Математическая биология и биоинформатика. 2010. Т. 5. № 2. С. 124-137.
URL: http://www.matbio.org/downloads/Galzitskaya2010(5_124).pdf

==========МАТЕРИАЛЫ III МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ==========
==========«МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА»==========

УДК: 577.322

Статистический анализ и предсказание неструктурированных остатков в белковых структурах

Лобанов М.Ю.[*], Галзитская О.В.[**]

Учреждение Российской академии наук Институт белка РАН,
142290 Пущино Московской обл., ул. Институтская, 4; факс: 8(4967)318-435

Аннотация. Получены данные по статистике встречаемости неструктурированных аминокислотных остатков на концах и в средней части белковой цепи. Создан алгоритм, основанный на модели Изинга (модели двух состояний) для предсказания неструктурированных остатков по аминокислотной последовательности. Тестирование данного метода на двух дополнительных базах показало, что наш метод входит в пятерку лучших.

Ключевые слова: *неструктурированные остатки, модель Изинга, чувствительность, специфичность*.

ВВЕДЕНИЕ

Предсказание структуры и функции белков является одним из главных направлений в структурной геномике. Особый интерес представляет предсказание так называемых неструктурированных участков белковой цепи (участков, не имеющих в нативном состоянии белка фиксированной пространственной структуры). Такие неструктурированные участки часто играют важную функциональную роль (см. обзоры [1, 2]). При этом неструктурированные участки могут структурироваться только тогда, когда связываются с другой молекулой. Предполагают, что отсутствие глобулярной структуры при физиологических условиях представляет значительное функциональное преимущество для нативно-развернутых белков, так как их большая доступная поверхность при небольшом размере белка и пластичность позволяет им более эффективно взаимодействовать с белками и нуклеиновыми кислотами по сравнению с глобулярными белками того же размера, обладающими ограниченной конформационной гибкостью [1].

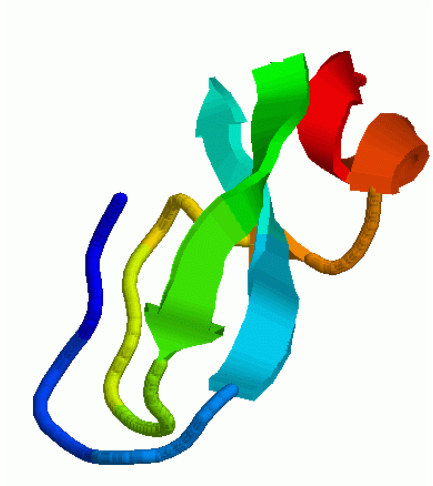
На сегодняшний день известно более 500 белков с неструктурированными участками [3]. Эти белки и домены либо целиком неструктурированны в нативном состоянии (так называемые нативно-развернутые белки), либо имеют протяжённые неструктурированные участки. При этом оказывается, что функционально важные белковые участки в таких белках часто находятся вне глобулярных доменов, то есть в тех самых неструктурированных участках [3, 4].

Поскольку неструктурированные участки белковой цепи играют важную роль в функционировании белка, то их предсказанию уделяется большое внимание. На

[*] mlobanov@phys.protres.ru
[**] ogalzit@vega.protres.ru

сегодня для этих целей разработаны специализированные программы, такие как FoldUnfold, PONDR, RONN, DisEMBL, PreLINK, IUPred, GlobPlot, FoldIndex и другие. Исходя из принципа, лежащего в основе их работы, все эти программы можно разделить на две группы. Программы FoldUnfold, PONDR, IUPred, GlobPlot, PreLINK и FoldIndex предсказывают неструктурированные участки белковой цепи, основываясь на физико-химических свойствах аминокислотных остатков в белке. В роли такого свойства может выступать локальный аминокислотный состав и гидрофильность (PONDR) [5, 6], число ожидаемых контактов (FoldUnfold) [7–9], способность участка цепи образовывать гидрофобный кластер (PreLINK) [10] или оценка энергетического взаимодействия между соседними аминокислотными остатками (IUPred) [11, 12]. GlobPlot оценивает тенденцию остатков находиться в регулярной вторичной структуре [13]. В основе программы FoldIndex лежит специально разработанная шкала заряд/гидрофобность для аминокислотных остатков [14].

Вторая группа программ использует выравнивания гомологичных белковых последовательностей. Программа RONN [15] использует нейронную сеть и сравнивает данную последовательность с рядом последовательностей, которые заранее отнесены к структурированным, к неструктурированным или к смеси того и другого. Программа DisEMBL использует нейронную сеть, натренированную на рентгеноструктурных данных [16]. DISOPRED – метод, использующий нейронную сеть, натренированную так, чтобы уметь отличать участки, которые пропущены в структуре, полученной рентгеноструктурным анализом [17].



1co7 I

KMSRLCLSVALLVLLGTLAASTPGCDTSNQAKAQRPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGAIG
PWENL

Рис. 1. Трехмерная структура ингибитора гидролазы, с PDB-кодом 1co7 цепь I. Внизу приведена аминокислотная последовательность. Цветом выделены участки, не разрешенные рентгеноструктурным анализом. Координаты структуры даны с разрешением 1.90 Å.

В работе проведен статистический анализ неструктурированных аминокислотных остатков в структурах белков, находящихся в банке белковых структур, PDB декабрь 2008 года. Получены данные по статистике встречаемости неструктурированных участков на концах и в средней части белковой цепи: в областях вблизи концов белковой цепи (на расстоянии до 30 остатков от N- либо от C-конца) находится 66% неструктурированных остатков (38% – вблизи N-конца и 28% – вблизи C-конца), при том, что эти краевые области включают всего 23% аминокислотных остатков. Интересен тот факт, что две шкалы, полученные из различных статистик (статистика контактов в глобулярных структурах и статистика неструктурированных остатков в банке белковых структур), коррелируют на уровне 95%. В данной работе создан новый алгоритм, основанный на модели Изинга [18], для предсказания неструктурированных

остатков. Параметры для программы получены и оптимизированы из статистики белковых структур. Тестирование данного метода на двух дополнительных базах показало, что наш метод позволяет делать надежные предсказания.

