21. *Якубке Х.-Д., Ешкайт Х.* Аминокислоты. Пептиды. Белки.— М. : Мир, 1985.— 455 с.
22. *Dayhoff M. O.* Atlas of protein sequence and structure.— Washington : Nat. Biomed. Res. Found, 1978.— V. 5.— 414 p.
23. *Eppstein D. A., March Y. V., Schreiber A. B.* Epidermal growth factor receptor occupancy inhibits vaccinia virus infection // Nature.— 1985.— **318**, N 6047.— P. 663—665.
24. *Meng M., Hogenkamp H. P.* Purification, characterization and amino acid sequence of thioredoxin from *Corynebacterium nephridii* // J. Biol. Chem.— 1981.— **256**, N 17.— P. 9174—9178.
25. *Calcitonin* messenger RNA encodes multiple polypeptides in a single precursor / I. W. Jacob, R. H. Goodman, W. W. Chin et al. // Science.— 1981.— **213**, N 4506.— P. 457—459.
26. *Pierschbacher M. O., Ruoslahti E.* Cell attachment activity of fibronectin can be dublicated by small synthetic fragments of molecule // Nature.— 1984.— **309**, N 5963.— P. 30—33.
27. *The protein* data bank: a computer-based archival file for macromolecular structures / F. C. Bernstein, T. F. Koetzle, G. J. B. Willams et al. // J. Mol. Biol.— 1977.— **112**, N 3.— P. 535—542.
28. *Janin J., Chotia C.* Stability and specificity of protein-protein interactions: the case of the trypsin-trypsin inhibitor complexes / Ibid.—1976.— **100**, N 2.— P. 197—211.
29. *Fox B. S., Walsh C. T.* Mercuric reductase: homology to glutathione reductase and lipoamid degydrogenase. Iodoacetamide alkylation and sequence of the active site peptide // Biochemistry.— 1983.— **22**, N 17.— P. 4082—4088.
30. *Schlegel R., Wade M.* Biologically active peptides of the vesicular stomatitis virus glycoprotein // J. Virol.— 1985.— **53**, N 1.— P. 319—323.
31. *Griffith M. J., Noyes C. M., Church F. C.* Reactive site peptide structural similarity between heparin cofactor II and antithrombin III // J. Biol. Chem.— 1985.— **260**, N 4.— P. 2218—2225.
32. *Traut T. W.* Do exons code for structural or functional units in proteins? // Proc. Nat. Acad. Sci. USA.— 1988.— **85**, N 9.— P. 2944—2948.
33. *Hantgan R. R.* Localization of the domains of fibrin involved in binding to platelets // Biochim. et biophys. acta.— 1988.— **968**, N 1.— P. 36—44.

ВНИИ молекуляр. биологии НПО «Вектор» минмедбиопрома СССР, пос. Кольцово, Новосиб. обл.

Получено 11.12.89

УДК 577.112

Е. В. Кунин, К. М. Чумаков, А. Е. Горбаленя

МЕТОД ПОИСКА СТРУКТУРНЫХ МОТИВОВ В АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ. ПРОГРАММА «SITE» ПАКЕТА «GENBEE»

Предложен метод поиска структурных мотивов в аминокислотных последовательностях, основанный на построении частотного профиля группы выравненных фрагментов последовательностей. На примере сканирования банка аминокислотных последовательностей мотивом, характерным для широкого класса NTP-связывающих белков, рассмотрена работа программы «SITE», написанной на основе предложенного алгоритма. Продемонстрированы преимущества предложенного подхода по сравнению со стандартными программами поиска паттернов в отношении полноты и избирательности извлечения из банка последовательностей, содержащих участки, сходные с рассматриваемым мотивом. Представлена предположительная идентификация NTP-связывающих центров в нескольких белках, где они ранее не были обнаружены. Обсуждается применение разработанного алгоритма для классификации банков аминокислотных последовательностей.

