Математическая биология и биоинформатика
2017. Т. 12. № 2. С. 354–374. doi: /10.17537/2017.12.354

========== ИНФОРМАЦИОННЫЕ И ВЫЧИСЛИТЕЛЬНЫЕ ==========
========== ТЕХНОЛОГИИ В БИОЛОГИИ И МЕДИЦИНЕ ==========

УДК: 123.4

Массовые вычисления электростатических потенциалов и карт молекулярных поверхностей белков и нуклеиновых кислот в распределенной компьютерной среде: организация и алгоритмы

Акишина Т.П.[1], Грохлина Т.И.*[2], Зрелов П.В.[1], Иванов В.В.[1,3], Полозов Р.В.[4], Сивожелезов В.С.[5], Степаненко В.А.[1], Чиргадзе Ю.Н.[6], Яковлев А.В.[1]

[1]*Объединенный институт ядерных исследований, Дубна, Россия*
[2]*Институт математических проблем биологии РАН – филиал Института прикладной математики им. М.В. Келдыша РАН, Пущино, Россия*
[3]*Национальный исследовательский ядерный университет ”МИФИ”, Москва, Россия*
[4]*Институт теоретической и экспериментальной биофизики РАН, Пущино, Россия*
[5]*Институт биофизики клетки РАН, Пущино, Россия*
[6]*Институт белка РАН, Пущино, Россия*

Аннотация. Важными факторами, определяющими ключевые процессы транскрипции и трансляции в клетке, являются распределения электростатических потенциалов и структуры молекулярных поверхностей нуклеиновых кислот, факторов транскрипции и их комплексов с ДНК. Однако расчеты электростатических потенциалов и построение карт молекулярных поверхностей требуют много времени и больших вычислительных ресурсов. Для решения этих задач нами разработаны технология и комплекс программ для вычислений в распределенной компьютерной среде, содержащей несколько тысяч процессорных ядер.

Ключевые слова: *ДНК, РНК, белки, факторы транскрипции, электростатический потенциал, карта молекулярной поверхности, массовые вычисления, распределенная компьютерная среда.*

ВВЕДЕНИЕ

Физико-химические свойства и, в особенности, характеристики поверхностей молекул ДНК, РНК и белков, участвующих в процессах транскрипции генов и трансляции РНК, существенным образом определяют функцию клетки. Взаимодействие ДНК с полимеразами, факторами транскрипции и другими белками, которые играют ключевую роль в транскрипции и ее регуляции – один из самых важных примеров молекулярного узнавания, где происходит селективное связывание белка со специфическим участком молекулы ДНК [1]. Специфика связывания может быть оценена разницей в величине свободных энергий связывания молекул, которая варьируется от 40 до более чем 80 кДж/моль [2]. Наиболее существенными при этом, в особенности в случае

*grokhlina@mail.ru

ДНК-белковых взаимодействий, являются электростатические взаимодействия молекул, которые имеют первостепенную важность в многошаговом процессе узнавания белком ДНК. В начале этого процесса, который представляет собой скольжение узнающего белка по поверхности ДНК на расстоянии около 15 Å от продольной оси ДНК [3], электростатические взаимодействия являются определяющим физическим фактором, так как на таких расстояниях энергии электростатических взаимодействий намного превышают энергии водородных связей, дисперсионных взаимодействий и т.п. Еще более существенно, что вычисление распределения электростатического потенциала вдоль длинных цепочек позволяет найти связи между функциональными свойствами ДНК и ее физическими свойствами [4]. Недостаток классификации функции ДНК по ее нуклеотидной последовательности состоит в том, что она не имеет явного физического обоснования, хотя известно, что электростатические свойства молекулярной поверхности ДНК коррелируют с ее нуклеотидной последовательностью.

В данной работе предлагается подход и алгоритм организации массовых вычислений электростатических потенциалов биополимеров (ДНК, РНК, белков) с использованием многосеточного метода решения уравнения Пуассона–Больцмана и рельефов карт молекулярных поверхностей. Это позволяет одновременно обрабатывать большое число фрагментов двуспиральной ДНК и белков. Ранее процесс вычисления требовал огромной подготовительной работы, а время вычисления электростатических потенциалов, например, большого числа длинных ДНК, доступными программами было недопустимо большим. Предложенная процедура позволяет также существенно расширить применение автоматизированного подхода для большого числа фрагментов ДНК и факторов транскрипции так, что становится возможным расчет электростатических потенциалов и карт поверхностей нескольких тысяч структур биополимеров или их фрагментов одновременно.

МЕТОДЫ ВЫЧИСЛЕНИЙ ЭЛЕКТРОСТАТИЧЕСКИХ ПОТЕНЦИАЛОВ БИОПОЛИМЕРОВ

Для расчетов электростатических потенциалов биополимеров использовалось нелинейное уравнение Пуассона–Больцмана. Парциальные заряды атомов биополимеров вычислялись согласно силовому полю AMBER94 [5]. Простанственное распределение зарядов противоионов приближалось распределением Больцмана, диэлектрическая константа биополимеров выбиралась равной 2.0, а растворителя – 80.0. Электролит для моделирования распределения противоионов был принят в соотношении 1:1 (заряды ионов $z_1 = 1, z_2 = -1$) при концентрациях 50÷150 ммоль, близких к физиологическим значениям. Решение находилось многосеточным методом [6, 7, 8]. Максимальное разрешение сетки $2000 \times 20 \times 20$ узлов выбиралось так, чтобы интервал между узлами сетки был меньше 1Å. Был разработан алгоритм решения нелинейного уравнения Пуассона–Больцмана, позволяющий эффективно вычислять электростатические потенциалы не только глобулярных белков, но и нуклеиновых кислот и мембран [9]. В описываемой реализации многосеточного решения нелинейного уравнения Пуассона–Больцмана отсутствуют ограничения на соотношение числа узлов сетки по каждой из координат. Это позволяет обрабатывать мембранные комплексы и длинные, до 1000 пар оснований, фрагменты ДНК.

Визуализация выполнялась с использованием модифицированной программы молекулярной графики MOLMOL [10]: электростатический потенциал визуализировался на ”поверхности скольжения” молекулы ДНК на удалении 15Å от ее продольной оси. Схематическое изображение B-формы ДНК представлено на рисунке 1. Разными цветами отмечены разные уровни электростатического потенциала. Созданное программное

