АНАЛИЗ АЛГОРИТМОВ ПРОГНОЗИРОВАНИЯ ТРЕТИЧНОЙ СТРУКТУРЫ ПРОТЕИНА НА БАЗЕ МЕТОДА ОПТИМИЗАЦИИ МУРАВЬИНЫМИ КОЛОНИЯМИ

Леонид Гуляницкий, Виталина Рудык

Аннотация: *Исследуются алгоритмы метода оптимизации муравьиных колоний для прогнозирования третичной структуры протеинов. Предлагается новый алгоритм на базе этого метода, анализируются его преимущества, проводятся экспериментальные исследования.*

Ключевые слова: *комбинаторная оптимизация, прогнозирование третичной структуры протеина, методы оптимизации муравьиными колониями, эвристики, NP-сложность.*

ACM Classification Keywords*: J.3 LIFE AND MEDICAL SCIENCES*

Введение

При исследовании различных процессов для избегания сложных, дорогостоящих и затратных по времени экспериментов принято использовать инструменты математического моделирования. Широкое применение нашла HP-модель Дилла [Dill, Bromberg, Yue, Fiebig, Yee, Thomas, Chan, 1995], которая хотя и упрощенно отображает процесс формирования трехмерной структуры белка, однако учитывает основные влияющие силы. Молекула белка состоит из аминокислотных остатков, соединенных в цепь пептидными связями. Последовательность остатков в цепи является первичной структурой белка. Третичная структура – это форма в пространстве, которую принимает молекула. Принято считать, что она однозначно определяется первичной структурой.

Для моделирования пространственной формы молекулы аминокислотные остатки располагаются в узлах некоторой дискретной решетки, соседние в пептидной цепи остатки – в соседних узлах (условие связности). В каждом узле может находиться не более одного остатка (условие отсутствия самопересечений). Экспериментально исследовано, что основной силой, формирующей третичную структуру белка (помимо пептидных связей), являются взаимодействия между гидрофобными аминокислотными остатками. В моделировании это используется так: все остатки делятся на два класса – гидрофобные и полярные, между гидрофобными остатками, расположенными в соседних узлах решетки, но не соседними в первичной последовательности, возникает HH-связь. Свободной энергией структуры принимается количество HH-связей в ней со знаком минус, и согласно термодинамической гипотезе считается, что молекула принимает ту форму, в которой достигается минимум ее свободной энергии. Так задача прогнозирования третичной структуры белка сводится к задаче комбинаторной оптимизации.

Метод оптимизации муравьиными колониями. Предпосылки

Метод оптимизации муравьиными колониями (ОМК) [Dorigo, Stützle, 2004] – эвристический популяционный алгоритм, который применяется для решения широкой круга задач оптимизации. В его основе лежит моделирование поведения муравьев при поиске кратчайших путей с использованием обмена информации между особями колонии через феромонные следы.

На базе метода ОМК разработано несколько алгоритмов для решения задачи прогнозирования третичной структуры молекул протеинов. Рассмотрим подробнее самые известные из них [Shmygelska, Hoos, 2005], [Chu, Till, Zomaya, 2005], [Fidanova, Lirkov, 2008]. В этих исследованиях строятся алгоритмы на базе метода ОМК для трехмерной кубической решетки. Для представления третичной структуры молекулы

используется относительное кодирование, где положение каждого следующего аминокислотного остатка задается относительно предыдущего [Shmygelska, Hoos, 2005].

Алгоритм метода ОМК обычно состоит из трех этапов. На ***этапе построения решений*** каждый муравей строит допустимую структуру молекулы, основываясь на феромонных следах, которые содержат в себе полезную информацию, собранную на предыдущих итерациях алгоритма. В [Shmygelska, Hoos, 2005], [Chu, Till, Zomaya, 2005] построение начинается со случайно выбранной позиции и продолжается в двух направлениях, в [Fidanova, Lirkov, 2008] молекула строится последовательно, начиная с первого остатка. Позиция аминокислоты выбирается случайным образом среди допустимых с вероятностью

$$p_{i,d} = \frac{[\tau_{i,d}]^{\alpha}[\eta_{i,d}]^{\beta}}{\sum_{l \in D}[\tau_{i,l}]^{\alpha}[\eta_{j,l}]^{\beta}}, \quad (1)$$