Введение. Один из основных эвристических приемов, используемых при теоретическом анализе структуры и функций биополимеров, — поиск в аминокислотных (нуклеотидных) последовательностях так называемых структурных мотивов и паттернов. В данной работе мы будем рассматривать только мотивы (паттерны), заданные на уровне первичной структуры. Мотив можно определить как относительно короткую аминокислотную последовательность (о нуклеотидных последовательно-

стях мы здесь говорить не будем), характеризующуюся определенной степенью консервативности в пределах достаточно обширной группы родственных белков. По определению, в состав мотива не входят пробелы. Группа разнесенных в границах одной полипептидной цепи мотивов образует паттерн; при формировании паттерна учитывается также расстояние между отдельными мотивами [1, 2]. В соответствии с представлениями современной теории молекулярной эволюции [3] естественно ожидать, что структурные мотивы (паттерны) должны ассоциироваться с определенными функциями, например, могут входить в сос-

Таблица 1

Примеры структурных мотивов, коррелирующих с определенными функциями в белковых молекулах

Examples of structural motifs associated with distinct functions in proteins

Мотив	Функция
G(D)SG(G)	Активный центр химотрипсин-подобных сериновых протеаз
GxCW	Активный центр тиоловых протеаз
[гидрофобные] остатки *D(T)D* [гидрофобные] остатки	Активный центр РНК-зависимых РНК-полимераз, ДНК-зависимых ДНК-полимераз и обратных транскриптаз
NADFDGD[QE]	Активный центр ДНК-зависимых РНК-полимераз
[RK] G [FY][GA]FVx[FY]	РНК-связывающий домен
$[FY]xCx_{2-4}Cx_3Fx_5Lx_2Hx_{3-4}H$	ДНК-связывающий домен типа «finger»

Примечание. Аминокислотные остатки, указанные в круглых скобках, встречаются во многих, но не во всех последовательностях соответствующего класса; в квадратных скобках указаны альтернативные остатки. В каждом случае приведенные последовательности образуют только часть активного центра; *x*—любой аминокислотный остаток.

тав активных центров ферментов. В ряде случаев это действительно подтверждается анализом конкретных экспериментальных данных (табл. 1).

Описано множество алгоритмов и программ поиска мотивов в последовательностях (см. обзор [2]). По существу это программы поиска регулярных выражений (в том числе вырожденных) в тексте. Так, например, для поиска одного из мотивов связывания ATP и GTP может быть использовано следующее выражение: $[GA]x_4GK[ST]$ (обозначения — как в табл. 1). Однако по мере роста массива известных аминокислотных последовательностей, из анализа которых выводятся формулы мотивов, подобные простейшие алгоритмы становятся все менее полезными в силу неизбежного «размывания» последних. Такое размывание можно продемонстрировать опять-таки на примере «эволюции» паттерна связывания ATP или GTP на протяжении последних 7 лет ([4, 5]; рис. 1). В результате «вырождения» мотивов складывается положение, когда использование достаточно «жесткого» мотива ведет к потере чувствительности, т. е. к неполному извлечению родственных последовательностей из базы данных, а использование «размытого» мотива — к утрате избирательности, т. е. к извлечению неродственных последовательностей. В такой ситуации вывод формул мотивов в значительной мере утрачивает смысл, и под мотивом приходится понимать собственно массив выравненных фрагментов последовательностей. Продуктивное использование таких мотивов для предсказания функций белков и выявления эволюционных связей требует применения более тонких методов, позволяющих максимально полно извлекать информа-

цию, содержащуюся в выравненных последовательностях. Такие методы были предложены рядом исследователей [1, 2]. Все они направлены на то, чтобы в той или иной степени учитывать неравномерность встречаемости отдельных аминокислотных остатков в каждой позиции массива выравненных последовательностей, составляющего обучающую выборку для формирования мотива. Наиболее последовательно такой

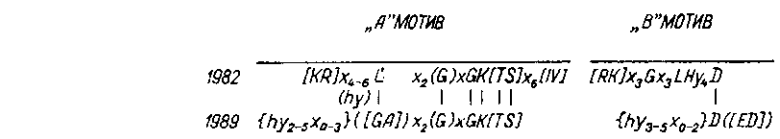

Рис. 1. «Размывание» NTP-связывающего паттерна по мере накопления родственных последовательностей, используемых для получения формулы паттерна. *hy* — гидрофобный аминокислотный остаток. Выражения, приведенные в фигурных скобках, означают, что из 5 последовательных аминокислотных остатков по крайней мере 2 («А» мотив) или 3 («В» мотив) являются гидрофобными. Остальные обозначения — как в табл. 1

Fig. 1. Deterioration of the NTP-binding pattern as the result of accumulation of related sequences used for generation of the pattern. *hy*, a hydrophobic amino acid residue. The expressions in figure brackets indicate that at least 2 («A» motif), or 3 («B» motif) out of 5 consecutive residues are hydrophobic. Other designation — as in Table 1

