МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2019, том 53, № 1, с. 120–131

СТРУКТУРНО-ФУНКЦИОНАЛЬНЫЙ АНАЛИЗ БИОПОЛИМЕРОВ И ИХ КОМПЛЕКСОВ

УДК 577.112.5

ИДЕНТИФИКАЦИЯ УНИКАЛЬНОГО АМИНОКИСЛОТНОГО СОСТАВА БЕЛКОВ ЧЕЛОВЕКА, СВЯЗЫВАЮЩИХ КРЕСТООБРАЗНЫЕ СТРУКТУРЫ[1]

© 2019 г. M. Bartas[a, 2], P. Bažantová[a, 2], V. Brázda[b], J. C. Liao[b, c], J. Červeň[a], P. Pečinka[a, *]

[a] *Department of Biology and Ecology/Institute of Environmental Technologies, Faculty of Science, University of Ostrava, Ostrava, 710 00 Czech Republic*

[b] *Institute of Biophysics, Academy of Sciences of the Czech Republic v.v.i., Brno, 612 65 Czech Republic*

[c] *School of Medicine, The University of Queensland, Greenslopes Private Hospital, Greenslopes, 4120 Australia*

**e-mail: petr.pecinka@osu.cz*

Поступила в редакцию 11.10.2017 г.
После доработки 20.02.2018 г.
Принята к публикации 20.02.2018 г.

Крестообразные структуры – предпочтительные мишени для многих структурных и регуляторных белков, а также ряда ДНК-связывающих белков со слабой специфичностью к последовательности. Некоторые из этих белков также способны индуцировать образование крестообразных структур при связывании ДНК. Нами проанализирован аминокислотный состав 18 белков *Homo sapiens*, связывающих крестообразные структуры (CBPs). При сравнении с обобщенной частотой встречаемости аминокислот во всех белках человека выявлены уникальные характеристики состава CBPs: обогащение остатками лизина и серина и/или пониженное содержание остатков аланина, глицина, глутамина, аргинина, тирозина и триптофана. Основываясь на анализе методом bootstrap resampling (повторное формирование выборки) и анализе нечетких кластеров, можно предложить несколько молекулярных механизмов взаимодействия CBPs с крестообразными структурами ДНК, включая те, которые связаны с репарацией ДНК, транскрипцией и регуляцией на уровне хроматина. Белки DEK, HMGB1 и TOP1, в частности, сформировали отдельную группу, отличающуюся от других белков. Кроме того, выявлена сеть сильных взаимодействий, объединяющая практически все исследованные CBPs. Полученные результаты могут быть востребованы для предсказании новых CBPs и даже для создания прогностического инструмента или интернет-приложения.

Ключевые слова: крестообразные структуры, ДНК–белковые взаимодействия, кластерный анализ, лизин, триптофан

DOI: 10.1134/S0026898419010026

Молекулы ДНК – носители генетической информации. Интересно, что регуляторные элементы часто находятся под управлением различных эволюционно консервативных коротких некодирующих последовательностей, которые могут образовывать разные локальные структуры, в том числе похожие на крест (крестообразные), квадруплексы или триплексы [1–4]. Крестообразные структуры ДНК наиболее известны на сегодняшний день. Эти структуры образуются последовательностями ДНК, содержащими инвертированные повторы (палиндромы), которые стабилизируются благодаря отрицательной суперспирализации ДНК [5] и/или связыванию с белками [6]. Образование крестообразных структур зависит от нуклеотидной последовательности и от наличия в ней совершенных или несовершенных инвертированных повторов из 6 и более нуклеотидов [7, 8]. Инвертированные повторы возникают неслучайным образом в ДНК всех организмов, и их часто обнаруживают в непосредственной близости от точек соединения разрывов ДНК, промоторных областей и в сайтах инициации репликации [9, 10]. Крестообразные структуры ДНК необходимы для реализации широкого круга биологических процессов, включая репликацию [11], регуляцию экспрессии генов [12], формирование нуклеосом [10] и рекомбинацию [13], поэтому неудивительно, что открыты белки, специфично распознающие крестообразные структуры [14].

Хорошо известно, что некоторые ДНК-связывающие белки имеют высокое сродство к крестообразным структурам ДНК. К ним относятся ас-

[1] Статья представлена авторами на английском языке.

[2] Эти авторы внесли одинаковый вклад в работу.