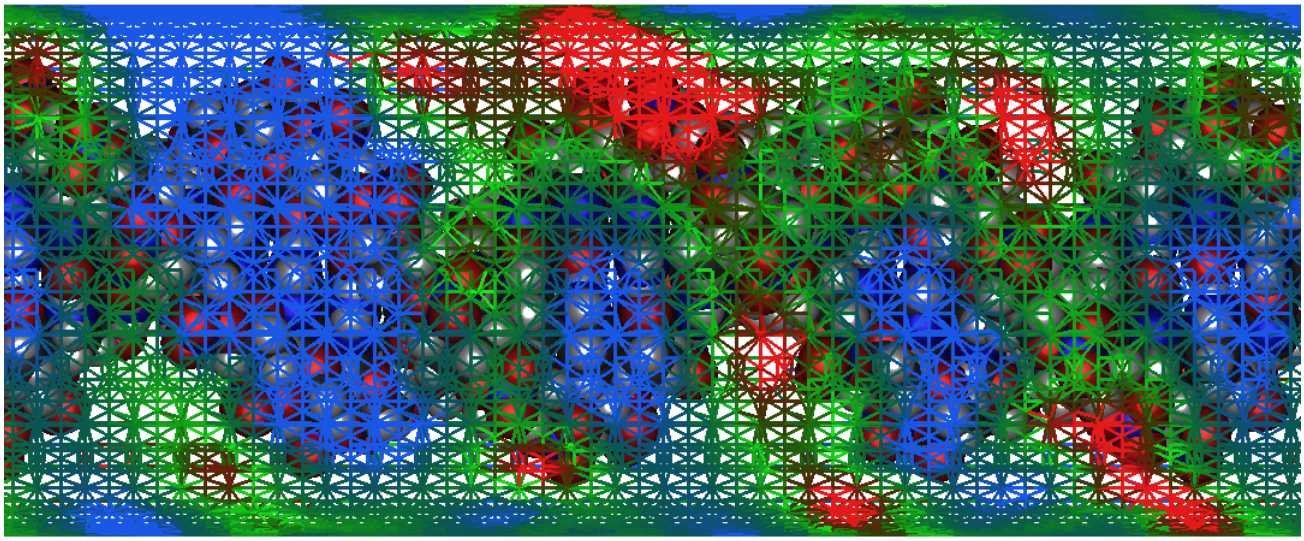

Рис. 1. Пример визуализации электростатического потенциала ДНК (B-форма).

обеспечение было применено для расчета электростатических потенциалов промоторных участков ДНК.

Решение уравнения Пуассона–Больцмана

Вычисление электростатического потенциала $\varphi(\boldsymbol{r})$ для фрагмента ДНК осуществлялось на основании решения уравнения Пуассона–Больцмана, которое описывает электростатический потенциал в растворе вокруг молекулы ДНК:

$$-\bigtriangledown(\epsilon(\boldsymbol{r})\bigtriangledown\varphi(\boldsymbol{r})) = 4\pi(\rho_0(\boldsymbol{r}) + \rho_1(\varphi(\boldsymbol{r}))), \tag{1}$$

здесь $\boldsymbol{r} = (x, y, z) \in R^3$,
φ – искомый электростатический потенциал,
ϵ – диэлектрическая проницаемость,
ρ_0 – распределение заряда ДНК.

$$\rho_0(\boldsymbol{r}) = \sum_i ez_i\delta(|\boldsymbol{r} - \boldsymbol{r}_i|), \tag{2}$$

где z_i – парциальный заряд i-ого атома молекулы в единицах элементарного заряда,
$\boldsymbol{r}_i$ – радиус-вектор i-ого атома,
e – элементарный заряд (абсолютное значение заряда электрона),
δ – дельта-функция Дирака и

$$\rho_1(\boldsymbol{r}) = \sum_i n_i ez_i \exp(ez_i\varphi/k_BT). \tag{3}$$

Для достаточно малых значений потенциала при $\varphi \ll k_BT/e$ получаем линеаризованное уравнение Пуассона–Больцмана:

$$-\bigtriangledown(\epsilon(\boldsymbol{r})\bigtriangledown\varphi(\boldsymbol{r}) + \kappa^2\varphi = 4\pi\rho_0(\boldsymbol{r}), \tag{4}$$

$$\kappa^2 = 4\pi e^2 \sum_i n_i z_i^2 / k_B T, \tag{5}$$

где n_i – концентрация i-го типа ионов,
z_i – заряд иона,
k_B – константа Больцмана,
T – абсолютная температура (T = 300 K).

Граничные условия для потенциала $\varphi(\infty)$ определялись приближением Дебая–Хюккеля:

$$\varphi(\boldsymbol{r})|_\Gamma = \sum_i \frac{e z_i \exp(-\kappa|\boldsymbol{r} - \boldsymbol{r}_i|)}{|\boldsymbol{r} - \boldsymbol{r}_i|}, \tag{6}$$

где Γ – поверхность достаточно большого по размерам параллелепипеда и $\boldsymbol{r} = (x, y, z) \in \Gamma$.

Для решения задачи (4) с граничными условиями (6) в области Γ применялся метод конечных элементов. Система линейных алгебраических уравнений (СЛАУ) решалась многосеточным методом (ММ) [6, 7, 8], а для решения нелинейного уравнения (1) использовались итерации:

$$-\bigtriangledown(\epsilon(\boldsymbol{r}) \bigtriangledown \varphi^{n+1}) + \alpha\varphi^{n+1} = 4\pi(\rho_0 + \rho_1(\varphi)) + \alpha\varphi^n, \tag{7}$$

где φ^n – приближение решения на n-ой итерации.

Начальным приближением для итерационного процесса (7) является решение задачи (4).

Многосеточный алгоритм

Для решения задачи (1)–(2) с граничными условиями (6) мы использовали метод конечных элементов, где приближенное решение находилось как сумма базисных функций $\Phi^i(x, y, z)$:

$$\varphi = \sum_i u^i \Phi^i. \tag{8}$$

Применяя метод Галеркина [11] к уравнениям (1), (7) и (8), получили СЛАУ $Au = f$, для которой находили оценки величин коэффициентов u_i. Далее, используя ММ, уточняется решение СЛАУ. Шаг вложенной сетки вычисляется следующим образом:

$$h_{l-1} = 2h_l; \quad S_{l-1} \subset S_l; \quad l = 1, \ldots, L, \tag{9}$$

где h – значение шага сетки,
S_i – пространство базисных функций на i-ой сетке.

Окончательное решение находится на сетке с номером L. Она строится итеративно, с использованием множества вспомогательных сеток $l = 0, 1, \ldots, L-1$. На каждой итерации за решение принимается минимальное значение функции φ на сетке $L-l$. Далее рекурсивно применяется этот же алгоритм до достижения сетки 0. Сетка 0 является самой грубой (с наибольшим значением шага h_0). Она содержит небольшое число неизвестных и, таким образом, СЛАУ для этой сетки довольно просто находится прямым методом, например, методом исключения Гаусса–Зейделя.

Процесс итераций завершается, когда норма ошибок удовлетворяет следующему критерию:

$$\frac{\| f_L - A_L u_L \|}{\| f_L \|} \leq 10^{-6}. \tag{10}$$

Обычно для сходимости итерационного процесса (7) с заданной точностью достаточно

4–5 шагов.

ЭЛЕКТРОСТАТИКА МОЛЕКУЛ НУКЛЕИНОВЫХ КИСЛОТ

Пространственные распределения электростатических потенциалов имеют различные геометрические формы, и их классификация может быть проведена, например, методами морфологии [12]. Наиболее точный метод вычисления электростатических потенциалов и энергий, доступный для макромолекулярных систем, в численном виде дает решение уравнения Пуассона–Больцмана на прямоугольной сетке [4]. Но этот метод ранее не использовался для расчетов длинных цепей ДНК, так как численные алгоритмы решения не позволяют использовать сильно различающиеся значения чисел узлов сетки по каждой из декартовых координат. Наш метод обрабатывает несколько сотен и даже тысячи пар нуклеотидов ДНК.