где D – множество возможных элементов кода, которое отражает направления свободных соседей текущего узла в заданной решетке, $\tau_{i,d}$ – элемент феромонной матрицы, соответствующий направлению d для позиции i, $\eta_{i,d}$ – эвристическая оценка, которая завит от уже построенной части структуры, α и β – параметры алгоритма, задающие степень влияния феромонов и эвристической информации на построение решения. В [Shmygelska, Hoos, 2005] эвристическая информация считается по формуле

$$\eta_{i,d} = e^{-\gamma h_{i,d}},$$

где $h_{i,d}$ – количество новых HH-контактов, которые возникают при позиционировании новой аминокислоты по направлению d. В [Chu, Till, Zomaya, 2005] принято $\eta_{i,d} = h_{i,d} + 1$, а в [Fidanova, Lirkov, 2008] полагается $\eta_{i,d} = h_{i,d}$, таким образом для полярных остатков $\eta_{i,d} = 0$, что значит, что на их позиционирование значения феромонной матрицы не влияют.

При построении структуры может возникнуть ситуация, когда все узлы, соседние с положением крайней в уже построенном фрагменте аминокислоты, заняты – в таком случае (если остаток не первый и не последний в первичной последовательности) используется откат. В [Shmygelska, Hoos, 2005] расформировывается половина построенной структуры и формируется заново, при этом первый (т.е. последний из расформированных) остатков располагается в направлении отличном от того, в котором он был в структуре, которая зашла в тупик. В [Chu, Till, Zomaya, 2005] процедура отката производится только на 1 шаг, в [Fidanova, Lirkov, 2008] – на некоторое фиксированное количество шагов.

На ***этапе локального поиска*** происходит оптимизация вариантов, построенных на первом этапе алгоритма. В [Shmygelska, Hoos, 2005] в процедуре локального поиска используются дальновидные сдвиги. При таком действии случайным образом выбирается аминокислотный остаток и меняется его направление. Для всех дальнейших остатков принимается решение не менять их направление, если оно допустимо, с некоторой вероятностью p. Если же направление меняется, то оно выбирается случайным образом с вероятностью, пропорциональной эвристической оценке $\eta_{i,d}$. В [Chu, Till, Zomaya, 2005] в локальном поиске происходят только локальные изменения – случайным образом меняется направление случайного остатка. В обоих алгоритмах новая структура принимается, если ее энергия меньше энергии начальной структуры и процедура останавливается после того, как некоторое количество шагов не привело к улучшениям. Схема в [Shmygelska, Hoos, 2005] требует больших затрат по времени, на каждой итерации она применяется только к определенному проценту наиболее оптимальных структур,

построенных на первом этапе. Алгоритм, описанный в [Fidanova, Lirkov, 2008] не использует локального поиска.

На ***этапе обновления феромонных следов*** происходит обновление феромонной матрицы с учетом энергии структур, полученных на первых двух этапах алгоритма. Этот этап обычно состоит из двух процедур – испарение феромона и отложение феромона на субоптимальных решениях. В [Shmygelska, Hoos, 2005] и [Chu, Till, Zomaya, 2005] эти этапы проходят последовательно. Процесс испарения задается формулой

$$\tau_{i,d} = (1-\rho)\tau_{i,d}, \tag{2}$$

где параметр ρ характеризует, какая доля информации, собранной на предыдущих этапах, сохраняется. Феромонные пути усиливают только муравьи с низкой энергией по формуле

$$\tau_{i,d} = \tau_{i,d} + \Delta_{i,d,c}, \tag{3}$$

где $\Delta_{i,d,c}$ – относительное качество структуры c, если остаток i расположен по направлению d, или 0 в противоположном случае.

В [Fidanova, Lirkov, 2008] обновления феромона происходит локально и глобально. Процесс локального обновления реализует испарение феромона и интегрирован в этап построения решений. После построения каждой структуры динамически изменяется количество феромона на всех участках, которые в нее входят, по формуле

$$\tau_{i,d} = (1-\rho)\tau_{i,d} + \rho\tau_0,$$

где τ_0 – некий параметр. Таким путем достигается более широкий поиск в окрестностях предыдущего оптимального решения. Феромонные пути в [Fidanova, Lirkov, 2008] усиливаются по формуле (3) только для одного лучшего решения, полученного на данной итерации.