социированные с хроматином белки (из группы белков с высокой подвижностью, белок DEK) и белки, участвующие в репарации ДНК (p53, BRCA1), репликации (топоизомераза 1) и транскрипции (рецептор эстрогена 1). Одним из первых описанных белков, связывающих крестообразные структуры (cruciform binding proteins; CBPs), был белок высокой подвижности 1 (HMG1) – эволюционно консервативный и функционально важный компонент клеточного ядра [15]. Среди других подобных белков можно назвать 14-3-3σ и DEK, которые участвуют в распознавании крестовидных структур в процессах соответственно репликации и ремоделирования хроматина [13, 16]. Сообщалось, что опухолевый супрессор p53 имеет сродство ко многим неканоническим структурам ДНК, таким как крестообразные структуры [17, 18] и CTG-шпильки [19].

Насколько нам известно, статистических исследований, посвященных аминокислотному составу CBPs, пока нет. Интересно, что в последнее время для прогнозирования сложных функций белка стали использовать анализ аминокислотного состава [20–22].

В проведенном исследовании мы проанализировали аминокислотный состав CBPs человека с целью выяснить, имеют ли эти белки общие признаки, по которым можно предсказать их способность связываться с крестообразными структурами. Для этого мы выбрали только известные и хорошо охарактеризованные CBPs. Основная цель, которую мы ставили при проведении этого исследования, – выявление общих структурных характеристик, которые можно использовать для предсказания новых CBPs и расширения исследований в этой области.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Анализ аминокислотного состава. Канонические последовательности 18 CBPs (табл. 1) получены из базы данных UNIPROT (*Homo sapiens*) [23] (последовательности включены в **Supplementary 1, см. Приложение на сайте http://www.molecbio.ru/downloads/2019/1/supp_Bartas_rus.pdf**). Эти последовательности встраивали в инструмент ProtParam (https://web.expasy.org/protparam/) [24], полученную аминокислотную последовательность переносили в Excel 2007 и строили матрицу аминокислотного состава индивидуальных CBPs. Вычисленные средние значения аминокислотного состава CBP сравнивали с ожидаемыми значениями, которые представляют собой средние частоты встречаемости аминокислот в белках человека [25], и рассчитывали величину относительного повышения (обогащения) или понижения содержания (истощения) отдельных аминокислот (**Приложение/Suppl. 2**). Чтобы иметь возможность применить статистические методы, мы использовали определенные наборы белков, а относительное обогащение или истощение отдельных аминокислотных остатков в 18 CBPs рассчитывали с использованием web-инструмента Composition Profiler (http:// www.cprofiler.org/) [26]. В программе R, используя функцию *sample*, мы произвольно выбрали 5000 образцов из всех канонических белковых последовательностей человека, полученных из реферируемой базы данных UNIPROT [23] **(Приложение/Suppl. 3A)**, а также составили набор из 284 ДНК-связывающих белков и набор из 446 мембранных белков, полученных из той же базы данных [23] (наборы с меньшим числом белков и их последовательности вошли в **Приложение/Suppl. 3B, 3C и 3D**), и сравнили эти последовательности с исследуемыми CBPs. Для получения аминокислотного состава CBPs по 2-mer и 3-mer использовали онлайн-инструмент Pse-in-One (http://bioinformatics.hitsz.edu.cn/Pse-in-One/) [27] (данные доступны соответственно в **Приложении/Suppl. 4** и **Suppl. 5)**. Для поиска более протяженных K-меров проверена база данных HRaP (http://bioinfo.protres.ru/hrap/) [28].

Корреляционный анализ аминокислотного состава CBPs. Матрицу аминокислотного состава вышеперечисленных 18 CBPs обрабатывали в среде R с использованием пакета “corrplot” [29]. Полный исходный код представлен в **Приложении/Suppl. 6.**

Анализ неупорядоченных участков. Данные по выбранным CBPs в формате FASTA были направлены в PrDOS (http://prdos.hgc.jp/cgi-bin/top.cgi) (за исключением KMT2A и PRKDC, последовательности которых оказались слишком длинными для предсказания), используя настройки по умолчанию. PrDOS – это веб-приложение для предсказания внутренних неупорядоченных областей в белках по их первичным аминокислотным последовательностям в формате FASTA [30]. Исходя из предсказанных неупорядоченных областей, мы вычисляли процентное содержание внутренних неупорядоченных областей для каждого белка и строили диаграммы размаха в R (в качестве контрольной группы использовано 16 случайно выбранных белков из подмножеств ДНК-связывающих белков, доступных в **Приложении/Suppl. 7**). Нормальность и соответствие вариаций обоих наборов данных проверяли в R с помощью соответственно shapiro.test или var.test (**Приложение/Suppl. 8**). Для статистического анализа различий по содержанию областей внутренней неупорядоченности в CBPs и контрольной группе белков использован критерий Стьюдента для двух выборок в R. Для сравнения тот же анализ проведен с использованием программы IsUnstruct (http://bioinfo.protres.ru/IsUnstruct/) [31]