Изначально модель Изинга была предложена для описания линейного массива ферромагнетиков с взаимодействиями между ближайшими соседями. Стоит отметить, что модель Изинга была успешно применена для описания перехода спираль-клубок для гомополипептидных цепей, модели двух состояний [19].

МАТЕРИАЛЫ И МЕТОДЫ

Под неструктурированными остатками мы понимаем остатки, не разрешённые методом рентгеноструктурного анализа. А точнее, аминокислотные остатки, у которых не разрешён C_α-атом. Нами рассматривались все белковые структуры, решённые методом рентгеноструктурного анализа, с разрешением лучше 3 Å, опубликованные до 20.12.2008. Все 100%-ные гомологи были сгруппированы. Вес одной цепи обратно пропорционален числу гомологов. То есть каждая уникальная цепь берётся с весом 1. Полученная база содержит 28 727 уникальных цепей и 7 487 366 остатков. Из этих остатков 4.6% – неструктурированные. Используя эту базу, были получены все статистические результаты и потенциалы. Тестовая база была составлена по тем же принципам, что и основная база, но взяты белковые структуры, опубликованные после 20.12.2008, и использовалась для оценки качества работы созданного нами метода для предсказания неструктурированных остатков. Число уникальных цепей и число аминокислотных остатков в них представлены в табл. 1. В 2008 году проходило соревнование по предсказанию пространственной структуры белков, в том числе по предсказанию неструктурированных остатков с использованием базы CASP8: 8th Critical Assessment of Techniques for Protein Structure Prediction. Данная база использовалась для сравнения наших результатов с результатами других научных групп.

Таблица 1. Основные параметры рассматриваемых баз данных

	число уникальных цепей	Число остатков	доля неструктурированных остатков
Основная база белков на которой оптимизировались потенциалы	28 727	7 487 578	0.046
База белков, вышедших после того, как была сформирована база A	1 723	467 437	0.062
Casp8	122	27 489	0.107
Casp8-T0500	121	26 660	0.079

Алгоритм. Для предсказания неразрешённых остатков мы использовали алгоритм, основанный на простой физической модели, согласно которой каждый остаток может находиться в двух состояниях: фиксированном и свободном. При этом фиксированные остатки разрешаются методом рентгеноструктурного анализа, а свободные не разрешаются. Энергия переноса из одного состояния в другое зависит от типа остатка. Далее мы ввели энергию границы (по цепи) между фиксированными и свободными остатками. Кроме того, мы рассмотрели энергию перехода из условных точек *N* и *C* в

свободное состояние. Таким образом, мы можем рассчитать энергию j-го состояния цепи:

$$E_j = \sum_{i=1}^{L} w\left(a_i, s_{ij}\right) + k_j \cdot w_g + \delta_{N,j} \cdot w_N + \delta_{C,j} \cdot w_C \,. \qquad (1)$$

Здесь a_i – тип аминокислотного остатка и s_i – фиксированное (0) или свободное (1) состояние остатка, w_g – энергия границы, k – число границ, w_N, w_C – энергия перехода из виртуальных позиций N и C в свободное состояние, δ_N, δ_C равны нулю, если соответствующий концевой остаток находится в фиксированном состоянии, и единице в противоположном случае, L – длина белковой цепи. Энергия полностью фиксированного состояния считалась равной нулю, то есть $w(a_i, 0) \equiv 0$.

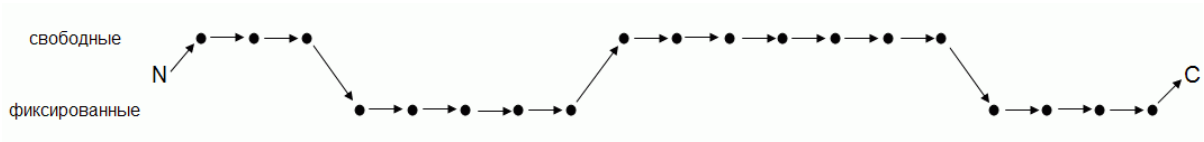


Рис. 2. Пример возможного состояния цепи. Видно, что конформацию полипептидной цепи можно представить как путь на графе. В этом примере $\delta_N = 1$, а $\delta_C = 0$.

Как известно из статистической физики, вероятность пребывания системы в одном из микросостояний пропорциональна $e^{-E_j/kT}$, где E_j – энергия микросостояния, T – температура, а k – константа Больцмана. Для всех возможных состояний мы можем вычислить статистическую сумму:

$$Z = \sum_{j=1}^{M} e^{-E_j/kT} \qquad (2)$$

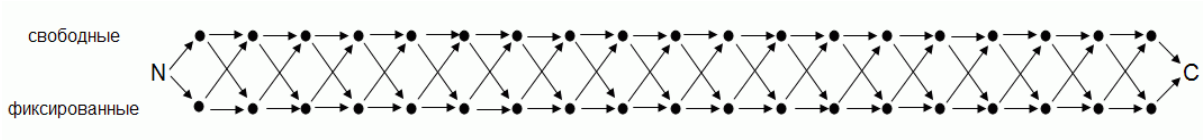


Рис. 3. Все возможные состояния белковой цепи.