подход реализуется путем построения так называемого «частотного профиля» мотива, т. е. таблицы, содержащей так или иначе преобразованные частоты встречаемости всех аминокислотных остатков в каждой позиции. Такой профиль затем используется для оценки сходства фрагментов последовательностей с данным мотивом. Нам, однако, неизвестны исследования, в которых были бы представлены результаты систематического использования таких подходов для скрининга банков аминокислотных последовательностей. Мы считали целесообразным в рамках пакета «GENBEE» разработать достаточно простой вариант «профильного» метода, обеспечивающий возможность чувствительного скрининга банков последовательностей, статистической оценки сходства фрагментов извлекаемых последовательностей с используемым мотивом и создания из этих последовательностей специализированных баз данных. Мы рассматривали такую работу как первый подход к построению классификации белков, основанной на результатах сравнительного анализа их последовательностей. Ниже кратко описаны предложенный алгоритм, написанная на его основе программа «SITE» и приведен пример ее использования.

Алгоритм поиска мотивов и программа «SITE». Схема предложенного алгоритма поиска мотива в аминокислотных последовательностях представлена на рис. 2. Мы использовали простейший вариант профиля мотива, представляющий собой прямоугольную матрицу размером $20 \times l$ (l — число позиций в мотиве), элементы которой подсчитывали по формуле

$$M_{ir} = \sum_{j=1}^{j=20} w_{ij} \cdot f_{jr},$$

где M_{ir} — значение профиля для i-го аминокислотного остатка в r-й позиции мотива; w_{ij} — вес сравнения i-го и j-го остатков по матрице Дэйхофф [6] или по какой-либо другой матрице сравнения аминокислотных остатков; f_{jr} — число вхождений j-го остатка в r-й позиции. Этот профиль (матрицу) использовали для сканирования банка аминокислотных последовательностей. Для каждой последовательности в банке вычисляли величину сходства каждого сегмента длины l с данным профилем по формуле

$$S = \sum_{r=1}^{r=l} M_{ir}$$

(т. е. если при сканировании последовательности в r-й позиции фрагмента длины l об-

наруживается i-й аминокислотный остаток, то в качестве слагаемого в сумму S, характеризующую величину сходства, входит M_{ir}). Для каждой последовательности в банке запоминалось максимальное значение S_m и вычислялось среднее

$$S_{av} = \sum S_m / N,$$

где N — число последовательностей в банке. Затем подсчитывали величину среднеквадратичного отклонения (σ) и величину «приведенной суммы» для каждой последовательности по формуле

$$S' = (S_m - S_{av})/\sigma.$$

Величина S' представляет собой число среднеквадратичных отклонений, на которое лучшее значение сходства с мотивом для данной последовательности превышает среднее значение по банку, и характеризует статистическую значимость сходства. Строится гистограмма распределения значений S' по банку и подсчитывается доля значений, превышающих заданный порог. Соответствующие последовательности отмечаются в исходном банке и могут быть выделены в отдельную базу данных для дальнейшего анализа. Очевидно, что время счета по данному алгоритму прямо пропорционально $l \cdot D$ (D — суммарная длина последовательностей в банке), но не зависит от числа последовательностей в мотиве.

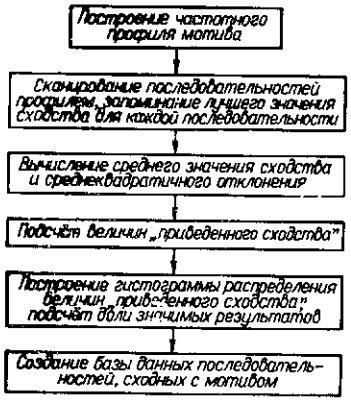

Рис. 2. Схема «профильного» алгоритма поиска мотивов в аминокислотных последовательностях
Fig. 2. A scheme of the «profile» algorithm for motif search in protein sequences.

Программа «SITE», написанная в рамках пакета «GENBEE», предусматривает три режима поиска: 1) «PATTERN», режим поиска мотивов по формулам, подобным приведенным на рис. 1 и в табл. 1, основанный на стандартном алгоритме поиска регулярных выражений (типа, например, программы GREP в системе UNIX); 2) «MOTIF», поиск, основанный на описанном «профильном» алгоритме; 3) «BESTFIT» аналогичный метод поиска с использованием вместо мотива одной последовательности.