Кластерный анализ с помощью дендрограмм. Древовидная диаграмма построена с использо-

Таблица 1. Названия и идентификационные коды 18 проанализированных CBPs

Краткое название	Идентификационный код (UNIPROT)	Полное название
14-3-3s	P31947-1	Белок 14-3-3, сигма
AF10	P55197-1	Белок AF-10
BRCA1	P38398-1	Белок 1 предрасположенности к раку молочной железы
DEK	P35659-1	Белок DEK
ESR1	P03372-1	Рецептор эстрогена
HMGB1	P09429-1	Белок B1 группы высокой подвижности
IFI16	Q16666-1	Индуцируемый γ-интерфероном белок 16
KMT2A	Q03164-1	Гистон-лизин-N-метилтрансфераза 2A
MUS81	Q96NY9-1	Эндонуклеаза MUS81 точки кроссинговера
p53	P04637-1	Клеточный опухолевый антиген p53
PARP1	P09874-1	Поли (ADP-рибоза)полимераза 1
PRKDC	P78527-1	ДНК-зависимая протеинкиназа, каталитическая субъединица
R51A1	Q96B01-1	RAD51-ассоциированный белок 1
RAD54L	Q92698-1	RAD54-подобный белок ДНК-репарации и рекомбинации
TERF2	Q15554-1	Фактор 2, связывающий теломерные повторы
TOP1	P11387-1	ДНК-топоизомераза 1
WRN	Q14191-1	АТР-зависимая хеликаза синдрома Вернера
XPF	Q92889-1	ДНК-репарационная эндонуклеаза XPF

ванием программы R версии 3.1.1. под пакетом “pvclust” [32]. Для построения кластерной дендрограммы (выбор метода наилучшего кластера проверен с помощью функции *seplot*) использовали метод bootstrap resampling (n = 10000) и кластерный метод ward2. Значения, представленные на каждом узле, – это приближенно несмещенные (approximately unbiased, AU) оценки. Для точного определения кластера в качестве критерия отсечения выбраны значения AU > 95.

Анализ нечетких кластеров. Анализ нечетких кластеров выполнен с использованием программы R, версия 3.1.1, в пакете “cluster” [33]. При использовании подхода с нечеткой кластеризацией точки, расположенные близко к центру кластера, могут находиться в кластере в большей степени, чем точки на краю кластера. Для анализа нечетких кластеров в качестве метрики был выбран SqEuclidean (квадрат стандартного евклидова расстояния). Для сравнения проведен простой анализ главных компонент, PCA, с полными параметрами изменчивости и важности отдельных компонент (**Приложение/Suppl. 9**).

Построение сети взаимодействий. Сеть взаимодействий 18 CBPs построена в рамках программы STRING, версия 10.0 (https://string-db.org/), с критериями по умолчанию для *Homo sapiens* [34]. Для сравнения был создан интерактом 18 случайно выбранных белков с использованием функции *sample* в рамках среды R, примененной ко всему протеому человека по базе данных RefSeq (**Приложение/Suppl. 3A**).

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Анализ аминокислотного состава CBPs человека

Опираясь на ранее опубликованные литературные данные, обобщенные в подробном обзоре Brazda и др. [14], мы выбрали 18 CBPs человека, у которых связывающая активность в отношении крестообразных структур была подтверждена в нескольких экспериментах *in vitro* и *in vivo*. Их аминокислотный состав анализировали с помощью инструмента protParam и сравнивали со средним аминокислотным составом протеома человека. Матрица аминокислотного состава этих CBPs представлена в **Приложении/Suppl. 2.** В последних четырех строках выделены рассчитанные средние и медианные значения для каждой аминокислоты, ожидаемые значения аминокислотного состава и относительное обогащение по сравнению с протеомом человека.

Подробные статистические характеристики (отклонения, выбросы) представлены на диаграммах рис. 1*a*. Исходя из относительного обогащения или истощения, превышающего 10% по сравнению с ожидаемыми значениями для протеома человека, наиболее существенное обогащение обнаружено для аспарагиновой кислоты (D), глутаминовой кислоты (E), лизина (K) и серина

