создана программа-монитор NetdBBio, которая позволяет управлять файлами и запуском программ в системе AFS. В дальнейшем после перевода программ картографирования на язык C++ предполагается независимое накопление карт белковых структур и в рамках ОС Linux. Это существенно расширит область применения разработанного комплекса программ.

4. АНАЛИЗ РЕЗУЛЬТАТОВ

Созданный комплекс программ расширяет возможности исследования и визуализации большого класса белковых комплексов на разных уровнях (атомы, основания, цепочки, рельеф). По сути дела, биологи впервые получили возможность изучать атомы поверхности молекул достаточно сложной формы с предельной точностью в рамках одного полнофункционального комплекса картографирования.

В процессе разработки программ комплекса были решены следующие важные задачи.

- Программа картографирования SURFACE переведена на единый современный язык программирования и оптимизирована. В рамках ОС Windows

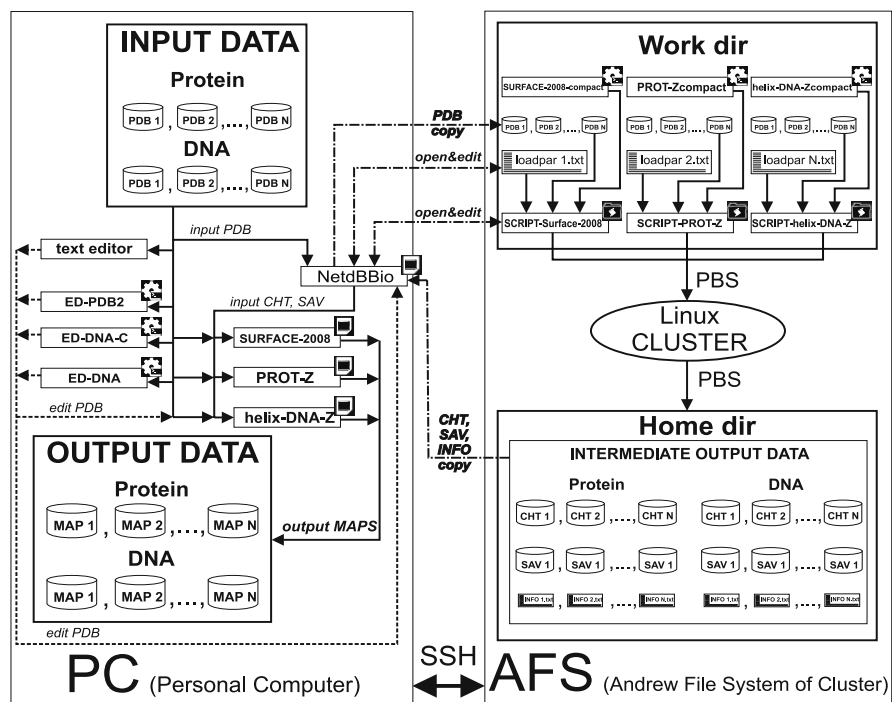


Рис. 18. Схема взаимодействия программ комплекса картографирования белковых структур

для работы исследователя достаточно иметь файл программы и PDB-файл данных для любой из программ комплекса (вместо пяти файлов ранее).

- Разработана оригинальная методика построения цилиндрической системы координат SC для визуализации атомов белковых структур с минимальными искажениями при сохранении интерфейса программ картографирования, основанного на использовании промежуточной сферической системы координат. Это позволило разработать методику картографирования спиральных белков, ДНК, РНК с введением шкалы измерений в ангстремах и устранить зависимость формы атомов от расстояния до центра масс.

- Сохранение карт в сжатом JPG-формате на дисках (вместо BMP) позволило в пять раз сократить нагрузку на внешние носители информации.

- Разработана методика и набор программ редактирования стандартных PDB-файлов для построения карт белковых структур из нескольких цепочек. Это позволило единым образом организовать ввод данных для ДНК, РНК и сложных белковых комплексов.

- Разработана методика и созданы программы для ускорения процесса массового картографирования в распределенной вычислительной среде.

Сохранены все основные режимы интерфейса и функциональные возможности исходной версии программы SURFACE: зуммирование фрагмента, размещение меток атомов и текста на карте, восстановление сеанса работы.

В настоящий момент рассчитаны карты для 27 файлов входных данных в среде распределенных вычислений:

- 9 комплексов с глобулярными белками, сильно отличающихся по типу структуры;
- 9 комплексов с узнающими альфа-спиралями спирализованных белковых факторов транскрипции, вырезанных из PDB-файлов координат комплексов белок–ДНК;
- 9 комплексов, содержащих фрагменты узнающих участков ДНК, вырезанных из PDB-файлов координат комплексов белок–ДНК.