Разработанный алгоритм ОМК для задачи прогнозирования третичной структуры протеина

Разработанный нами алгоритм строит структуру в трехмерной треугольной решетке. В таком выборе есть несколько преимуществ. Во-первых, это отсутствие известной проблемы парности, которая проявляется в двумерной квадратной и трехмерной кубической решетках. Ее суть состоит в том, что в силу особенностей решетки НН-контакты могут возникнуть только между остатками с четным и нечетным номерами в первичной последовательности, что противоречит естественным представлениям. Второе преимущество – в том, что у каждого узла большее количество соседей по сравнению с кубической решеткой, что выливается в более широкое множество возможных структур для фиксированной первичной последовательности. Недостатком является то, что по сравнению с квадратной и кубической эта решетка менее исследована, что ограничивает возможность экспериментального сравнения разработанного алгоритма с известными. Исследуются варианты с абсолютным и относительным кодированием, построение этих кодировок для трехмерной кубической решетки подробно описано в [Рудык, 2011].

Разработанный алгоритм реализует стандартные этапы методов ОМК, его схема приведена на рис. 1.

Процедура *ДопустимоеРешение*(Φ) на основе феромонной матрицы поэлементно, начиная с первого остатка, генерирует новую структуру с вероятностью, которая задается формулой (1). Эвристическая оценка $\eta_{i,d}$ вычисляется по правилу

$$\eta_{i,d} = \begin{cases} n_P\,\eta_{HP} + n_H\,\eta_{HH} + (n - n_P - n_H)\,\eta_{H0}, & s_{i+1} = H, \\ n_P\,\eta_{PP} + n_H\,\eta_{PH} + (n - n_P - n_H)\,\eta_{P0}, & s_{i+1} = P. \end{cases}$$

Тут s_{i+1} – тип аминокислотного остатка, который позиционируется на данном шаге, n, n_P и n_H – количество узлов, соседних к рассматриваемому, соседних узлов, занятых полярными и гидрофобными остатками соответственно, а η_{HH}, η_{HP}, η_{H0}, η_{PP}, η_{PH}, η_{P0} – параметры, удовлетворяющие условиям

$$0 \le \eta_{HP} \le \eta_{H0} < \eta_{HH},$$
$$0 \le \eta_{PH} \le \eta_{P0} \le \eta_{PP}.$$

```
procedure ACO()
    fold_rec := null;
    ИнициализироватьФеромоннуюМатрицу(Φ);
    while (!УсловиеЗавершения()) do
        for i = 1,..., N_P do
            fold_i :=ДопустимоеРешение(Φ);
        end for;
        Отсортировать(fold_i);
        for i = 1,..., N_LS do
            fold_i :=Локальный поиск(fold_i);
        end for;
        fold_rec :=ОптимальноеЗначение(fold_rec, fold_1,..., fold_N_LS);
        ИспаритьФеромоннуюМатрицу(Φ, v_i);
        for i = 1,..., N_P do
            ОбновитьФеромоннуюМатрицу(Φ, v_i);
    end while;
    return fold_rec;
end procedure.
```

Рис. 1. Схема алгоритма ОМК для задачи определения структуры протеина

Такое определение эвристической оценки обобщает схему из [Fidanova, Lirkov, 2008] при

$$\eta_{HP} = \eta_{H0} = \eta_{PP} = \eta_{PH} = \eta_{P0} = \tfrac{1}{n},$$
$$\eta_{HH} = 1 + \tfrac{1}{n},$$

и схему из [Chu, Till, Zomaya, 2005] при

$$\eta_{HP} = \eta_{H0} = \eta_{PP} = \eta_{PH} = \eta_{P0} = 0,$$
$$\eta_{HH} = 1.$$

Если остаток расположить не удалось (все возможные позиции заняты), производится откат на один шаг назад.

В качестве локального поиска используется детерминированная процедура, в которой итеративно производится переход к структуре, отличающейся на один элемент, с меньшей энергией до тех пор, пока такие существуют. Локальный поиск совершается для лучших N_{LS} решений.