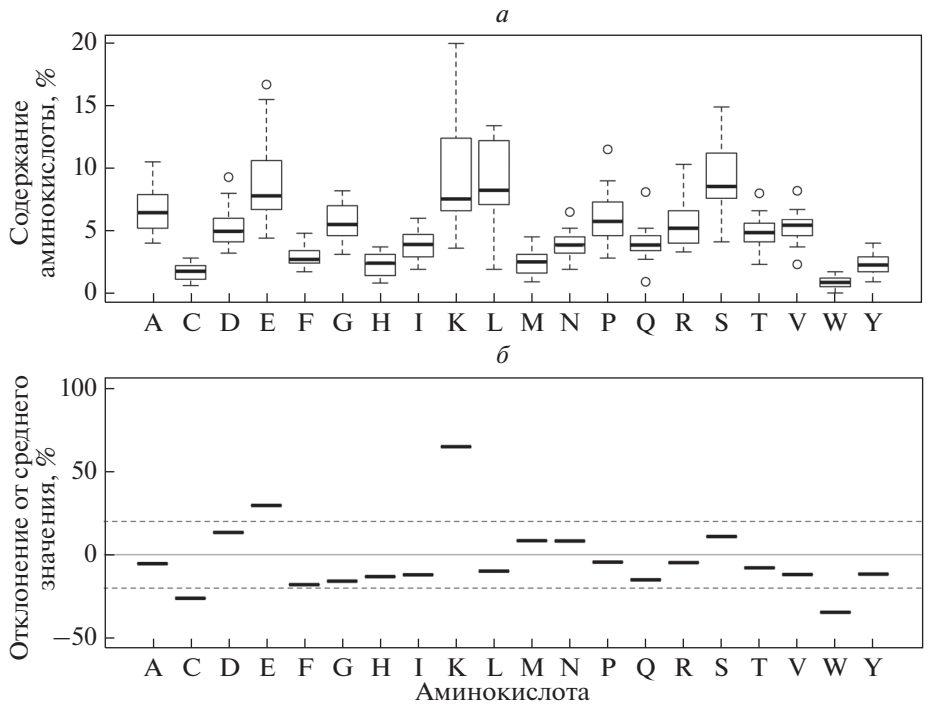

Рис. 1. Различия в аминокислотном составе среди выбранных CBPs. *а* – Диаграммы размаха для оценки аминокислотного состава 18 CBPs. Толстые горизонтальные линии в прямоугольниках соответствуют средним значениям аминокислотного состава CBPs человека. Пустые кружки – это статистические выбросы; например, p53 можно считать выбросом из-за обогащения пролином, а HMGB1 – по истощению глутамина. *б* – Относительные обогащения или истощения 18 CBPs представлены в виде средних значений для каждой аминокислоты, пунктирные линии указывают на 20%-ный уровень обогащения или истощения.

(S), в то время как наиболее заметное истощение выявлено для цистеина (C), фенилаланина (F), глицина (G), гистидина (H), изолейцина (I), глутамина (Q), валина (V), триптофана (W) и тирозина (Y) (рис. 1*б*).

В связи с тем, что эти результаты не отражают тот факт, что у каждой аминокислоты своя дисперсия, сравнение наблюдаемых средних значений аминокислотного состава этих 18 CBPs с ожидаемыми значениями частоты встречаемости аминокислот в протеоме человека не позволило нам провести достоверную и значимую статистическую оценку различий. Мы проанализировали различия в аминокислотном составе с использованием более сложного подхода – программы Composition Profiler. Для этого сравнивали аминокислотный состав тех же 18 CBPs с тремя разными группами белков: (1) случайное подмножество (5000 белков) из протеома человека; (2) относительно небольшая группа хорошо описанных ДНК-связывающих белков (284 белка) и (3) мембранные белки (446 белков), – которые получены из реферируемой базы данных UNIPROT. Относительные величины обогащения или истощения аминокислотного состава CBPs по сравнению с этими белковыми группами показаны на рис. 2. Различия в аминокислотном составе CBPs в сравнении со случайным подмножеством из протеома человека и усредненным аминокислотным составом потеома человека были идентичными (рис. 2*а*). Наибольшие изменения обнаружены для K (обогащение) и W (истощение). Статистически достоверные изменения выявлены для E, K, N, S (обогащение) и для A, C, G, W, Y (истощение). Большинство изменений в аминокислотном составе по сравнению с ДНК-связывающими белками оказались меньше, чем для набора случайных белков (рис. 2*б*). Однако изменения по K, S (обогащение) и A, G, Y (истощение) тоже были значимыми. Сравнение CBPs с мембранными белками дало статистически достоверные изменения для большинства аминокислот, что, вероятно, связано с совершенно определенными функциями белков этой группы (рис. 2*в*). Все идентификационные коды анализируемых CBPs приведены в табл. 1. Подробная статистика обогащений или истощений (p-value) показана в **Приложении/Suppl. 10.** Значительное обогащение по лизину и серину в CBPs обнаружено при всех вариантах сравнения (CBPs против протеома человека, ДНК-связывающих и мембранных белков). Наибольшее относительное обогащение по лизину обнаружено в белках HMGB1 (3.51), DEK (3.14) и TOP1 (3.07). Во всех вариантах сравнения выявлено значительное истощение CBPs по аланину,