Кроме того, в рамках нашей модели мы можем определить вероятность того, что конкретный остаток свободен (неструктурирован) ($s_i = 1$). Для этого нам надо подсчитать статистическую сумму подмножества состояний:

$$p_{i,1} = \frac{Z_{i,1}}{Z} = \frac{\sum_{j=1}^{M} s_{ij} \cdot e^{-E_j/kT}}{\sum_{j=1}^{M} e^{-E_j/kT}} \,. \qquad (3)$$

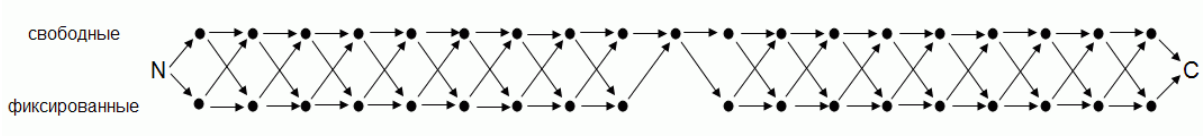


Рис. 4. Все состояния цепи с одним остатком в развернутом состоянии.

На первый взгляд, чтобы воспользоваться вышеприведёнными формулами, надо перебрать $M = 2^L$ возможных состояний, однако существует теория, позволяющая рассчитать Z и все p_i за время порядка L [18, 20, 21]. Из формулы (2) следует, что

$Z = Z^N_{L,1} + Z^N_{L,0}$, то есть статистическая сумма, равна сумме статсумм путей, идущих через последний свободный остаток ($Z^N_{L,1}$) и последний фиксированный остаток ($Z^N_{L,0}$). Нетрудно показать, что:

$$Z^N_{L,1} = e^{-w(a_L,1)/kT} \cdot e^{-w_C/kT} \cdot \left(Z^N_{L-1,1} + Z^N_{L-1,0} \cdot e^{-w_g/kT} \right), \quad (4)$$

$$Z^N_{L,0} = e^{-w(a_L,0)/kT} \cdot \left(Z^N_{L-1,0} + Z^N_{L-1,1} \cdot e^{-w_g/kT} \right). \quad (5)$$

Выполняя рекурсивно данную процедуру, мы можем рассчитать Z. В своих вычислениях мы шли от последнего остатка к первому, но ничего не изменится, если мы будем выполнять рекурсию от первого остатка к последнему: $Z = Z^C_{1,1} + Z^C_{1,0}$, где

$$Z^C_{1,1} = e^{-w(a_1,1)/kT} \cdot e^{-w_N/kT} \cdot \left(Z^C_{2,1} + Z^C_{2,0} \cdot e^{-w_g/kT} \right), \quad (6)$$

$$Z^C_{1,0} = e^{-w(a_1,0)/kT} \cdot \left(Z^C_{2,0} + Z^C_{2,1} \cdot e^{-w_g/kT} \right). \quad (7)$$

Более того, мы можем рассмотреть статсумму путей, стартуя от состояния (здесь свободного) произвольного остатка:

$$Z_{i,1} = \left(Z^N_{i-1,1} + Z^N_{i-1,0} \cdot e^{-w_g/kT} \right) \cdot e^{-w(a_i,1)} \cdot \left(Z^C_{i+1,1} + Z^C_{i+1,0} \cdot e^{-w_g/kT} \right). \quad (8)$$

Зная Z и $Z_{i,1}$, мы легко можем рассчитать вероятность того, что i-тый остаток будет в свободной конформации, т.е. будет неструктурирован (см. (3)).

Поскольку мы знаем, что 6 подряд идущих гистидинов практически всегда развёрнуты, мы не рассматривали пути, идущие через фиксированную конформацию хотя бы одного из таких гистидинов.

Оценка качества предсказаний неструктурированных участков. Для оценки качества предсказания использованы стандартные определения чувствительности и специфичности [22]:

$$S_n = TP/N_d, \quad (9)$$

$$S_p = TN/N_o. \quad (10)$$

Здесь S_n – чувствительность, S_p – специфичность, TP ("true positives") – число правильно предсказанных неструктурированных аминокислотных остатков, N_d – общее число неструктурированных остатков, TN ("true negatives") – число правильно предсказанных структурированных остатков, N_o – суммарное число структурированных остатков. Таким образом, чувствительность – это доля правильно предсказанных неструктурированных остатков, а специфичность – доля правильно предсказанных структурированных остатков [22].

Кроме того, в данной работе мы рассмотрели меру оценки качества предсказаний, которая используется на соревнованиях CASP ("Community Wide Experiment on the **C**ritical Assessment of Techniques for Protein Structure Prediction" – соревнования, проводимые по оценке качества методов предсказания пространственной структуры белков) в категории, посвящённой оценке предсказаний неструктурированных участков [23, 24] (http://predictioncenter.org/casp8/doc/presentations/CASP8_DR_Sussman.pdf):

$$Sw = \frac{W_1 TP - W_2 FP + W_2 TN - W_1 FN}{W_1 N_d + W_2 N_o}, \quad (11)$$

где *FP* ("false positives") – число ложноположительных предсказаний (число остатков, предсказанных как неструктурированные, на самом деле, являющихся структурированными), *FN* ("false negatives") – число ложноотрицательных предсказаний: число остатков, предсказанных как структурированные, по сути, являющихся неструктурированными, а W_1 и W_2 – коэффициенты, вычисленные следующим образом:

$$W_1 = \frac{N_o}{N} \cdot 100\%,$$

$$W_2 = \frac{N_d}{N} \cdot 100\%$$

($N = N_d + N_o$ – общее число аминокислотных остатков). /в формулах заменила * на умножение/

Нетрудно заметить, что формулу для расчёта *Sw* можно переписать, используя меньшее количество характеристик, нежели чем использовали авторы [23]. Подставив выражения для W_1 и W_2, получаем:

$$Sw = \frac{N_o(TP - FN) + N_d(TN - FP)}{2N_d N_o}. \quad (12)$$

Учитывая, что $FN = N_d - TP$, а $FP = N_o - TN$, получаем:

$$Sw = \frac{N_o(2TP - N_d) + N_d(2TN - N_o)}{2N_d N_o} = \frac{TP}{N_d} + \frac{TN}{N_o} - 1. \quad (13)$$

Или, используя определения чувствительности и специфичности, см. формулы (9), (10):

$$Sw = S_n + S_p - 1. \quad (14)$$

Этот критерий используется для оценки качества работы программ более 5 лет.