Испарения феромонов проходит по формуле (2), а для обновления феромонных следов рассматриваются две схемы. Первая – аналогична [Shmygelska, Hoos, 2005], [Chu, Till, Zomaya, 2005], [Fidanova, Lirkov, 2008] и задается формулой

$$\tau_{i,d} = \tau_{i,d} + \Delta_{i,d,c}{}^{\gamma}, \tag{4}$$

где γ – параметр алгоритма. Вторая разработана с учетом того, что необходимо усиливать только те феромонные пути, которые влияют на энергию заданной молекулы. Сила $\varphi_{i,d,c}$ определенного направления в структуре c определяется как количество связей, в которых это направление задействовано (Рис. 2). Белые вершины обозначают гидрофобные аминокислотные остатки, а пунктиры – HH-связи. Феромонная матрица обновляется по формуле

$$\tau_{i,d} = \tau_{i,d} + \varphi_{i,d,c}{}^{\gamma}. \tag{5}$$

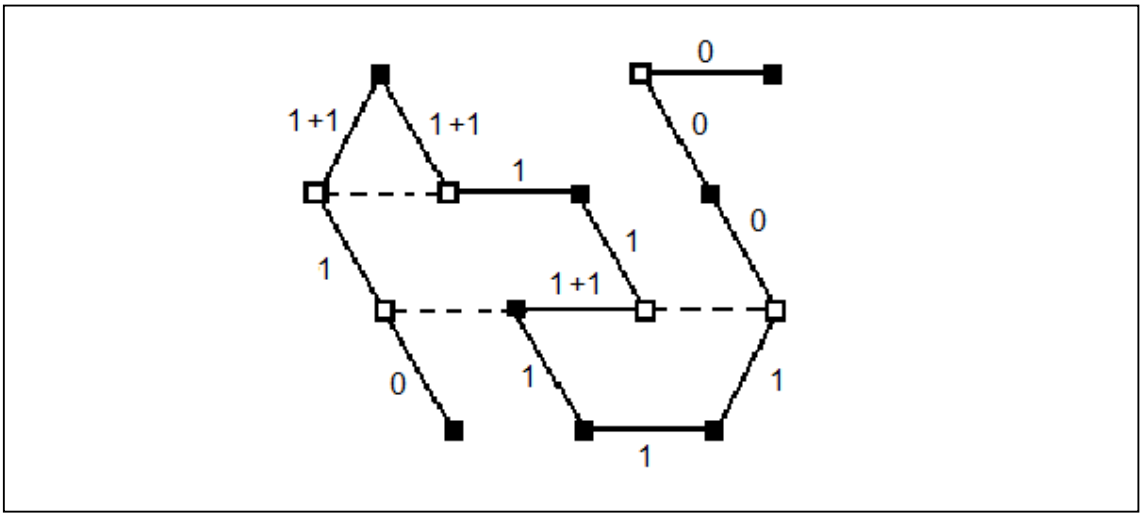

Рис. 2. Оценка силы направлений в структуре.

Такая схема позволяет более точно учитывать приемлемость той или иной части структуры молекулы. Одна из целей проведения вычислительного эксперимента – подтвердить это на практике.

В описанном алгоритме метода ОМК можно выделить следующие параметры:

- enc – используемый тип кодировки структуры (абсолютная или относительная);
- N_P – количество элементов в популяции;
- τ_0 – начальное значение элементов феромонной матрицы;
- η_{HH}, η_{HP}, η_{H0}, η_{PP}, η_{PH}, η_{P0} – параметры для эвристической оценки;
- α, β – степень влияния феромонных путей и эвристических оценок на строящуюся структуру;
- N_{LS} – количество решений, для которых применяется процедура локального поиска;
- ρ – характеристика испарения феромона;

- *UpdType* – вариант обновления феромонной матрицы;

- γ – параметр обновления феромонных путей;

- K_{stop} – количество итераций, на которых не должно измениться оптимальное решение, чтобы алгоритм остановился.

Вычислительный эксперимент

Алгоритмам метода ОМК свойственна чувствительность к параметрам, поскольку при несбалансированном их выборе возникает или предварительная сходимость, или в противоположном случае феромонная матрица перестает существенно влиять на процесс решения и алгоритм превращается в случайный поиск. К тому же, особенностью расссматриваемой задачи является то, что субоптимальные решения представляют собой компактные структуры (ведь внутри них образуются связи), и поэтому в их окрестностях содержится много недопустимых вариантов. Для алгоритмов ОМК это значит, что во многих случаях комбинации нескольких субоптимальных решений будут недопустимыми. Целью вычислительного эксперимента был поиск значений варьируемых параметров, при которых алгоритм становится максимально эффективным.