Таблица 2. Наиболее длинные паттерны и гомоповторы аминокислотных остатков в анализируемых CBPs

CBP	Паттерны	Гомоповторы (более 3 подряд)
14-3-3s	Не найдено в базе данных	
AF10	–	S4
BRCA1	RRGKKK	K4
DEK	EEEEEEE, EEEED	E4, E7
ESR1	–	A5
HMGB1	EDEREE, EEEEEEE, EEEED, EDDEDED, SKKKK, DEEDE, EEDDD	E4, E5, D4, K4
IFI16	Не найдено в базе данных	
KMT2A	Не найдено в базе данных	
MUS81	–	–
P53	APAPA	–
PARP1	KKSKK	–
PRKDC	–	A4
R51A1	GGSRS, KKEKK	S4
RAD54L	Не найдено в базе данных	
TERF2	–	G4
TOP1	PSPPP, EEEED, KKPKNK, KKEKK	E4, K4
WRN	EEEED, DDDKD, EEDDD	E4
XPF	–	K4

глицину и тирозину. Замечено значительное истощение по триптофану при сравнении со случайным подмножеством и группой мембранных белков. Последний вывод хорошо согласуется с тем фактом, что в специфичной по структуре ДНК-связывающей области BRCA1 нет остатков триптофана (расстояние между двумя ближайшими остатками, W385 и W1508, составляет 1123 аминокислотных остатка). Очень большой интервал между триптофанами также обнаружен в KMT2A (W2056 и W3649, т.е. 1 593 остатка) и в TERF2 (W97 и W453, т.е. 356 остатков). Центральная область белка AF10 также не содержит триптофана (W100 и W724. т.е. разделены 623 остатками). Первые 303 аминокислотных остатка R51A1 не содержат триптофана. Подобное явление наблюдалось в специфичном по структуре ДНК-связывающем домене в С-концевой части белка p53 (среди последних 247 аминокислотных остатков нет ни одного триптофана). В белке DEK, состоящем из 375 аминокислотных остатков, тоже нет остатков триптофана. Наиболее экстремальным примером истощения по триптофану может служить белок IFI16, так как в этом белке, состоящем из 785 аминокислотных остатков, нет ни одного триптофана.

Анализ K-меров показал некоторые интересные факты: почти все CBPs (кроме MUS81) имеют KK-димеры в своих последовательностях. Фактически выявлено очень большое обогащение (162%) по сравнению с набором 248 ДНК-связывающих белков. С другой стороны, для отдельных димеров, в основном IW, CW, WD, WQ и FW, истощение составляло более 70% по сравнению с набором 248 ДНК-связывающих белков. Кроме того, при анализе тримеров выявлено, что в последовательностях большинства CBPs (11/18, включая BRCA1, p53, WRN и IFI16) есть тримеры SSS. Анализ более длинных паттернов аминокислотных остатков и гомоповторов также выявил некоторые интересные факты: белки DEK, HMGB1, TOP1 и WRN содержат один и тот же паттерн EEEED, а BRCA1, HMGB1, KMT2A, TOP1 и XPF – один и тот же гомоповтор K4 (четыре лизина подряд). Полные результаты представлены в табл. 2.

Корреляционный анализ аминокислотного состава CBPs человека

На корреляционной диаграмме (рис. 2*г*) представлены сложные взаимосвязи каждой индивиду-