Оптимизация параметров. Результатом работы описанного выше алгоритма для одной цепи является вероятность пребывания каждого остатка в свободном ($p_{i,1}$) и фиксированном ($p_{i,0}$) состояниях ($p_{i,1} + p_{i,0} \equiv 1$). В реальной же структуре остаток может быть неструктурированным, то есть не разрешённым методом рентгеноструктурного анализа (1) и структурированным (0). Мы предсказываем остаток структурированным, если $p_{i,1} > p_m$, и неструктурированным, если $p_{i,1} \le p_m$. Естественной границей является $p_m = 0.5$.

Варьируя вероятность (p_m), выше которой мы считаем остаток неструктурированным, мы получаем разные пары чувствительности и специфичности, которые представлены на рис. 9. Значение площади под данной ROC-кривой характеризует надежность работы программы. Этот параметр используется как один из критериев качества работы программы, AUC-критерия. Этот критерий тоже используется более 5 лет. Считается, что если метод дает значение по этому параметру больше 0.8, то данный метод можно рассматривать как один из методов, дающих надежные предсказания.

Описанные выше критерии удобны для оценки качества работы программы, но они не очень удобны для оптимизации параметров, потому что: (а) меняются дискретно, и (б) локальный проигрыш может обернуться выигрышем в дальнейшем. Сказанное иллюстрируют графики на рис. 5.

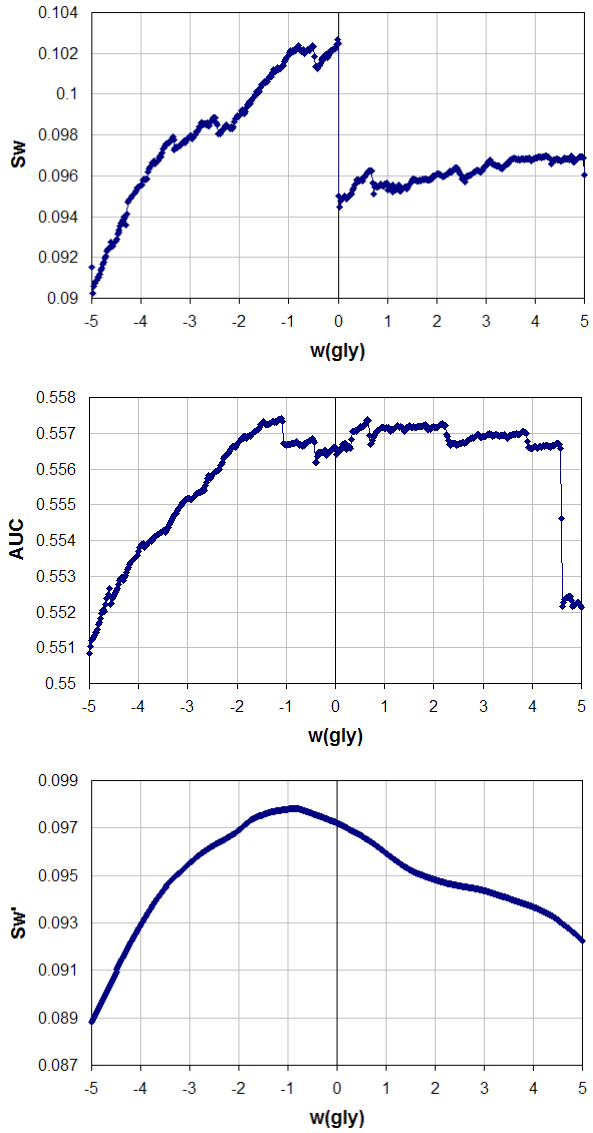


Рис. 5. Зависимости AUC, *Sw*, *Sw*' (пояснения в тексте ниже) при варьировании одного параметра *w*(Gly,1) (потенциал для глицина находиться в развернутом состоянии) при рассмотрении основной базы. Значения *w*(a, 1) для других аминокислотных остатков случайно приравнивались 5 и –5. Энергия границы равнялась 10. Энергия инициации на концах равнялась –10.

На рис. 5 показано изменение критериев качества при варьировании потенциала глицина. Для остальных остатков потенциалы были равны 5 или –5, при этом знак выбирался случайным образом, $w_g = 10$, $w_C = w_N = -10$. Здесь хорошо видно, что хотя оптимум по Sw' достигается при значениях *w*(Gly,1) чуть ниже нуля, однако если мы выберем положительное значение для потенциала, то так и не достигнем этого оптимума. Так же обстоят дела и с оптимизацией по AUC-критерию. Поэтому мы использовали для оптимизации критерий, тесно связанный со стандартными. Вместо N_{11} мы рассчитывали Q_{11}, суммируя по всем неразрешённым остаткам: $Q_{11} = \sum w_{chain} \cdot p_{i,1}$. Аналогично вычисляли (по всем разрешённым остаткам) $Q_{00} = \sum w_{chain} \cdot p_{i,0}$. Таким образом, получили параметр $Sw' = Q_{11} / N_d + Q_{00} / N_o - 1$. Значения этого параметра меняются плавно, в отличие от *Sw*, и имеют один хорошо выраженный максимум.