Результаты и обсуждение. Для отработки предложенного метода поиска структурных мотивов в аминокислотных последовательностях мы провели скрининг части (включающей около 25 % всех последовательностей) банка PIR (выпуск 15) с использованием в качестве зонда «А»-мотива NTP-связывающего паттерна (см. выше). Конкретно используемый мотив представлял собой 186 выравненных фрагментов последовательностей, как правило, удовлетворяющих формуле из работы [5], приведенной на рис. 1. Последовательности включали в мотив на основании имеющихся экспериментальных данных об участии данного фрагмента (или по крайней мере соответствующего белка) в связывании NTP либо на основании консервации последовательности в семействах достаточно сильно дивергировавших белков. Ниже мы будем говорить об NTP-связывающих белках, подразумевая белки, обладающие соответствующими свойствами и содержащие последовательности, достоверно родственные данному мотиву; не следует при этом забывать, что существуют NTP-связывающие белки с другими консервативными мотивами. На рис. 3 представлена гистограмма распределения величин «приведенной суммы» (см. выше) при скрининге используемой базы данных. Результаты поиска предложенным методом сравнивали с результатами поиска стандартной программой с использованием формулы «А»-мотива, приведенной на рис. 1.

Данные табл. 2 достаточно убедительно демонстрируют преимущества «профильного» метода поиска. Стандартная программа не обеспечивает полной селективности (что очевидно и из элементарных стати-

стических соображений): около половины отобранных с ее помощью белков, по всей вероятности, не обладают NTP-связывающими свойствами. Не менее важно то, что при использовании стандартной программы из рассмотрения выпадали некоторые белки, безусловно, обладающие NTP-связывающими свойствами, но содержащие в своей последова-

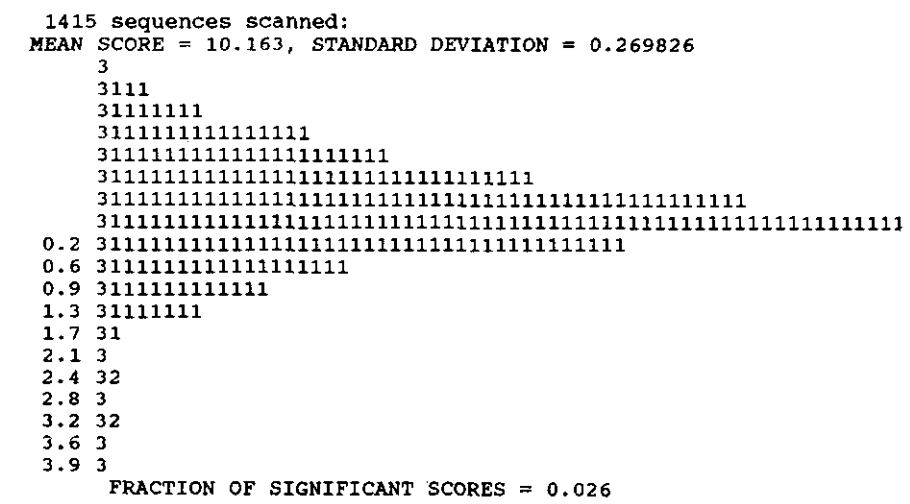

Рис. 3. Гистограмма распределения последовательностей в базе данных по сходству с «А» мотивом NTP-связывающего паттерна. По вертикали указаны значения «приведенного сходства» (в единицах σ), а по горизонтали — соответствующие доли последовательностей. Указаны среднее значение сходства с мотивом в расчете на один аминокислотный остаток, величина σ и доля последовательностей, для которых сходство с мотивом превышает заданный порог (в данном случае 2σ)

Fig. 3. The distribution of the sequences in the database by similarity to the «A» motif of the NTP-binding pattern. Ordinate: adjusted score values (number of standard deviations above the mean); abscissa: respective fractions of the sequences. The mean per residue score for comparison of the sequences in the database with the motif, the SD value, and the fraction of scores above the user-defined cut-off level (2 SD, in this case) are indicated.