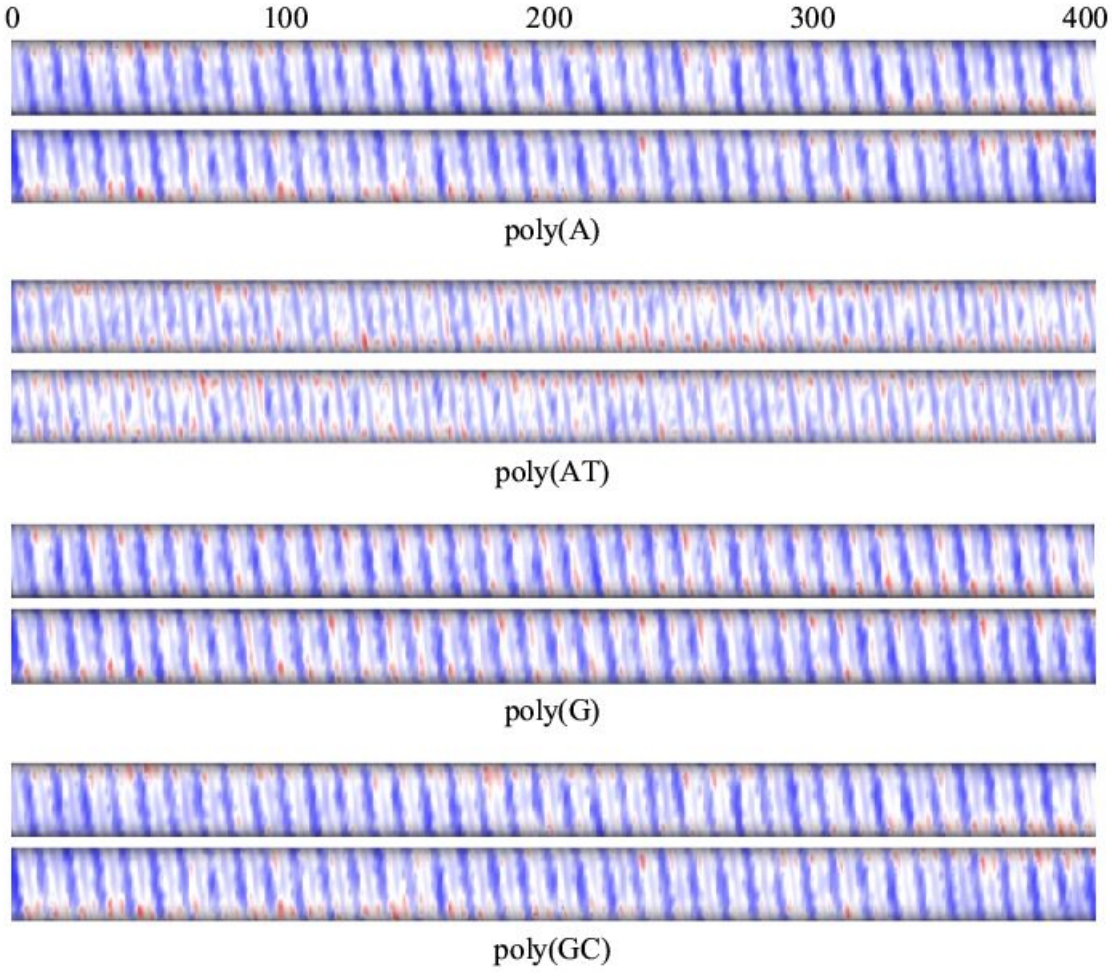

Рис. 2. Пространственные распределения электростатического потенциала вокруг молекул ДНК (B-форма) с периодической нуклеотидной последовательностью. Каждая молекула показана с двух сторон, отличающихся вращением на 180^o вокруг продольной оси.

Иллюстрируемые в этой работе промоторные области ДНК из *E. coli* имеют до 411 пар оснований. Горизонтальная ось на рисунках 2 и 3 совпадает с осью спирали ДНК. Цветовая гамма отражает распределение электростатического потенциала в единицах k_BT/q, где k_BT – тепловая энергия, а q – заряд протона. В этих единицах красный цвет соответствует значению -1.3, синий – -0.8, а белый – -1.05, и все промежуточные значения потенциалов окрашивались в соответствии с результатами интерполяции [10].

Разница между красным и синим цветами в этих единицах составила 0.5 k_BT/q. Для узнавания ДНК белком достаточно десяти контактов белок-ДНК, которые обычно имеют место [13, 14, 15]. Это соответствует пяти единицам k_BT/q.

Нуклеотидные последовательности промоторных областей ДНК *E. coli* были взяты из работ [16, 17]. Точка старта транскрипции – позиция 257, кодирование идет в сторону убывания номеров позиций, а область промотора – от той же точки в сторону возрастания номеров.

На рисунке 2 представлены электростатические потенциалы молекулы ДНК с периодическими нуклеотидными последовательностями *poly*(A), *poly*(AT), *poly*(G) и *poly*(GC).

Как видно из рисунка 2, электростатические потенциалы тоже имеют периодическую природу. Тот факт, что эта периодичность не очень четко проявилась на цилиндрической поверхности, обусловлен пространственным строением B-формы ДНК. Качественно электростатические потенциалы *poly*(AT) заметно отличаются от потенциалов других периодических последовательностей. Основное различие заключается в наличии сильной дипольной составляющей в электростатическом потенциале поперек двойной спирали ДНК. Действительно, интенсивные синие пятна (меньший отрицательный потенциал) расположены далеко от интенсивных красных пятен (больший отрицательный потенциал). Напротив, периодические нуклеотидные последовательности *poly*(A), *poly*(G) и *poly*(GC) (рис. 2) показывают однородное распределение электростатического потенциала поперек двойной спирали, визуально более похожее на квадрупольное распределение потенциала.

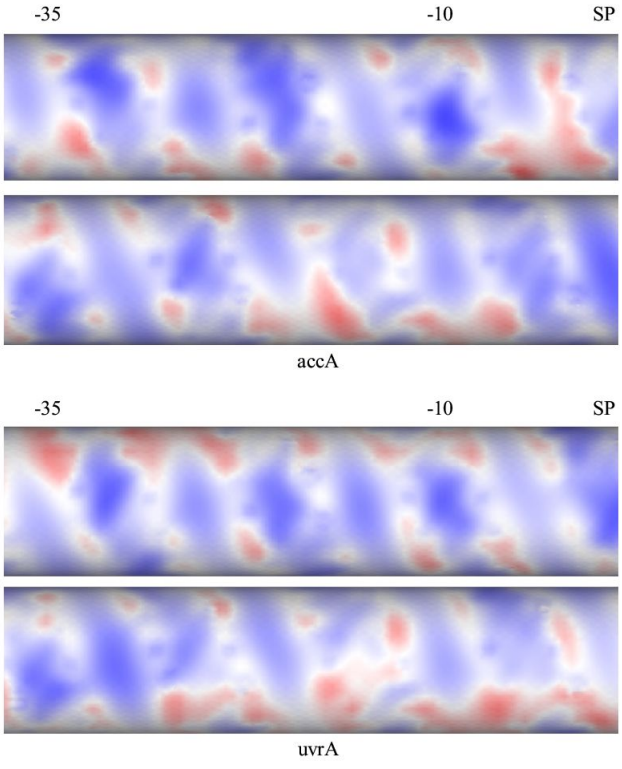

Рис. 3. Пространственные распределения электростатического потенциала вокруг промоторов accA и uvrA *E. coli* представлены в двух проекциях, отличающихся вращением на 180^o вокруг продольной оси ДНК.