У оптимизации есть ещё один подводный камень. Сделаем один из параметров равным 1000. При его изменении на единицу (1001 или 999) Sw', *Sw*, *AUC* не

изменятся, если считать с обычной (double) для компьютера точностью. Так, если считать в Excel'e $\left(10^{16}+1\right)-10^{16}=0$. Таким образом, если мы в процессе оптимизации получим очень большое (по модулю) число, то изменить его будет очень сложно. Эта же проблема существует и для более реальных значений. Более того, при разумных параметрах (дающих $Sw>0$) $Sw' \le Sw$. А Sw' приближается к Sw при температуре, стремящейся к нулю (то есть при синхронном увеличении значений всех потенциалов). По этой причине мы оптимизировали не саму величину Sw', а $Sw''=Sw'-0.02/(kT)^2$ (см. рис.6). При этом мы считали, что:

$$\frac{1}{kT}=\left(\sum_{i=1}^{20} w^2(i,1)\cdot p(i)+p_g w_g^2+\frac{N_{chain}}{N_{residue}}\left(w_N^2+w_C^2\right)\right)^{1/2}, \qquad (15)$$

где $p(i)$ – частота встречаемости i аминокислоты. Естественно, $\sum_{i=1}^{20} p(i)\equiv 1$. p_g – частота встречаемости границы между разрешёнными и неразрешёнными остатками.

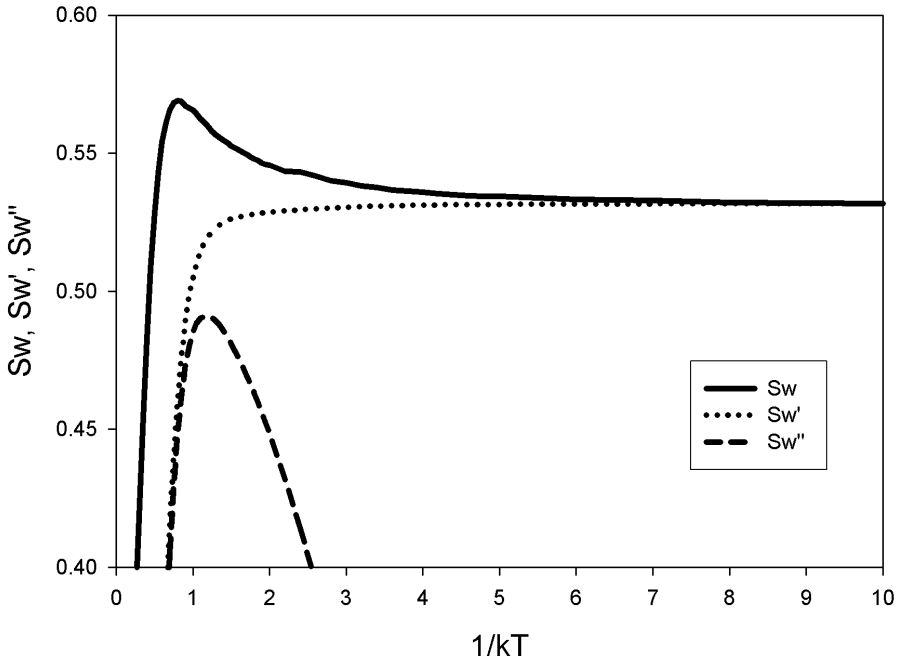


Рис. 6. Зависимость S*w*, S*w′*, S*w″* от обратной температуры для потенциалов, которые оптимальны для S*w*''.

Sw'' мы максимизировали методом Монте-Карло. Случайным образом меняли один из потенциалов. Если изменение вело к увеличению *Sw″*, то оно принималось. В противном случае отвергалось.

В дальнейшем мы оптимизировали *Sw*, *AUC* и *Sw*+*AUC*, стартуя от потенциалов, полученных при оптимизации *Sw″* (масштабируя по температуре).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Нами было проанализировано распределение неструктурированных остатков в полученной базе данных. Была подсчитана статистика встречаемости неструктурированных участков различной длины. Отдельно были рассмотрены *N*-концевые неструктурированные участки, отдельно – *C*-концевые, отдельно – внутренние неструктурированные петли (неструктурированные участки, на обоих краях которых находятся структурированные области). Чаще всего встречаются

неструктурированные участки длиной в 1 остаток на *N*- и на *C*-концах белков. В средней части белковой цепи чаще всего встречаются участки размером четыре аминокислотных остатка.

Статистика распределения неструктурированных остатков в белковых цепях показала, что 2/3 (66%) всех неструктурированных аминокислотных остатков находятся на краях белковых цепей (в пределах 30 остатков от *N*- или *C*-конца белковой цепи), при том, что эти краевые остатки включают всего 23% аминокислотных остатков белковой молекулы. Поэтому для дальнейшего изучения встречаемости неструктурированных остатков мы рассматривали отдельно краевые части (находящиеся в пределах 30 остатков либо от *N*-, либо от *C*-конца) и среднюю часть белковой цепи (все остальные остатки).