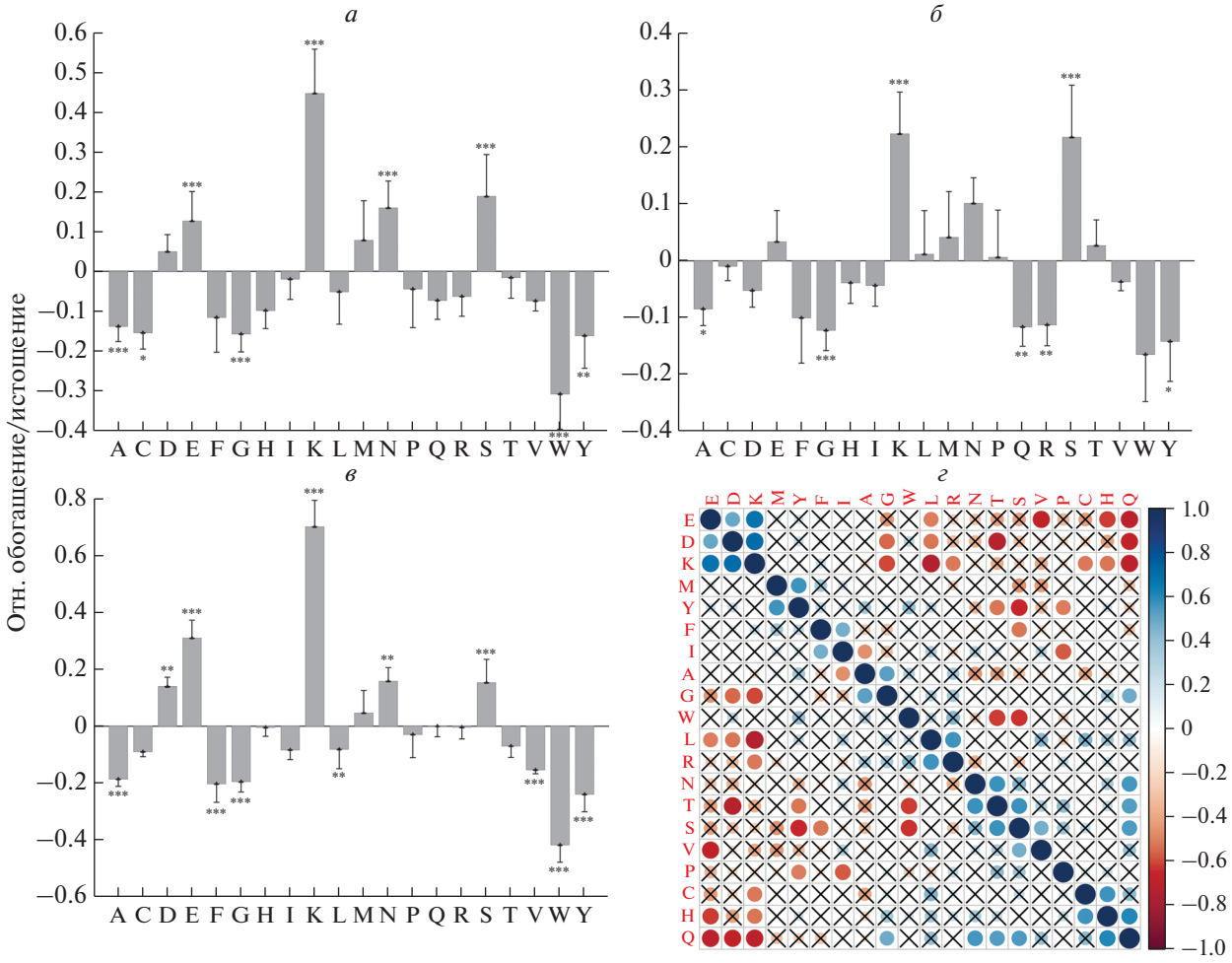

Рис. 2. Сравнение частота встречаемости аминокислот в CBPs и в различных наборах белков. Относительное обогащение или истощение (повышение или понижение содержания аминокислоты) в 18 CBPs по сравнению с 5 000 случайно выбранных белков человека из базы данных UNIPROT (*а*), с 284 ДНК-связывающими белками из базы данных UNIPROT RefSeq (*б*) и с набором из 446 мембранных белков из базы данных UNIPROT RefSeq (*в*). Анализы (*а*, *б*, *в*) выполнены в программе Composition Profiler (10000 итераций методом бутстрэпа с поправкой Бонферрони для тестирования нескольких гипотез). При использовании поправки Бонферрони для множественного тестирования значимыми считались только величины ниже 0.0025 (**P* < 0.0025; ***P* < 0.0010; ****P* < 0.0001). *P*-value приведены в отдельных таблицах в **Приложениях/Supplements**. *г* – Корреляционная диаграмма содержания аминокислот в анализируемых CBPs. Хорошо видно, что лизин положительно коррелирует с глутаминовой и аспарагиновой кислотой (или наоборот). С другой стороны, глутамин отрицательно коррелирует с лизином, глутаминовой и аспарагиновой кислотами, валин отрицательно коррелирует с глутаминовой кислотой, треонин отрицательно коррелирует с аспарагиновой кислотой, а лейцин отрицательно коррелирует с лизином (или наоборот). Незначимые корреляции (*p*-value < 0.05) исключены.