Распределения электростатических потенциалов в функционально важных областях для двух промоторов accA и uvrA *E. coli* (это позиции -35 и -10, стартовая точка SP) представлены только в этих областях (рис. 3)

Длинные и многочисленные нуклеотидные последовательности ДНК представляют собой естественную область приложения для распределенных вычислений. Нуклеотидные последовательности промоторов, ответственные за регулирование транскрипции гена, требуют особого внимания. Заметим, что классификация ДНК, основанная на попарном сравнении их электростатических потенциалов, будет требовать применения распределенных или облачных вычислений. Возможно, что это единственный путь для анализа колоссального количества известных генов. Так, только в геномах млекопитающих имеется около 35000 генов. Нуклеотидная последовательность одного промотора включает несколько сотен пар оснований.

Ниже рассмотрена процедура, которая описывает подготовку и организацию массовых вычислений электростатических потенциалов [18].

ОРГАНИЗАЦИЯ МАССОВЫХ ВЫЧИСЛЕНИЙ ЭЛЕКТРОСТАТИЧЕСКИХ ПОТЕНЦИАЛОВ БИОПОЛИМЕРОВ

Для построения поверхности электростатического потенциала существуют программы, например, MOLMOL, но в них нет возможности расчета протяженных, как ДНК, и уплощенных, как мембрана, структур. Поэтому для этих целей был реализован модифицированный многосеточный метод конечных элементов [6, 7]. В расчетах, кроме известных программных продуктов HyperChem и MOLMOL, использовалась программа elefull [9] , созданная для вычисления электростатического потенциала ДНК с применением сетки с переменным шагом. Был создан также ряд вспомогательных процедур, оформленных в виде программ, скриптов (сценариев) и макросов, для автоматизации подготовительной работы и сокращения ошибок ручного ввода. На рисунке 4 представлена блок-схема организации вычислений электростатического потенциала молекул ДНК. Описание схемы и вспомогательных процедур приведено в Приложениях A-E.

ОРГАНИЗАЦИЯ МАССОВЫХ ВЫЧИСЛЕНИЙ КАРТ МОЛЕКУЛЯРНЫХ ПОВЕРХНОСТЕЙ ФРАГМЕНТОВ ДНК И БЕЛКОВ В РАСПРЕДЕЛЕННОЙ СЕТИ ОБЪЕДИНЕННОГО ИНСТИТУТА ЯДЕРНЫХ ИССЛЕДОВАНИЙ (ОИЯИ)

Построение карт молекулярной поверхности (структурных карт) белков и фрагментов ДНК и РНК существенно расширяет возможности изучения пространственного распределения зарядов и рельефов поверхности этих молекул. Развитые нами методы и алгоритмы в значительной степени дополняют возможности качественного анализа структурных и физико-химических характеристик этих биополимеров и их комплексов. Карты молекулярной поверхности биополимеров (см. рис. 5) с функциональной раскраской (т.е. окрашенными по определенным правилам атомами, образующими рельеф молекулярной поверхности), метками и наложением уровня рельефа дают возможность точно оценить положение атомов в исследуемом фрагменте, провести аналогию с распределением электростатических потенциалов по поверхности цилиндра (см. рис. 1). Карты создаются в прямоугольной декартовой системе координат.

Массовые вычисления структурных карт были организованы в режиме многозадачности в сети распределенных вычислений, где можно использовать несколько тысяч процессорных ядер. В этом случае копии одних и тех же программ для различных фрагментов биополимеров запускались в пакетном режиме с персонального компьютера, а полученные результаты передавались обратно для детального анализа или выводились в виде карт фрагментов ДНК/РНК с функциональной раскраской и масштабной шкалой. Для построения структурных карт глобулярных белков мы используем проекцию Аитова–Хаммера [21, 22] для фибриллярных структур (ДНК, РНК, фибриллярных