Для проведения эксперимента из базы данных структур протеинов [http://www.rcsb.org] было выбрано 8 структур с уникальными идентификаторами 7RXN, 3VUB, 4I1B, 3STR, 3SBZ, 3RLG, 3TEC, 3TX9 с длиной от 52 до 400 аминокислот. Вычисления проводились в одном потоке на кластере вычислительного комплекса СКИТ-3 ИК им. В.М.Глушкова НАН Украины [https://icybcluster.org.ua]. Для каждой из 8 задач производилось 10 перезапусков алгоритма при каждом из приведенных ниже наборов параметров.

Поскольку эксперименты проводились для задач с широким интервалом значений длины входящей последовательности, то для того, чтобы параметры в равной степени влияли на алгоритм, было принято решения нормировать количество феромона на каждом участке и эвристические оценки в диапазоне от 0 до 1.

В вычислительном эксперименте исследовались параметры α, β, *UpdType* и γ. Остальные параметры во всех приведенных вычислениях были фиксированы и принимали следующие значения:

- $enc = relative$ (относительная);
- $\tau_0 = 0.2$;
- $\rho = 0.05$;
- $N_P = 30$;
- $N_{LS} = 0$;
- $K_{stop} = 100$;
- $\eta_{HH} = 0.6$;
- $\eta_{HP} = 0.15$;
- $\eta_{H0} = 0.25$;
- $\eta_{PP} = 0.4$;
- $\eta_{PH} = 0.2$;
- $\eta_{P0} = 0.4$.

На первом этапе эксперимента использовалось обновление феромонной матрицы по формуле (4). Значениями параметров (α, β, γ) были выбраны следующие комбинации: (1; 1; 1), (1; 1; 0.5), (1; 1; 2), (0; 1; 1), (0.5; 1; 1), (1; 0.5; 1). Результаты вычислений показаны на рис. 3. По горизонтальной оси – размер входящей последовательности, по вертикальной – среднее (по 10 запускам) значение оптимальной энергии, найденное алгоритмом при соответствующих значениях параметров.

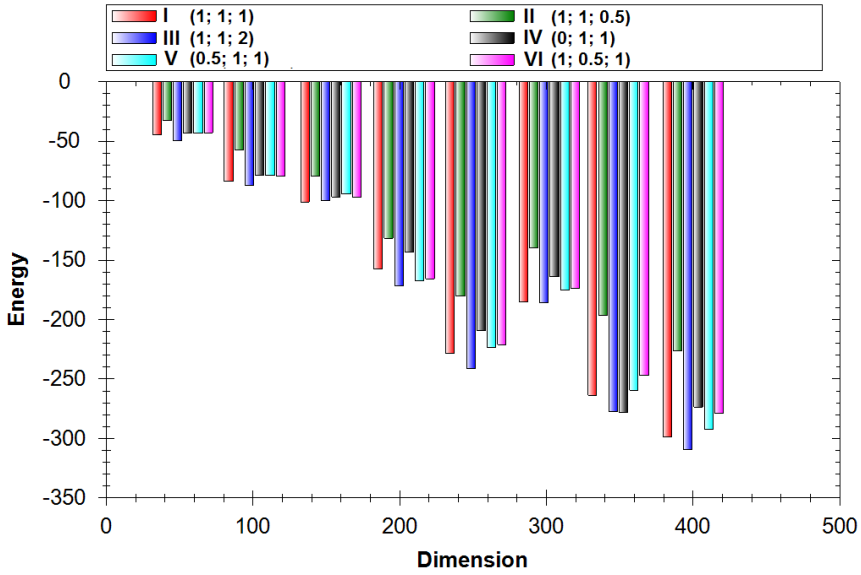

Рис. 3. Результаты первого этапа вычислительного эксперимента

На втором этапе феромонной матрица обновлялась по формуле (5). Значения параметров (α, β, γ) – (1; 1; 1), (1; 1; 2), (0; 1; 1), (0.5; 1; 1) (Рис. 4).