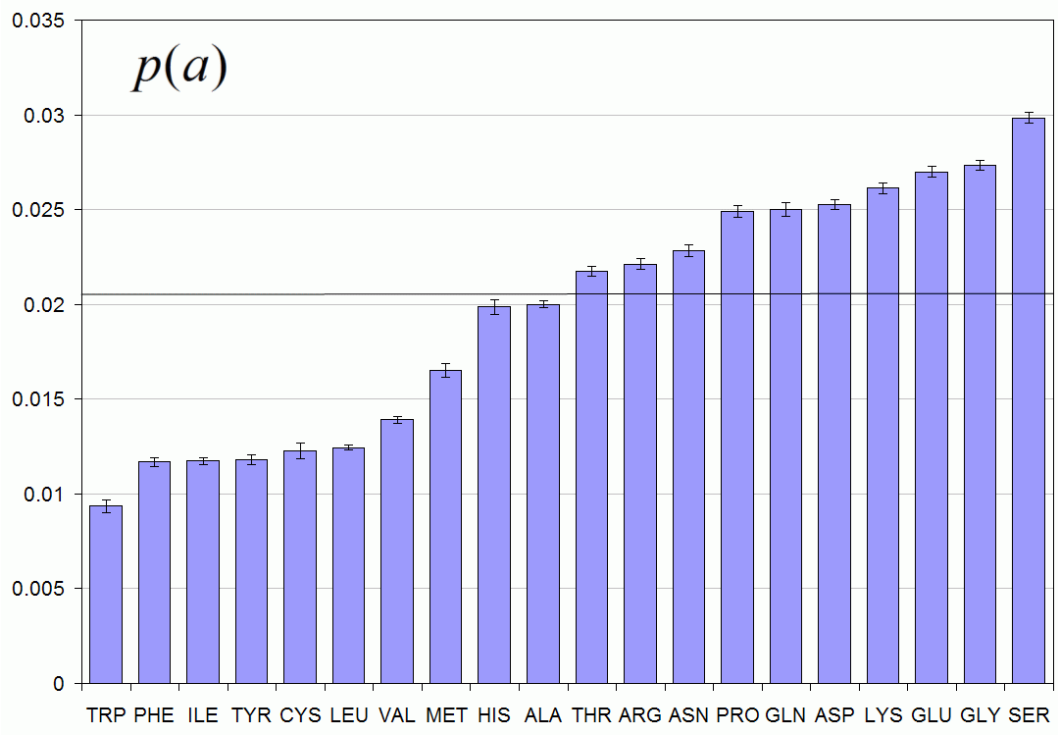


Рис. 7. Доля неструктурированных остатков в центре белковой цепи. Горизонтальная линия соответствует среднему значению.

На рис. 7 представлена доля неструктурированных аминокислотных остатков каждого из 20 типов в средней части белковой цепи. Как видно из приведенной гистограммы, доля неструктурированных аминокислотных остатков в средней части белковой цепи варьирует от 0.01 (для триптофана) до 0.03 (для серина). Как и следовало ожидать, доля неструктурированных аминокислотных остатков ниже для гидрофобных остатков и выше для гидрофильных. Интересно отметить, что серин чаще бывает неструктурированным, чем любой другой тип аминокислотных остатков (в том числе глицин и пролин, которые (или хотя бы один из них) обычно [7, 13] выделяют как остатки с наибольшей "предрасположенностью" к нахождению в неструктурированных участках). Как видно из погрешностей, указанных на гистограмме, это отличие достоверно.

Работа программы FoldUnfold для предсказания неструктурированных аминокислотных остатков была (в исходном варианте программы [7, 9]) основана на шкале предсказанных (ожидаемых) контактов, которая была получена нами [7–9] при анализе контактов, наблюдаемых в глобулярных структурах белков, и использована нами для поиска участков, образующих аномально малое число контактов. Сравнение

двух шкал (контактной и статистической) показало, что полученная шкала встречаемости неструктурированных остатков в средней части белковой цепи коррелирует со шкалой контактов на уровне 95%. Интересен тот факт, что две шкалы, полученные из различных статистик (статистика контактов в глобулярных структурах и статистика неструктурированных остатков в банке белковых структур), коррелируют на уровне 95%.

В нашей работе мы использовали алгоритм, основанный на модели Изинга. Каждый остаток может быть в двух состояниях: структурированном и неструктурированном. Энергия каждого остатка в том или ином состоянии зависит от типа остатка и состояния. Энергия полностью структурированного состояния была взята за ноль. Используя созданную базу, нами были подобраны оптимальные параметры для предсказания вероятности нахождения аминокислотного остатка в структурированном или неструктурированном состояниях. Кроме того, мы ввели энергию границы между структурированными и неструктурированными остатками. Нами было рассмотрено 23 параметра: 20 значений потенциалов для аминокислотных остатков быть в развернутом состоянии, отдельно энергии для *N*- и *C*-концов и одна энергия границы (см. табл. 2).

Таблица 2. 23 энергетических потенциала для неструктурированных аминокислотных остатков, полученных в процессе оптимизации

w_g	w_N	w_C							
3.67	-5.33	-6.65							
SER	**GLU**	**PRO**	**GLN**	**ASP**	**LYS**	**HIS**	**ASN**	**ARG**	**GLY**
-0.58	-0.54	-0.52	-0.39	-0.34	-0.32	-0.31	-0.30	-0.25	-0.21
ALA	**THR**	**MET**	**LEU**	**VAL**	**ILE**	**CYS**	**TYR**	**PHE**	**TRP**
0.07	0.08	0.11	0.68	0.70	0.93	0.94	1.02	1.02	2.00

В дальнейшем мы подсчитывали статистическую сумму всех возможных состояний цепи, а также статистическую сумму всех состояний при заданном состоянии одного остатка. Разделив одно на другое, мы получали вероятность пребывания остатка в этом состоянии.

На рис. 8 показан один из примеров работы нашего метода в сравнении с экспериментальными данными. Для данной структуры в банке белковых структур представлено 6 цепей, и для каждой цепи есть некая вариация по набору неструктурированных остатков. В таких случаях мы делали усреднение по всем белковым цепям, представленным в банке белковых структур.