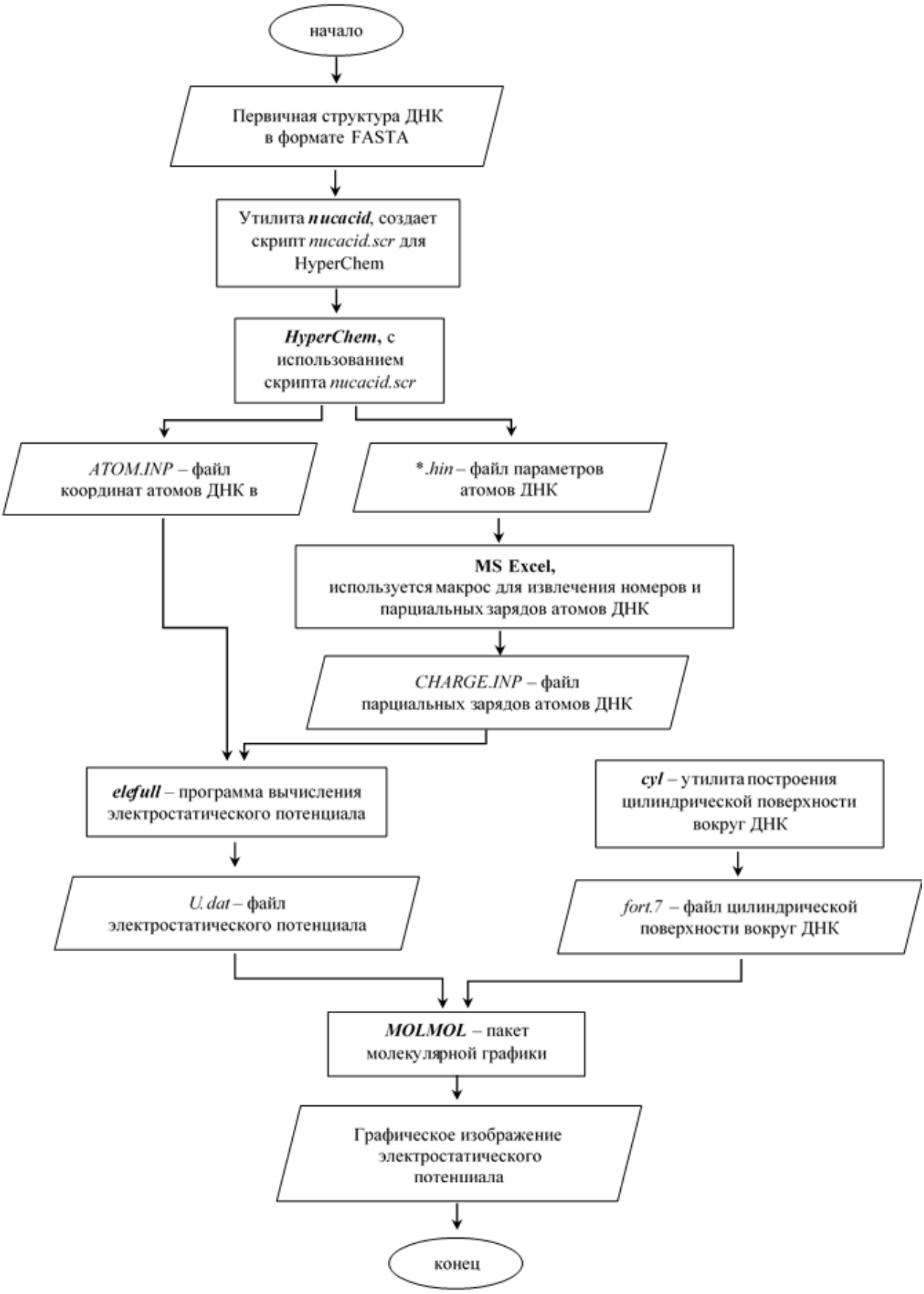

Рис. 4. Схема организации вычислений электростатического потенциала молекул ДНК.

белков) – псевдоконформные преобразования [23].

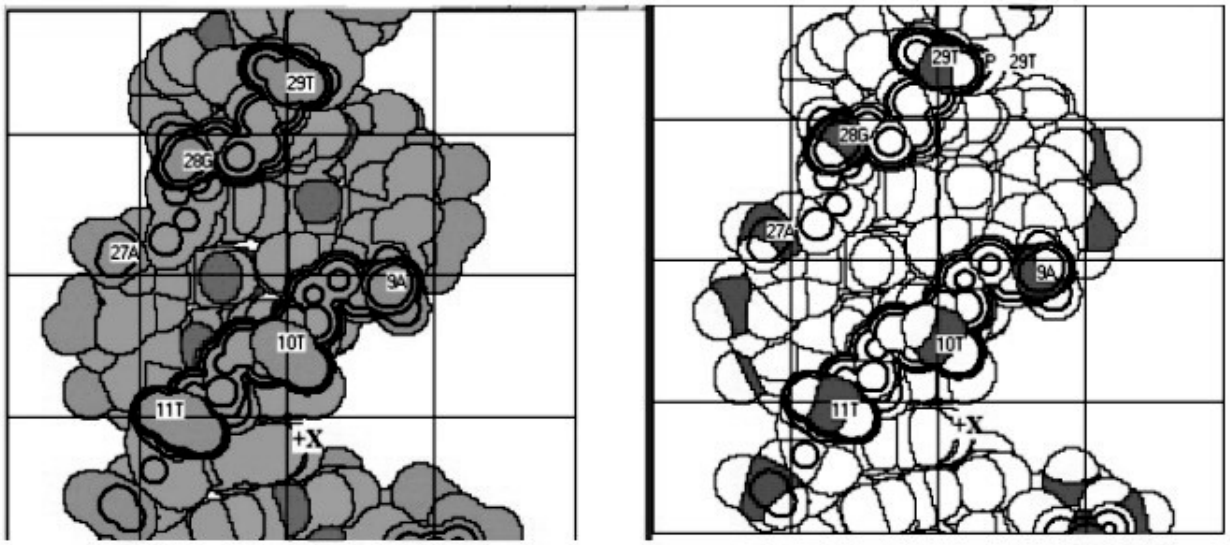

Рис. 5. Пример двух структурных карт фрагмента ДНК. На обеих картах показаны уровни рельефа поверхности отдельных атомов или атомных групп молекулы.

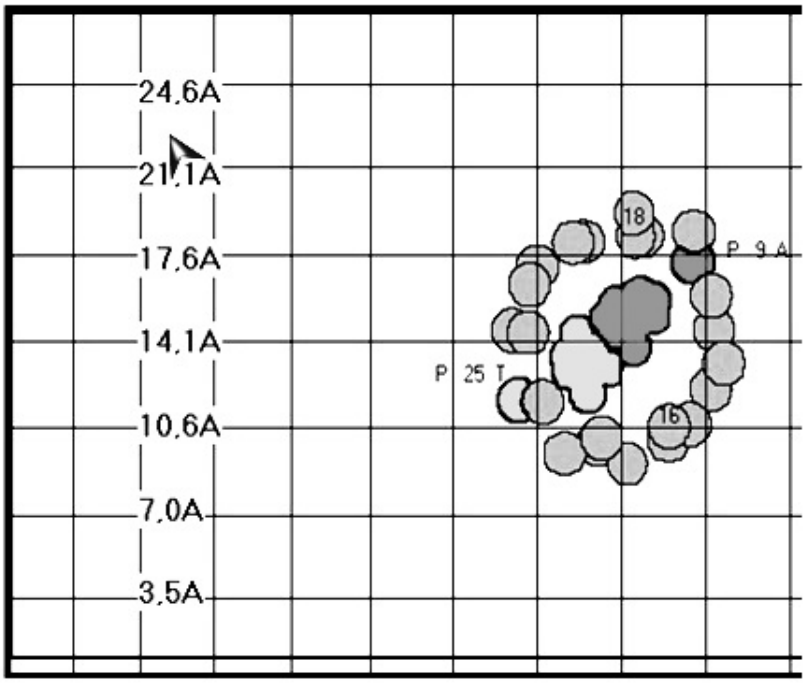

Рис. 6. Карта пространственного расположения атомов фосфора и пары оснований фрагмента ДНК (вид с торца вдоль длинной оси молекулы ДНК в B-форме). Приведена шкала расстояний.

Вычислительные фермы Центрального вычислительного комплекса ОИЯИ образуют Linux-кластер [24], [25] с распределенной файловой системой AFS (Andrew File System) [26]. Запуск задач в пакетном режиме, получение пользователем результатов выполняются с помощью системы пакетной обработки заданий PBS (Processing Batch System)[27].

Организация вычислений в распределенной сети потребовала развития ряда программ, работающих под ОС Unix, и создания программ для обмена файлами с компьютерами под ОС Windows. Для этого мы использовали скрипты – файлы, содержащие команды PBS для запуска и управления процессами в пределах распределенной сети вычисления. Программа управления графическим интерфейсом работает под ОС Windows и контролирует процесс вычислений.

Вычисления структурных карт биополимеров

Консольные версии программ картографирования созданы на языке Delphi с использованием межплатформенной среды программирования Lazarus [28].

Программы SURFACE-2008-compact, PROT-Zcompact и helix-DNA-Zcompact предназначены для вычисления файлов данных для карт глобулярных, спиральных белков и ДНК/РНК соответственно [29, 30]. При запуске в сети распределенных вычислений они требуют, кроме названия программы, ввода трех парамеров: имени pdb-файла исследуемого фрагмента, имени файла эмуляции параметров настройки интерфейса, названия файла результатов – файла с функциональной картой фрагмента. Содержание файла эмуляции параметров настройки интерфейса зависит от вида молекулы (фрагмента биополимера).