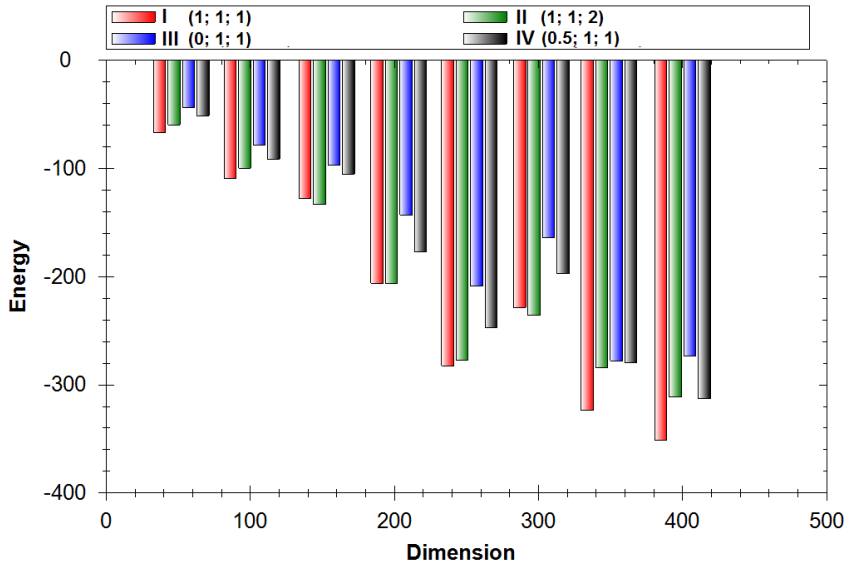

Рис. 4. Результаты второго этапа вычислительного эксперимента

Выводы и направления дальшейших исследований

Вычислительный эксперимент показал, что алгоритм, разработанный на базе метода ОМК, в среднем находит решение с энергией в 150% лучше начальной. Варьируя параметры алгоритма, удается достичь более высокой эффективности. При прочих одинаковых параметрах обновление феромонной матрицы разработанным новым методом более эффективно, чем ее обновление методом, аналогичным тому, который встречается в литературе.

Открытыми остаются следующие направления исследований:

- подбор оптимального набора значений параметров алгоритма, при котором он максимально эффективен;

- реализация разработанного алгоритма для кубической решетки с целью проведения сравнительного анализа с известными алгоритмами метода ОМК на эталонных задачах;

- проведение сравнительного анализа разработанного алгоритма с известными алгоритмами, разработанными для трехмерной треугольной решетки.

Благодарности

Работа опубликована при финансовой поддержке проекта **ITHEA XXI** Института информационных теорий и приложений FOI ITHEA Болгария www.ithea.org и Ассоциации создателей и пользователей интеллектуальных систем ADUIS Украина www.aduis.com.ua.

Список литературы

[Dill, Bromberg, Yue, Fiebig, Yee, Thomas, Chan, 1995] K.Dill, S.Bromberg, K.Yue, K.M.Fiebig, D.Yee, P.Thomas, H.Chan. Principles of protein folding - a perspective from simple exact models // Protein Science. – 1995. – 4. – P. 561– 602.

[Dorigo, Stützle, 2004] M.Dorigo M.,T. Stützle. Ant Colony Optimization. – Cambridge: MIT Press, MA, 2004. – 348 p.

[Shmygelska, Hoos, 2005] A. Shmygelska, H. Hoos. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem // BMC Bioinformatics. – 2005. – 6(30). – P.30–52.

[Chu, Till, Zomaya, 2005] D. Chu, M. Till, A. Zomaya. Parallel Ant Colony Optimization for 3D Protein Structure Prediction using the HP Lattice Model // 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05). – 2005. – 7. – P.193-200.

[Fidanova, Lirkov, 2008] S. Fidanova, I. Lirkov. Ant Colony System Approach for Protein Folding // Int. Conf. Multiconference on Computer Science and Information Technology. – 2008. – P.887–891.

[Рудык, 2011] В. Рудык. Представление структуры белка в трехмерных дискретных решетках произвольного типа // Теорія оптимальних рішень. – 2011. – №10. – С. 38–47.

[http://www.rcsb.org] http://www.rcsb.org/pdb/home/home.do Protein Data Bank.

[https://icybcluster.org.ua] https://icybcluster.org.ua/index.php?lang_id=1&menu_id=5 Сайт суперкомпьютера ИК НАН Украины.

Об авторах

Леонид Гуляницкий *(Hulianytskyi) – д.т.н., заведующий отделом Института кибернетики им. В.М.Глушкова НАН Украины, пр-т Глушкова, 40, Киев, 03680, Украина. e-mail: leonhul.icyb@gmail.com*

Виталина Рудык *(Rudyk) – аспинтантка, младший научный сотрудник Института кибернетики им. В.М.Глушкова НАН Украины, пр-т Глушкова, 40, Киев, 03680, Украина. e-mail: vitalina.rudyk@gmail.com*