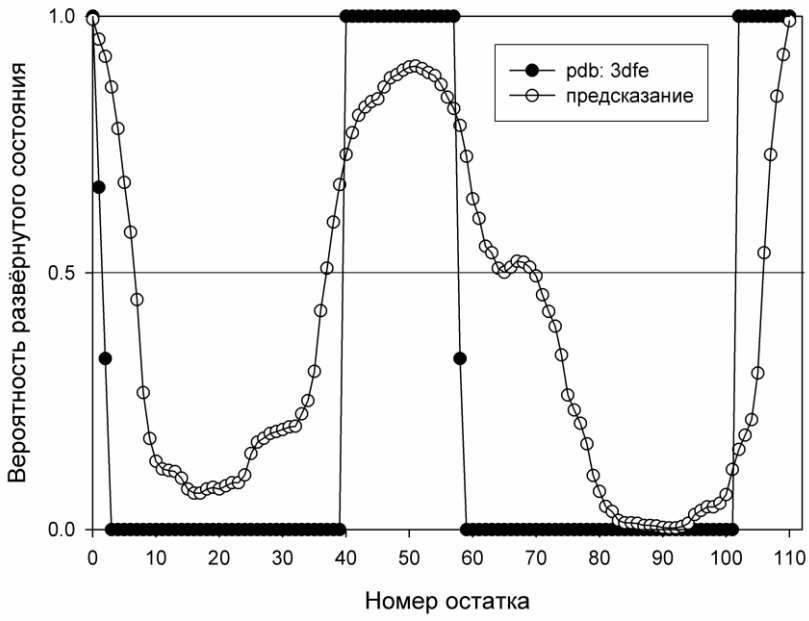


Рис. 8. Сравнение профилей вероятностей для каждого остатка быть в развернутом состоянии для трехмерной структуры (вероятность для остатка быть развернутым усреднена по 6 белковым цепям, представленным в PDB-файле 3dfe) и предсказанным нашим методом. Остаток считается развернутым, если вероятность больше 0.5.

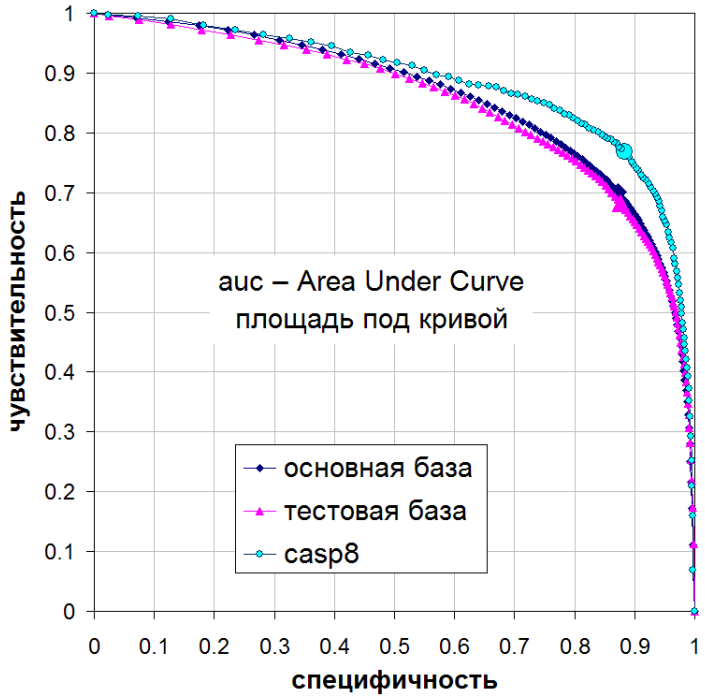


Рис. 9. ROC-кривые для предсказания неструктурированных остатков нашим методом для разных баз данных. Символы на кривых соответствуют вероятности $p = 0.5$.

Мы использовали стандартные критерии оценки качества предсказания: чувствительность (доля правильно предсказанных неструктурированных остатков) и специфичность (доля правильно предсказанных структурированных остатков). Сумма этих параметров минус 1 даёт значение для параметра *Sw*, используемого как один из

критериев оценки качества работы программ по предсказанию неструктурированных остатков. Варьируя вероятность, выше которой мы считаем остаток неструктурированным, можно получить разные пары чувствительности и специфичности, которые и показаны на рис. 9.

Таблица 3. Эффективность работы нового метода

База данных	чувствительность	специфичность	*Sw*	*AUC*
База, на которой оптимизировались потенциалы	0.70	0.87	0.57	0.86
База белков, вышедших после того, как была сформирована база A	0.68	0.88	0.56	0.85
Casp8	0.77	0.88	0.65	0.89
Casp8-T0500	0.67	0.88	0.56	0.84

Наши результаты можно сравнить с результатами других людей, полученными на международном соревновании CASP8 по предсказанию трехмерных структур в категории предсказания неструктурированных остатков (см. рис. 10). Как видно, по параметру AUC у нас результаты выше среднего, зато по критерию S*w* мы среди лучших четырех методов, три из которых это мета-серверы [25]. Отдельно хочется упомянуть цель T0500, которая сильно повлияла на оценку параметра *Sw*. Данный белок оказался полностью неструктурированным, и к тому же длиной около 500 аминокислотных остатков. Из табл. 3 видно, что после удаления данной цели, значения всех параметров близки к значениям параметров, полученным по двум другим базам.

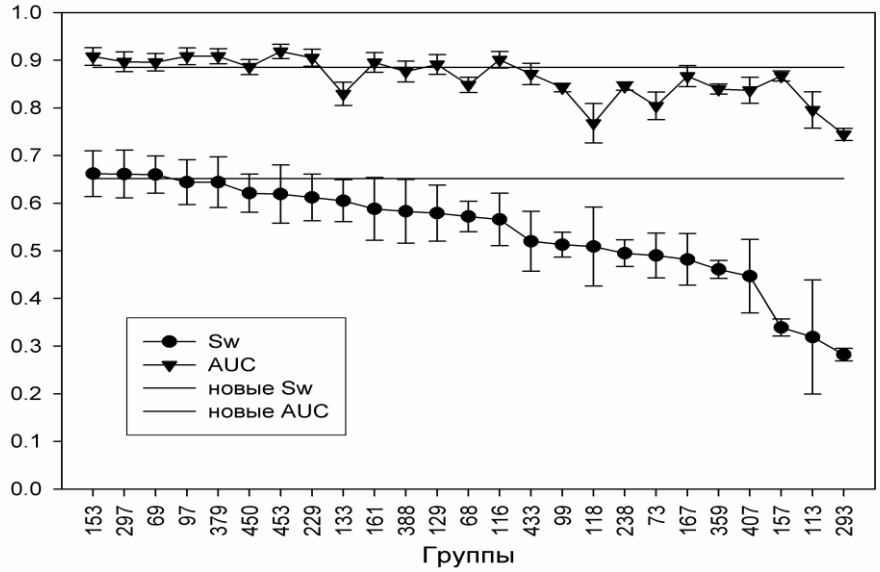


Рис. 10. Сравнение результатов нашего метода с предсказаниями других групп, участвующих в соревновании CASP8 [25]. Горизонтальные линии показывают качество работы нашего метода.