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

Ограничения на параметры обработки

- число разбиений площади карты – 900;
- число атомов – около 2000;
- (или число остатков – около 200);
- число уровней изолиний – 100;
- число адресов атомов для раскраски – 100;
- число адресов остатков для раскраски – 100;
- общее число надписей и меток – 100.

Функциональная и рельефная раскраска атомов

- положительно заряженные;
- отрицательно заряженные;
- все заряженные;
- полярные;
- неполярные;
- атомы главной цепи;
- раскраска атомов по их адресу;

- контуры поверхностного рельефа на атомах или атомных группах.

Для автоматизации вычислений карт большого числа фрагментов написано несколько скриптов (см. Приложение F).

ЗАКЛЮЧЕНИЕ

Одними из ключевых факторов, определяющих процессы транскрипции и трансляции в клетке, являются пространственные распределения электростатических потенциалов вокруг молекул ДНК, РНК, белков и их комплексов, а также структура их поверхностей. Имея информацию о распределении электростатических потенциалов и рельефов молекулярной поверхности функционально важных участков биомолекул, можно эффективно изучать и визуально представлять свойства и структуры биополимеров.

Задачи вычисления электростатических потенциалов и рельефов молекулярной поверхности исследуемых биомолекул являются сложными и требуют много времени и больших вычислительных ресурсов. С целью преодоления указанных трудностей нами были разработаны технология и соответствующий комплекс программ для проведения массовых расчетов электростатических потенциалов и карт молекулярных поверхностей молекул указанных биополимеров в распределенной компьютерной среде, содержащей несколько тысяч процессорных ядер.

Возможности развитой нами методики вычислений были продемонстрированы на примерах расчетов электростатических потенциалов и соответствующих карт, важных для изучения процессов молекулярного распознавания и связывания ДНК с белками. Отметим, что в процессе выполнения данной работы с помощью разработанных процедур и программ были проведены расчеты электростатических потенциалов и структурных карт для более 100 молекул биополимеров.

Работа поддержана грантом РФФИ №17-07-01331.

ПРИЛОЖЕНИЯ

Приводимые ниже инструкции и описания скриптов и файлов относятся к конкретному случаю реализации общего подхода к организации массовых вычислений электростатических потенциалов и карт молекулярных поверхностей белков и нуклеиновых кислот. Выполнение описанного процесса на других вычислительных платформах требует надлежащей адаптации.

Приложение А. Описание схемы вычислений электростатического потенциала

В вычислениях использованы следующие программные разработки:

- NUCACID – программа для автоматизации ввода данных для HyperChem и исключения ошибок при ручном вводе нуклеотидной последовательности до несколько сотен нуклеотидов.

  Исходные данные: текстовый файл, содержащий строку с нуклеотидной последовательностью в направлении от 5’ к 3’.

  Результат: скрипт для HyperChem (файл с расширением src) (см. Приложение В).

- elefull – программа для вычисления электростатического потенциала ДНК. *Исходные данные*:

  – ATOM.INP – файл координат атомов в формате pdb;
  – CHARGE.INP – файл, содержащий заряды атомов;
  – ARAD.INP – файл, содержащий параметры сетки;
  – PARAM.INP – файл параметров вычислений электростатического потенциала;

  Результат: U.DAT – файл с рассчитанным электростатическим потенциалом;

- cyl –- утилита для построения цилиндрической поверхности вокруг ДНК.

  Исходные данные: один из макросов showcyl.mac, cyl_-35.mac, cyl_-75.mac. Это варианты расчета параметров цилиндрической поверхности вокруг молекулы ДНК для визуализации потенциалов.

  Результат: fort.7 – файл параметров визуализации цилиндрической поверхности вокруг ДНК.

- HyperChem (нами использовалась версия HyperChem 7.01 [19, 20]), согласно принципу комплементарности, добавляет к исходной вторую цепочку ДНК и определяет парциальные заряды для всех атомов фрагмента. В рачетах мы использовали структуру ДНК в В-форме. Парциальные заряды атомов нуклеотидов локализуются на центрах каждого атома, их значения соответствуют силовому полю AMBER [5] и записываются в файл с расширением hin. Параллельно записывается ATOM.INP – файл в pdb-формате, один из входных файлов программы elefull.

- Абсолютные значения зарядов на атомах O1 и O2 фосфатных групп нуклеотидов уменьшались на величину $0.25q$, где q – это заряд протона, принимая во внимание известный эффект конденсации противоионов около заряженных атомов фосфатных групп. Диэлектрические константы были приняты равными двум – для ДНК и 80 – за пределами ее молекулярной поверхности. Электростатический потенциал визуализировался на поверхности цилиндра с радиусом 15 Å относительно оси фрагмента ДНК, приблизительно в 5 Å от сахаро-фосфатного

Таблица 1. Список файлов и их назначение

Файл	**Описание**
ARAD.INP	полуширина нормального распределения заряда по самым близким точкам сетки
atom.ent	декартовы координаты атомов промотора
atom.hin	декартовы координаты атомов промотора и их зарядов
atom.tmp, atom.xls	временные файлы для преобразования формата
charge.inp	файл атомных зарядов, используемый для электростатических вычислений
elefull.out	файл регистрации процесса электростатических вычислений
fort.7	содержит данные о сетке цилиндра для проецирования электростатических потенциалов
PARAM.INP	параметры электростатических вычислений
showcyl.mac, cyl-35.mac, cyl-75.mac	файлы макросов для выбора визуализации потенциалов из файла U.dat для функционально важных участков ДНК
U.dat	файл для фрагмента ДНК, содержит данные для расчета значений потенциалов на поверхности цилиндра

остова, что соответствует поверхности скольжения ДНК и первой стадии распознавания ДНК белком.

- После дополнительного редактирования с использованием макроса MS EXCEL, из hin-файла получается файл CHARGE.INP – второй файл данных программы elefull, и с ее помощью вычисляется электростатический потенциал ДНК с применением сетки с переменным шагом. Результаты этих вычислений записываются в файлы U.dat, а результат работы утилиты cyl.exe – в файл fort.7, которые и позволяют получить изображение в программе MOLMOL с помощью макроса (см. Приложение C). Кроме перечисленных, для работы программы elefull необходим файл с параметрами сетки ARAD.INP и файл с параметрами вычисления потенциалов PARAM.INP, который содержит число узлов сетки, значения диэлектрических констант и ионной силы.

- Действия, которые следует выполнить до начала описанных процедур, включая резервирование места на диске, настройку программы HyperChem, копирование необходимых файлов и т.д., подробно приведены в Приложении D.