Таблица 2

Сравнение результатов поиска NTP-*связывающего мотива стандартным и профильным методами*

Comparison of the results of searches for the NTP-binding motif made by standard and profile methods

Метод поиска	N			C	S_1	S_2	$E=C\cdot S$	
	I	II	III				1	2
Стандартный	30	7	36	0,86	0,41	0,51	0,35	0,44
Профильный порог								
3σ	19	0	0	0,54	1,00	1,00	0,54	0,54
2,5σ	26	4	0	0,74	0,87	1,00	0,64	0,74
2σ	32	4	(1)	0,91	0,86	0,97	0,78	0,88
1,8σ	35	4	13	1,00	0,67	0,75	0,67	0,75

Примечание. N—число селектированных последовательностей: I—NTP-связывающие белки; II—белки, вероятно связывающие NTP; III—белки, вероятно не обладающие NTP-связывающими свойствами; C—чувствительность метода, т. е. полнота извлечения NTP-связывающих белков (за 1 принимали все NTP-связывающие белки, селектированные каким-либо из использованных методов); S_1, S_2—избирательность извлечения NTP-связывающих белков (S_1=I/I+II+III; S_2=I+II/I+II+III). E—эффективность метода (1: $E=S_1\cdot C$; 2: $E=S_2\cdot C$). При использовании порога 2σ единственная последовательность, по-видимому, не входящая в рассматриваемый класс NTP-связывающих белков, принадлежала аланил-аминоацил-тРНК синтетазе; хотя аминоацил-тРНК синтетазы утилизируют ATP, для них характерна консервация последовательностей, отличных от «А» мотива.

тельности отклонения от формулы мотива (например, аденилаткиназы). «Профильный» метод с оптимальным порогом значимости (в нашем случае 2σ) позволяет добиться практически полной селективности при извлечении из базы подавляющего большинства последовательностей NTP-связывающих белков. При более низком пороге ($1{,}8\sigma$) селектировались все последовательности NTP-связывающих белков, но возникал некоторый «фон». Интересно, впрочем, что из 13 селектированных при этом пороге белков, не обладающих NTP-связывающими ствойствами, 8 являются NAD-зависимыми дегидрогеназами, т. е. относятся к другому классу нуклеотид-связывающих белков. «Профильный» метод, кроме того, однозначно выявлял упомянутые выше последовательности аденилаткиназ.

Он позволил также предположительно идентифицировать NTP-связывающий участок в dNMP-киназе бактериофага *T4*, в которой этот мотив ранее не был обнаружен.

Нам показалось интересным применить разработанный метод для анализа последовательностей некоторых белков, в которых предполагалось наличие NTP-связывающих мотивов, отклоняющихся от стандартной формулы. Мы рассмотрели последовательности белка *F* вирусов иммунодефицита человека и обезьян [7], белка *RecO*, участвующего в рекомбинации ДНК у *Escherichia coli* [8], и белка бактериофага *T4 UvsX*, также вовлеченного в рекомбинацию и репарацию фагового генома [9]. Оказалось, что в первых двух белках фрагменты, наиболее сходные с «А»-мотивом, не совпадают с предсказанными авторами соответствующих работ и характеризуются значениями сходства, лишь незначительно превышающими средние в исследованной нами базе данных. Это делает идентификацию в них NTP-связывающего мотива весьма сомнительной. Для белка *UvsX* участок, наиболее близкий к «А»-мотиву, также отличается от указанного в оригинальной работе, но совпадает с предсказанным нами на основании сравнения с последовательностями бактериальных белков *RecA* [5]. Хотя в этом случае уровень сходства с мотивом не очень высок («приведенное сходство» около $1{,}2\sigma$), представляется достаточно вероятным, что идентифицированный сегмент участвует в связывании NTP.

Таким образом, хотя профильный метод (по крайней мере в описанном здесь первоначальном виде) не решает всех проблем, связанных с поиском мотивов в аминокислотных последовательностях, он позволяет намного более надежно селектировать участки, сходные с такими мотивами.

Мы видим два основных связанных между собой аспекта применения предложенной программы поиска мотивов в аминокислотных последовательностях: 1) скрининг банков для выделения групп родственных белков и 2) анализ новых аминокислотных последовательностей с целью выявления в них структурных мотивов и отнесения их к той или иной группе. Для решения этих задач необходимо создать «библиотеку» мотивов и соответствующих средних значений сходства и величин σ, полученных при скрининге банка программой «SITE». Такая работа ведется в настоящее время.