Используя разные параметры для оценки эффективности работы нашего метода (чувствительность, специфичность, *Sw*, *AUC*), можно сказать, что он позволяет делать надежные предсказания.

ЗАКЛЮЧЕНИЕ

Нами создана база неструктурированных остатков, включающая все белковые цепи, имеющиеся в банке белковых структур на момент декабря 2008 года. Создан алгоритм, основанный на модели Изинга, для предсказания неструктурированных остатков. Используя созданную базу, подобраны оптимальные параметры для предсказания. Тестирование на двух дополнительных базах показало, что наш метод входит в пятерку лучших.

Работа выполнена при финансовой поддержке РФФИ (грант № 08-04-00561), при поддержке Российской академии наук (программа «Молекулярная и клеточная биология» (01200959110) и «Фундаментальные науки – Медицине»), Федерального агентства по науке и инновациям (02.740.11.0295).

СПИСОК ЛИТЕРАТУРЫ

1. Tompa P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* 2002. V. 27. P. 527–533.
2. Wright P.E., Dyson H.J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol*. 1999. V. 293. P. 321–331.
3. Sickmeier M., Hamilton J.A., LeGall T., Vacic V., Cortese M.S., Tantos A., Szabo B., Tompa P., Chen J., Uversky V.N., Obradovic Z., Dunker A.K. DisProt: the database of disordered proteins. *Nucl. Acids Res.* 2007. V. 35. P. D786–793.
4. Dunker A.K., Brown C.J., Lawson J.D., Iakoucheva L.M., Obradovic Z. Intrinsic disorder and protein function. *Biochemistry.* 2002. V. 41. P. 6573–6582.
5. Romero P., Obradovic Z., Kissinger C.R., Villafranca L.E., Dunker A.K. Identifying disordered regions in proteins from amino acid sequence. *Proc. of the IEE International Conference on Neural Networks*. 1997. P. 90–95.
6. Li X., Romero P., Rani M., Dunker A.K., Obradovic A.Z. Prediction protein disordered for N-, C-, and internal regions. *Genome Inform.* 1999. V. 10. P. 30–40.
7. Galzitskaya O.V., Garbuzynskiy S.O., Lobanov M.Yu. Prediction of amyloidogenic and disordered regions in protein chain. *PLoS Comput. Biol.* 2006. V. 2. P. e177.
8. Galzitskaya O.V., Garbuzynskiy S.O., Lobanov M.Yu. FoldUnfold: web server for the prediction of disorderred regions in protein chain. *Bioinformatics*. 2006. V. 22. P. 2948–2949.
9. Галзитская О.В., Гарбузинский С.А., Лобанов М.Ю. Предсказание нативно-развернутых участков белковой цепи. *Молекуляр. биология*. 2006. Т. 40. С. 341–348.
10. Coeytaux K., Poupon A. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*. 2005. V. 21. P. 1891–1900.
11. Dosztányi Z., Csizmók V., Tompa P., Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* 2005. V. 347. P. 827–839.
12. Dosztányi Z., Csizmók V., Tompa P., Simon I. IUPred: web server for the prediction of intrinsically unstructured regions based on estimated energy content. *Bioinformatics.* 2005. V. 21. P. 3433–3434.
13. Linding R., Russell R.B., Neduva V., Gibson T.J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucl. Acids Res.* 2003. V. 31. P. 3701–3708.
14. Prilusky J., Felder C.E., Zeev-Ben-Mordehai T., Rydberg E.H., Man O., Beckmann J.S., Silman I., Sussman J.L. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*. 2005. V. 21. P. 3435–3438.

15. Yang Z.R., Thomson R., McNeil P., Esnouf R.M. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*. 2005. V. 21. P. 3369–3376.
16. Linding R., Jensen L., Diella F., Bork P., Gibson T., Russell R. Protein disorder prediction: implications for structural proteomics. *Structure*. 2003. V. 11. P. 1453-1459.
17. Ward J.J., McGuffin L.J., Bryson K., Buxton B.F., Jones D.T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics*. 2004. V. 20. P. 2138–2139.
18. Ising E. Beitrag zut Theorie des Ferromagnetizmus. *Zeitschr. Phys.* 1925. V. 31. P. 253–258.
19. Zimm B.H., Bragg J.K. Theory of the phase transition between helix and random coil in polypeptide chains. *J. Chem. Phys.* 1959. V. 31. P. 526–535.
20. Finkelstein A.V. Theory of protein molecule self-organization. III. A calculating method for the probabilities of the secondary structure formation in an unfolded protein chain. *Biopolymers*. 1977. V. 16. P. 525–529.
21. Finkelstein A.V., Roytberg M.A. Computation of biopolymers: A general approach to different problems. *BioSystems*. 1993. V. 30. P. 1–19.
22. Melamud E., Moult J. Evaluation of disorder predictions in CASP5 *Proteins*. 2003. V. 53 (Suppl. 6). P. 561–565.
23. Jin Y., Dunbrack R.L., Jr. Assessment of disorder predictions in CASP6. *Proteins*. 2005. V. 61 **(**Suppl. 7**).** P. 167–175.
24. Bordoli L., Kiefer F., Schwede T. Assessment of disordered predictions in CASP7 *Proteins*. 2007. V. 69 (Suppl. 8). P. 129–136.
25. Noivirt-Brik O., Prilusky J., Sussman J.L. Assessment of disordered predictions in CASP8. *Proteins*. 2009. V. 77 (Suppl. 9). P. 210–216.

Материал поступил в редакцию 22.11.2010, опубликован 02.12.2010.