Приложение В. Сценарий (скрипт) для HyperChem: автоматизация ввода нуклеотидной последовательности ДНК и генерация трехмерной структуры ДНК

```
nucleic-double-strand yes
translate-whole-molecules yes
nucleic-b-form
add-nucleic-acid dA
add-nucleic-acid dC
add-nucleic-acid dG
add-nucleic-acid dA
...
add-nucleic-acid dT
add-nucleic-acid dT
add-nucleic-acid 3cap
file-format hin
write-file atom.hin
file-format pdb
write-file atom.ent
menu-file-exit
```

Приложение С. Макрос для получения изображения в программе MOLMOL

```
InitAll yes
ReadPdb Atom.inp
RotateY -90
SelectAtom ':257,559'
AddText 0.04 '| start'
SelectAtom ':247,569'
AddText 0.04 '| -10'
SelectAtom ':222,594'
AddText 0.04 '| -35'
SelectPrim 'text'
MovePrim 0 0.5 0
ColorPrim 0 0 0
SelectPrim 'none'
DrawPrec 2 1
AddIsosurface 'fort.7' 0.067
ReadPot 'U.DAT'
PaintSurface pot 0 1.4 0.2 3 '-1.3 1 0 0 -1.1 .9 .9 .9 -0.8 0 0 1'
PlotPar 21 29.7 18 0 2500 0 0 0 0 1 1 75
SelectPrim 'surface'
MaterialPrim 0.2 0.9 0. 30 0 1 1
MoveX -120
ZoomAbs 0.08
PlotPng -35front.png
RotateX 180
PlotPng -35back.png
InitAll yes
Quit yes
```

Приложение D. Запуск программы HyperChem

1. Создать рабочую папку программы HyperChem. В ней должны храниться все рабочие файлы;
2. Установить программу HyperChem (версия 7.0 и старше).
3. В HyperChem выполнить следующие действия:
   - Setup ≫ Molecular Mechanics ≫ AMBER
   - Меню ≫ Select Parameters Set ≫ Amber94
4. В директорию Windows\System32 скопировать программы nucacid.exe, elefull.exe и unix-утилиты для Windows grep.exe, cut.exe.

Построить трехмерную структуру ДНК, для чего мы использовали программу HyperChem, можно другими способами, например, с помощью Macromoluculebuilder (MMB) [31].

Приложение Е. Организация массовых вычислений электростатических потенциалов биополимеров

Для реализации массовых вычислений в интерактивном режиме выполнялся ряд шагов с помощью созданного нами в среде MS Excel программного файла. Этот файл представляет собой набор последовательно вызываемых макросов, которые формируют и запускают на выполнение командные файлы операционной системы (*.*bat*). Функция командных файлов - выполнение однотипных операций для каждого из исследуемых фрагментов ДНК одним командным действием.

1. Создать в рабочей папке папку (подкаталог) для каждого исследуемого фрагмента ДНК и скопировать с нее все необходимые файлы – текстовый файл с последовательностью, ARAD.INP, PARAM.INP, showcyl.mac, cyl_-35.mac, cyl_-75.mac;

2. Создать в каждом подкаталоге файл сценария (скрипта для HyperChem, см. Приложение В) *.*scr*;

3. Модифицировать файлы *.*scr* для того, чтобы в структуру ДНК были включены водороды и исключены ковалентные связи (строки CONECT в файлах *.*pdb*).

4. Выполнить скрипты HyperChem (файлы *.*scr*) в каждом подкаталоге.

5. С помощью утилит renum, grep и cut выполнить в каждом подкаталоге следующие действия:

    - Записать файл atom.inp – копию файла ATOM.ENT с измененной нумерацией строк;
    - Скопировать строки ATOM из файла ATOM.ENT в файл atom.hin;
    - Создать из файла atom.hin вспомогательный файл atom.xls с сохранением номера, заряда и имени атомов.

6. Создать в каждом подкаталоге файл CHARGE.INP - копию файла atom.xls с измененными на 0.25 по абсолютной величине значениями зарядов на атомах O1P и O2P;

7. Выполнить программу elefull и утилиту cyl для каждого фрагмента ДНК, в результате чего в каждом подкаталоге появятся файлы U.dat со значениями электростатических потенциалов и fort.7 с данными о поверхности цилиндра, на который проецируется электроститический потенциал;

8. Получить изображение результатов вычислений для каждого фрагмента в программе MOLMOL с помощью одного из макросов showcyl.mac, cyl_-35.mac или cyl_-75.mac.

Приложение F. Управление массовыми вычислениями в сети распределенных вычислений с помощью скриптов

Описанные ниже скрипты (сценарии) позволяют запустить процесс вычисления карт молекулярных поверхностей большого числа фрагментов в пакетном режиме. Эти модули представляют собой набор команд PBS для выбора фермы кластера Linux и узла в его пределах, управления задачей, контроля ее выполнения и пересылки пользователю результатов в указанный каталог или уведомления по электронной почте. Приведенные сценарии реализованы в распределенной файловой системе AFS (версия 3.0). Согласно принципу функционирования AFS, дисковое пространство пользователя разделяется на три каталога: home – наиболее защищенный от несанкционированного доступа и отказов оборудования, предназначенный для долговременного хранения данных, scratch – для хранения больших объемов данных и tmp(src) – для кратковременного хранения, например, во время работы программы в распределенной сети.

В результате обработки создаются:

- Файл с расширением CHT – сохраненная карта биополимеров и их комплексов;
- Файл с расширением SAV – файл в pdb-формате, необходимый для визуализации карты на персональном компьютере;
- Файл INFO.TXT, содержащий краткую статистику результатов вычислений: тип карты молекулярной поверхности, продолжительность вычислений в сети, дату и время запуска программы, имя каталога результатов.

Для визуализации и разметки функциональных карт используют соответствующие версии программ для ДНК, РНК и белков [29, 30].

Функции первого сценария (скрипта):

1. Задание всех необходимых переменных.
2. Задание расположения pdb-файлов, файла эмуляции интерфейса loadpar.txt, файла консольной программы.
3. Создание структуры каталогов, соответствующих именам pdb-файлов во временном каталоге:

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

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

   /[temp dir]/[users]/[username]/[helix_dnarna] (для фрагментов ДНК/РНК).
4. Копирование требуемого файла loadpar.txt, файла программы и второго сценария (скрипта) в соответствующие каталоги.
5. Выполнение второго сценария с параметрами для каждого из pdb– файлов из временного каталога командой qsub системы PBS.
6. Создание иерархии каталогов для записи результатов в домашний каталог пользователя HOME.

Функции второго сценария (скрипта):

1. Копирование выполнимой программы, pdf-файла данных и модуля эмуляции интерфейса из временного каталога в cluster.farm.

2. Запуск консольной программы с последующей записью в файл INFO.TXT информации о дате создания и времени вычислений для заданного фрагмента.

3. Копирование результатов в указанный каталог основного пользовательского каталога HOME на отдаленном сервере.

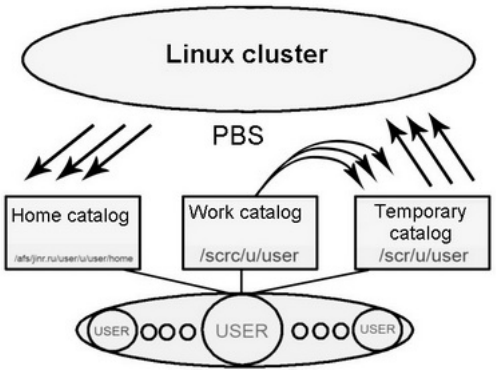

Рис. 7. Блок-схема обмена данными в каталогах файловой системы AFS.