альной аминокислоты со всеми другими аминокислотами в проанализированном наборе данных CBPs. Неудивительно, что содержание положительно заряженной аминокислоты лизина значимо коррелирует с содержанием отрицательно заряженных остатков – глутаминовой и аспарагиновой кислот. Однако это совсем не так для аргинина и гистидина (содержание гистидина даже отрицательно коррелирует с содержанием глутаминовой кислоты). С другой стороны, глутамин отрицательно коррелирует с лизином и глутаминовой/аспарагиновой кислотами. Наконец, валин, треонин и лейцин отрицательно коррелируют соответственно с глутаминовой кислотой, аспарагиновой кислотой и лизином.

Анализ неструктурированных областей

Согласно Tompa [35], частота встречаемости отдельных аминокислотных остатков может служить показателем внутренней неструктурированности (неупорядоченности) белка. Действительно, на основании полученных нами результатов, истощение по цистеину, триптофану и тирозину

и обогащение лизином и глутаминовой кислотой можно рассматривать как показатель увеличения доли внутренних неструктурированных областей в выбранном наборе CBPs. В связи с этим мы проанализировали содержание предсказанной внутренней неупорядоченной области (предсказанной с помощью PrDOS) в выбранных нами CBPs и сравнили его с подмножеством случайно выбранных ДНК-связывающих белков. К нашему удивлению, не обнаружено статистически значимых различий в процентном содержании внутренне неупорядоченных областей в этих двух группах белков (**Приложение/Suppl. 11**). Это может быть связано с тем, что на сегодняшний день для многих ДНК-связывающих белков пока не обнаружена способность связывать неканонические структуры ДНК (в том числе крестообразные); скорее всего, это дело времени. В результате анализа неупорядоченных областей с помощью инструмента IsUnstruct [31] получены сходные данные, за исключением белков 14-3-3σ, HMGB1 и PARP1, которые, согласно предсказанию, оказались примерно в два раза более неупорядоченными при сравнении с данными PrDOS. Для детального графического анализа неупорядоченных областей (предсказанных с помощью IsUnstruct) всех проанализированных CBPs см. **Приложение/Suppl. 12.**

Кластерный анализ

Используя статистическую кластеризацию (R-пакет *pvclust*), мы сравнили CBPs по аминокислотному составу. Основываясь на дендрограмме кластеров (рис. 3*а*), мы смогли достаточно четко различить 3 близкородственных кластера белков на основе приближенно несмещенных значений. В соответствии с этим подходом метод нечеткой кластеризации (рис. 3*б*) показал, что в первый кластер входят, в основном, белки ДНК-репарации (эндонуклеазы MUS81 и XPF), ДНК-зависимые ферменты (ДНК-зависимая киназа PRKDC и ДНК-зависимая трансфераза PARP1) и хеликазы (Rad54, WRN). Во второй кластер попали, в основном, транскрипционные факторы (p53, BRCA1, IFI16, AF10) и коактиватор транскрипции (KMT2A). Наконец, в третьем кластере оказались белки DEK, TOP1 и HMGB1, которые часто связаны с ДНК хроматина и могут взаимодействовать с несколькими факторами транскрипции. Даже если белки DEK и TOP1 имеют разные механизмы изменения плотности суперспирализованной ДНК, они попадают в одну группу в соответствии с кластерной дендрограммой и результатами анализа методом нечетких кластеров (рис. 3*а*,*б*); причем белки DEK, TOP1 и HMGB1 сформировали отдельную группу, отличающуюся от других исследованных CBPs. Такой результат дает основание полагать, что молекулярный механизм, лежащий в основе взаимодействия этих белков с крестообразными структурами ДНК, отличается от других CBPs.

Анализ сети взаимодействий

Поскольку прямые (физические), а также косвенные (функциональные) белок–белковые взаимодействия могут быть показаны базой данных STRING [34], для анализа CBPs мы использовали онлайн-инструмент STRING. В результате обнаружена сильная сеть взаимодействий, объединяющая почти все 18 исследуемых CBPs (рис. 4). Напротив, при анализе контрольного набора из 18 случайно выбранных белков, как и ожидалось, значимых взаимодействий не выявлено.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В последнее время во многих исследованиях выявлена взаимосвязь между функциями белков и их аминокислотным составом. Сообщалось, что на основе аминокислотного состава можно предсказать прионную [38] или противомикробную [20] активность белка. Основываясь на результатах, полученных нами при детальном анализе аминокислотного состава CBPs, можно предполагать наличие более чем одного молекулярного механизма взаимодействия CBP с крестообразными структурами ДНК. Для выяснения точных механизмов взаимодействия между ними потребуется проведение сложных экспериментов. К сожалению, в настоящее время не получено надежных экспериментальных доказательств относительно локусов связывания крестообразных структур ДНК и аминокислот, ответственных за это взаимодействие. В исследовании, проведенном на мышином белке Rif1, показано, что Arg2294 и Lys2303 домена CRII – ключевые аминокислотные остатки при связывании крестообразных ДНК [39]. В дальнейших исследованиях мы планируем идентифицировать аминокислотные остатки, ответственных за связывание крестообразных ДНК, в 18 CBPs, проанализированных нами с использованием биоинформационных подходов, аналогичных подходам Adhikari [40], а также инструменту BindN+ [41]. Интересно также сравнить CBPs человека с другими известными ортологами – чтобы понять, существует ли эволюционно консервативный паттерн предполагаемого связывания крестообразной ДНК с аминокислотными остатками.