Результаты вычислений сохраняются в домашний каталог на сервере в заданной структуре (рис. 19). В названиях файлов включены также обозначения интервалов вырезанных участков белков/ДНК.

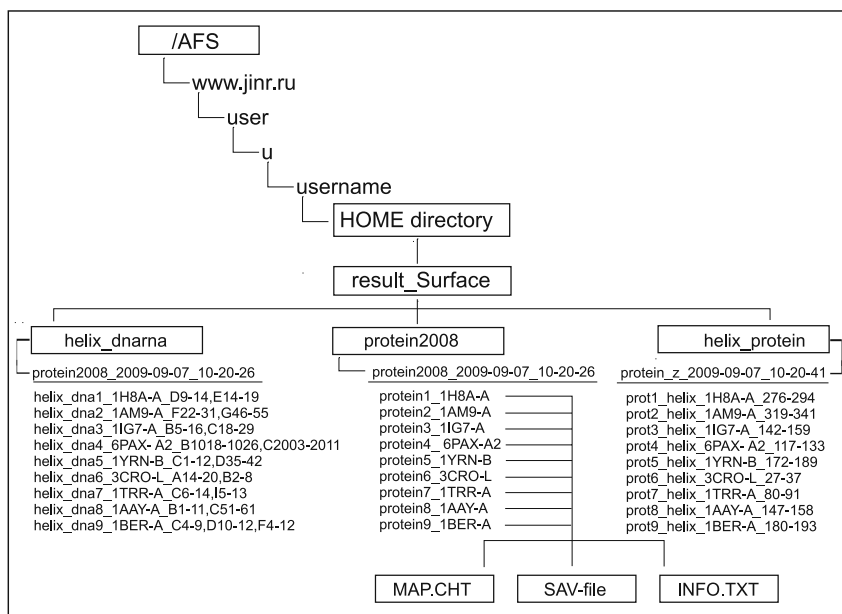


Рис. 19. Структура домашней (HOME) директории с результатами

На стадии отладки отдельных блоков находится программа NetdBBio, предназначенная для удобства администрирования результатов, полученных в результате счета в распределенной вычислительной среде, и управления массовым счетом на удаленной машине. Такой подход сокращает время картографирования белков и нуклеиновых кислот по сравнению с рутинной ручной работой в программах с графическим интерфейсом.

Программы комплекса и разработанные методы картографирования белковых структур не имеют аналогов в мире. Комплекс может использоваться при разработке принципиально нового класса лекарств, при детальном изучении свойств новых белков, при решении задач генной инженерии и экологии.

Работа поддержана РФФИ, грант № 11-07-00374.

Приложение

1. ТАБЛИЦЫ ФУНКЦИОНАЛЬНОЙ ПРИНАДЛЕЖНОСТИ ОСТАТКОВ И АТОМОВ

Таблица 1. Однобуквенный аминокислотный код

D Asp	E Glu	K Lys	R Arg	H His	S Ser	T Thr	Y Tyr	N Asn	Q Gln	C Cys
G Gly	A Ala	I Ile	L Leu	V Val	M Met	P Pro	F Phe	W Trp		

Таблица 2. Химическое деление остатков

Кислые	D	E	C-конец							
Щелочные	K	R	H		N-конец					
Гидроксильные	S	T								
Амидные	N	Q								
Гидрофобные	G	A	I		L		V	F	W	Y
Серосодержащие	C	M								
Пролинсодержащие	P									

Таблица 3. Функциональное деление остатков

1. Полярные заряженные отрицательные	D	E	C-конец							
положительные	K	R	H		N-конец					
2. Полярные незаряженные	S	T	Y		N		Q	C		
3. Неполярные	G	A	I		L		V	M	P	F

2. ТАБЛИЦЫ ФУНКЦИОНАЛЬНОГО ДЕЛЕНИЯ АТОМНЫХ ГРУПП И ИХ ОКРАСКА В ПОЛИНУКЛЕОТИДНЫХ ЦЕПЯХ ДНК И РНК

Таблица 4. Функциональное деление атомных групп

1. Полярные заряженные отрицательные	D	OD1	OD2					
	E	OE1	OE2					
	C-конец							
	положительные	K	NZ					
		R	NH1	NH2				
		H	ND1	NE2				
N-конец								
2. Полярные незаряженные пептидная группа	N	C	O					
	S	OG						
	T	OG1						
	placeN	OD1	ND2					
	Q	OE1	NE2					
	Y	OH						
	C	SG						
	R	NE						
	placeW	NE1						
	3. Неполярные атомы для всех остатков	CA	CB	CG	CD	CE	CZ	CH
		SD	CD1	CD2	и др.			