Интересно также, что данный подход можно использовать для выявления белков, содержащих наиболее типичные для исследуемого мотива последовательности, т. е. в известном смысле для реконструкции «предковой» последовательности.

Авторы глубоко признательны разработчикам пакета «GENBEE» Л. И. Бродскому, А. Л. Драчеву и Р. Л. Татузову за всестороннюю поддержку в работе; А. С. Кондрашову — за ценные замечания, сделанные при обсуждении описанного алгоритма, а также М. Н. Линьковой — за техническую помощь.

A METHOD FOR SEARCH OF STRUCTURE MOTIFS IN AMINOACID SEQUENCES PROGRAM SITE OF THE GENBEE PACKAGE

E. V. Koonin, K. M. Chumakov, A. E. Gorbalenya

Institute of Microbiology, the USSR Academy of Sciences, Moscow;
Institute of Poliomyelitis and Viral Encephalitides,
Academy of Medical Sciences of the USSR, Moscow Region

Summary

A method is suggested to search for the structure motifs in amino acid sequences, based on their scanning by frequency profiles generated from aligned sequence segments. The program «SITE» implementing the proposed algorithm is discussed, exemplified by the search of an amino acid sequence database for the motif typical of a vast class of purine NTP-binding proteins. The superiorities of the proposed approach as compared to standard pattern-searching routines is demonstrated with respect to selectivity and completeness of extraction of relevant sequences.

СПИСОК ЛИТЕРАТУРЫ

1. *Staden R.* Methods to difine and locate patterns of motifs and sequences // CABIOS.— 1988.— **4**, N 1.— P. 53—60.
2. *Hodgman T. C.* The elucidation of protein function by sequence motif analysis // Ibid.— 1989.— 5, N 1.— P. 1—13.
3. *Кимура М.* Молекулярная эволюция: теория нейтральности.— М. : Мир, 1985.— 398 с.
4. *Distantly* related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold / J. E. Walker, M. Saraste, M. J. Runswick, N. J. Gay // EMBO J.— 1982.— **1**, N 8.— P. 945—951.
5. *Gorbalenya A. E., Koonin E. V.* Viral proteins containing the NTP-binding sequence pattern // Nucl. Acids Res.— 1989.— **17**, N 22.— P. 8413—8440.
6. *Dayhoff M. O., Barker W. C., Hunt L. T.* Establishing homologies in protein sequences // Meth. Enzymol.— 1983.— **91**.— P. 524—549.
7. *HIV F/3 orf* encodes a phosphorylated GTP-binding protein resembling an oncogene product / B. Guy, M. P. Kieny, Y. Riviere et al. // Nature.— 1987.— **330**, N 6145.— P. 266—269.
8. *Molecular* analysis of the *Escherichia coli recO gene* / P. T. Morrison, S. T. Lovett, L. E. Gilson, R. Kolodner // J. Bacteriol.— 1989.— **171**, N 7.— P. 3641—3649.
9. *Fujisawa H., Yonesaki T., Minagawa T.* Sequence of the *T4* recombination gene, *uvsX*, and its comparison with that of the *recA* gene of *Escherichia coli* // Nucl. Acids Res.— 1985.— **13**, N 20.— P. 7443—7481.

Ин-т микробиологии АН СССР, Москва
Ин-т полиомиелита и вирус. энцефалитов АМН СССР,
Моск. обл.

Получено 28.05.90

УДК 577.32

С. Г. Галактионов, В. М. Цейтин, И. А. Ваксер

КОМПЬЮТЕРНОЕ КОНСТРУИРОВАНИЕ БИОЛОГИЧЕСКИ АКТИВНЫХ ПЕПТИДОВ: НЕКОТОРЫЕ НОВЫЕ ВОЗМОЖНОСТИ

Развиты алгоритмы теоретического конформационного анализа для расчета стабильных конформаций молекул пептидов на поверхности раздела фаз «вода — липофильная среда». Поскольку взаимодействие многих пептидных биорегуляторов со специфическими рецепторами осуществляется главным образом за счет гидрофобных сил, развитая процедура может быть использована для компьютерного проектирования аналогов природных биорегуляторов, обладающих повышенным сродством к рецептору, измененным спектром проявления биологической активности, пролонгированным действием.

Введение. Существующие стратегии конформационно направленного конструирования биологически активных пептидов основываются на поиске так называемых «биологически активных» конформаций.