На рисунке 7 представлена схема всего вычислительного процесса, реализованного с использованием приведенных файлов сценариев. Результаты сохраняются в каталоге home – в созданном для этого отдельном каталоге со следующей структурой:

1. [catalog_results];
2. [catalog_data_type (глобулярные белки, спиральные белки, ДНК/РНК)];
3. [catalog_creation_data];
4. [catalog_pdb-имя-файла (содержит полученные результаты)];
5. файлы *.CHT, *.SAV и INFO.TXT.

СПИСОК ЛИТЕРАТУРЫ

1. Romberg R.D. Eukaryotic transcriptional control. *Trends Cell Biol.* 1999. V. 9. P. M46–M49.
2. Coleman R.A., Pugh B.F. Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA. *J. Biol. Chem.* 1995. V. 270. P. 13850–13859.
3. Parsons J.D. Improved tools for DNA comparison and clustering. *Comput. Appl. Biosci.* 1995. V. 11. P. 603–613.
4. Guan X, Du L. Domain identification by clustering sequence alignments. *Bioinformatics.* 1998. V. 14. P. 783–788.
5. Cornell W.D., Cieplak P., Bayly C.I., Gould I.R., Merz K.M., Ferguson D.M., Kollman P.A. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* 1995. V. 117. P. 5179–5197.
6. Федоренко Р.П. Релаксационный метод решения разностных эллиптических уравнений. *Ж. вычисл. матем. и матем. физ.* 1961. Т. 1. № 5. С. 922–927.
7. Федоренко Р.П. Итерационные методы решения разностных эллиптических уравнений. *УМН*. 1973. V. 28. № 2(170). С. 121–182.
8. Hackbusch W. *Multi-grid methods and applications.* Springer–Verlag, Berlin – Heidelberg – New York – Tokyo, 1985. 377 p. (Springer Series in Computational Mathematics. V. 4).
9. Fedoseyev A.I., Lazarev P.I., Sivozhelezov V.S., Chernyaev E.V., Petrenko I.I., Purtov S.V. Mathematical modelling of 3D protein molecule potential in nonlinear media. In: *Abstracts of the International Conference "Physique en Herbe'92"*. Marseille, 1992. P. 233–236.
10. Koradi R., Billeter M., Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* 1996. V. 14. P. 51–55.
11. Флетчер К. *Численные методы на основе метода Галёркина.* М.: Мир, 1988. 352 с.
12. Визильтер Ю., Желтов С., Князь В., Ходарев А., Моржин А. *Обработка и анализ изображений с примерами на LabVIEW IMAQ Vision.* М.: ДМК Пресс, 2007. 465 с.
13. Chirgadze Y.N., Zheltukhin E.I., Polozov R.V., Sivozhelezov V.S., Ivanov V.V. Binding regularities in complexes of transcription factors with operator DNA: homeodomain family. *Journal of Biomolecular Structure and Dynamics*. 2009. V. 26. No. 6. P. 687–700.
14. Chirgadze Y.N., Sivozhelezov V.S., Polozov R.V., Stepanenko V.A., Ivanov V.V. Recognition rules for binding of homeodomains to operator DNA. *Journal of Biomolecular Structure and Dynamics*. 2012. V. 29. No. 4. P. 715–731.
15. Polozov R.V., Sivozhelezov V.S., Chirgadze Y.N., Ivanov V.V. Recognition rules for binding of Zn-Cys2His2 transcription factors to operator DNA. *Journal of Biomolecular Structure and Dynamics*. 2015. V. 33. No. 2. P. 253–266.
16. Kornberg R.D. Eukaryotic transcriptional control. *Trends Cell Biol.* 1999. V. 9. P. M46–M49.
17. Ozoline O.N., Deev A.A., Arkhipova M.V. Noncanonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase. *Nucleic Acids Research*. 1997. V. 25. No. 23. P. 4703–4709.
18. Akishina T.P., Ivanov V.V., Stepanenko V.A. Massive calculations of electrostatic potentials and structure maps of biopolymers in a distributed computing environment. In: *Communication of JINR*. E11-2013-137. Dubna, 2013. 12 p.
19. HyperChem Professional 8.0. URL: http://http://www.hyper.com/?tabid=360 (дата обращения 25.10.2017).
20. Ivanciuc O. HyperChem release 4.5 for windows. *Journal of Chemical Information and Computer Sciences*. 1996. V. 36. No. 3. P. 612–614.
21. Porter W. *Introduction to Map Projections*. N.Y.: Marcel Dekker, 1979.
22. Мещеряков Г.А. *Теоретические основы математической картографии*. М.: Недра, 1968.
23. Лаврентьев М.А., Шабат Б.В. *Методы теории функций комплексного переменного*.

М.: Наука, 1973.
24. ГалактионовВ.В., Голоскокова Т.М., Громова Н.И., Гусев А.В., Мицын В.В., Мусульманбеков Ж.Ж., Некрасова И.К., Позе В.Д., Сергеев А.В., Тихоненко Е.А. *Руководство для пользователей Linux кластера ЛИТ ОИЯИ*. 2004. URL: http://lit.jinr.ru/ccic/usersguide/ (дата обращения 25.10.2017).
25. Scientific Linux. URL: http://www.scientificlinux.org/ (дата обращения 25.10.2017).
26. Галактионов В.В., Голоскокова Т.М., Громова Н.И., Гусев А.В., Мицын В.В., Мусульманбеков Ж.Ж., Некрасова И.К., Позе В.Д., Сергеев А.В., Тихоненко Е.А. Файловая система AFS. В: Руководство для пользователей Linux кластера ЛИТ ОИЯИ. 2004. URL: http://lit.jinr.ru/ccic/usersguide/index.php?link=3_ (дата обращения 25.10.2017).
27. Галактионов В.В., Голоскокова Т.М., Громова Н.И., Гусев А.В., Мицын В.В., Мусульманбеков Ж.Ж., Некрасова И.К., Позе В.Д., Сергеев А.В., Тихоненко Е.А. 2.4 Пакетная обработка счетных задач. В: *Руководство для пользователей Linux кластера ЛИТ ОИЯИ*. 2004. URL: http://lit.jinr.ru/ccic/usersguide/index.php?link=2_#2.4_ (дата обращения 25.10.2017).
28. *Lazarus. The professional Free Pascal RAD IDE.* URL: http://www.lazarus-ide.org/ (дата обращения 25.10.2017).
29. Афанасьев О.А., Зрелов П.В., Иванов В.В., Полозов Р.В., Сивожелезов В.С., Степаненко В.А., Чиргадзе Ю.Н. *Комплекс программ картографирования и исследования белков и нуклеиновых кислот.* Сообщения ОИЯИ. Р10–2011–108. Дубна, 2011. 35 с.
30. Бедняков И.В., Зрелов П.В., Иванов В.В., Полозов Р.В., Сивожелезов В.С., Степаненко В.А., Чиргадзе Ю.Н. Картографирование структур белков и нуклеиновых кислот. *Письма в ЭЧАЯ*. 2013. Т. 10. № 5(182). С. 744–755.
31. Flores S.C., Bernauer J., Shin S., Zhou R., Huang X. Multiscale modeling of macromolecular biosystems. *Briefings in Bioinformatics*. 2012. V. 13. No. 4. P. 395–405.

Рукопись поступила в редакцию 01.10.2017.
Дата опубликования 30.10.2017.