Таблица 5. Окраска по выделенному типу атомов

Тип группы	Обозначение	Атомы	Цвет (по умолч.)
1. Заряженные отрицательные	negative	Атомы кислорода O1P, O2P	Светло-малиновый
2. Полярные отрицательные	polar positive	Все N оснований	Светло-синий
3. Полярные электроотрицательные	polar negative	Все O, O3*-O5*	Светло-розовый
4. Неполярные	non-polar	Все C, C1*-C5*, H	Светло-серый
5. Основная цепь	main chain	Все атомы P	красный

Таблица 6. Окраска по выделенным химическим группам атомов

Тип группы	Обозначение	Атомы	Цвет (по умолч.)
Основания	bases	Только атомы оснований	см. выше
Фосфаты	phosphates	P, O1P, O2P	см. выше
Сахара	sugar	O5*, C5*, C4*, O4* C3*, O3*, C2*, C1*	см. выше

ЛИТЕРАТУРА

1. Альбертс Б. и др. Молекулярная биология клетки. В 3-х т. М.: Мир, 1994.
2. <http://www.rcsb.org/pdb/statistics/holdings.do>
3. Westbrook J., Fitzgerald P.M. The PDB format, mmCIF formats and other data formats. Structural Bioinformatics / P. E. Bourne and H. Weissig. Hoboken, N. Y.: John Wiley & Sons, Inc., 2003. P. 161–179.
4. <http://www.pymol.org/>
5. Лаврентьев М. А., Шабат Б. В. Методы теории функций комплексного переменного. 4-е изд., доп. и испр. М.: Наука, 1973.
6. Porter W. Introduction to map projections. N. Y.: Marcel Dekker, 1979.
7. Меццержаков Г. А. Теоретические основы математической картографии. М.: Недра, 1968.
8. Чиргадзе Ю. Н., Ларионова Е. А. Определяющая роль кластеров полярных остатков в структурах белковых факторов при узнавании большой бороздки двухспиральной В-ДНК // Мол. биология. 2003. Т. 37, № 2. С. 266–276.
9. Аглымов С. Р., Чиргадзе Ю. Н. Инструкция к программе SURFACE (version 0.1) для построения карт молекулярной поверхности глобулярных белков. Пушино: Изд-во Ин-та белка РАН, 1999.
10. Сухарев М. В. Основы Delphi. Профессиональный подход. СПб.: Наука и техника, 2004.
11. <http://microsoft.com>
12. Eitter D.M. Structured FORTRAN 77 for Engineers and Scientists (3rd ed.). The Benjamin/Cummings Publishing Company, Inc., 1990.
13. <http://lit999.narod.ru/images.html>
14. <http://lit.jinr.ru/view.php?var1=comp&var2=ccic&lang=rus&file=ccic/main-&menu=ccic/menu>
15. <http://lit.jinr.ru/ccic/usersguide/>
16. http://lit.jinr.ru/ccic/usersguide/index.php?link=3_
17. http://lit.jinr.ru/ccic/usersguide/index.php?link=2_#2.4_
18. <http://www.scientificlinux.org/>
19. <http://www.lazarus.freepascal.org/>
20. <http://www.python.org/download/releases/2.5/>
21. <http://www.riverbankcomputing.com/software/pyqt/intro>
22. <http://www.lag.net/paramiko/docs/paramiko-module.html>
23. <http://www.ghisler.com/>
24. Afanasiev O.A. et al. Software Complex for Computing Surface Maps of Helical Biopolymer Molecule Proteins and Nucleic Acids // Book of abstr. of Intern. Conf. «Mathematical Modeling and Computational Physics», July 7–11, 2009. Dubna, 2009. P. 171.

Получено 28 октября 2011 г.

Редактор *Е. В. Сабеева*

Подписано в печать 13.12.2011.

Формат 60 × 90/16. Бумага офсетная. Печать офсетная.

Усл. печ. л. 2,37. Уч.-изд. л. 2,82. Тираж 260 экз. Заказ № 57524.

Издательский отдел Объединенного института ядерных исследований
141980, г. Дубна, Московская обл., ул. Жолио-Кюри, 6.

E-mail: publish@jinr.ru

www.jinr.ru/publish/