Недавно показано, что N-концевая область хеликазы RecQL4 человека внутренне неупорядочена и проявляет высокое сродство к неканоническим структурам ДНК, преимущественно квадруплексным [42]. С-концевой домен p53 также внутренне неупорядочен и проявляет высокую аффинность к неканоническим структурам ДНК [43]. Если проанализировать последние 30 ами-

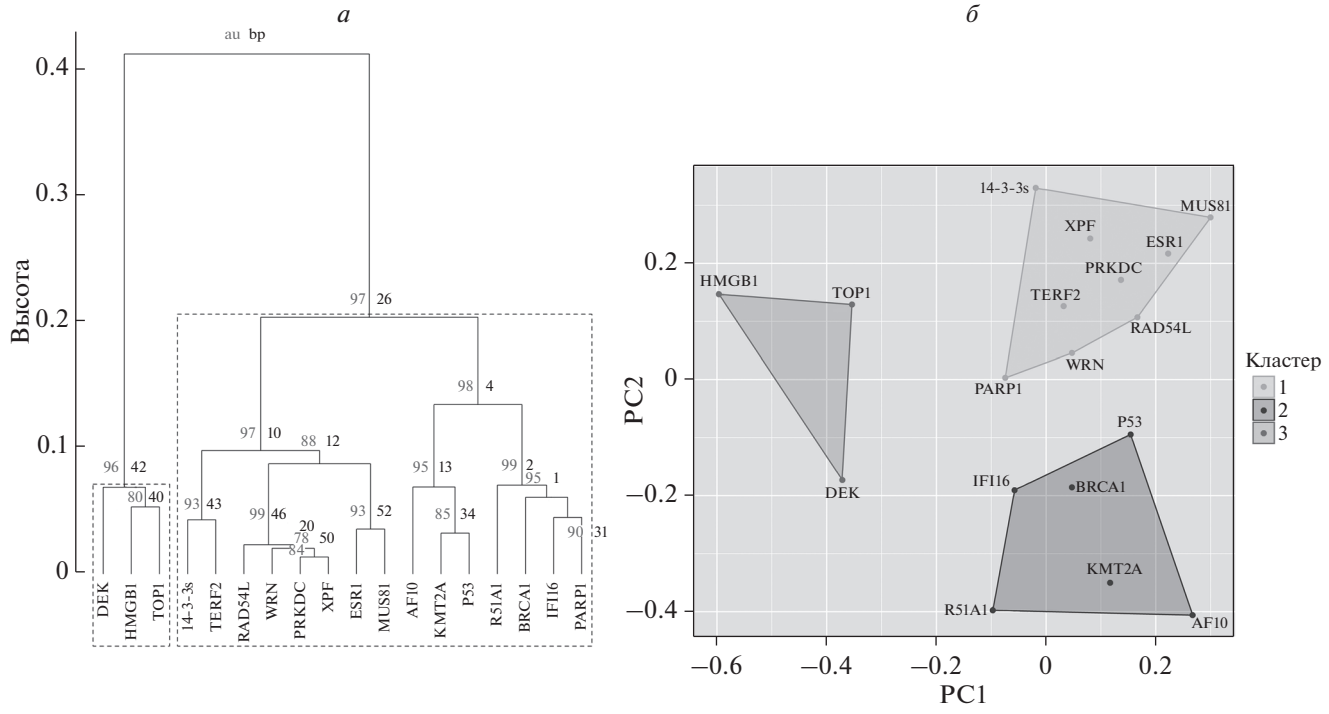

Рис. 3. Анализ сходства CBPs по аминокислотному составу. *а* – Дендрограмма кластерного анализа 18 белков, связывающих крестообразные структуры ДНК, построена на основе их точного аминокислотного состава с использованием R-пакета *pvclust* и bootstrap resampling (n = 10 000) с помощью метода кластерного анализа ward2. Значения, представленные на каждом узле, являются приближенно несмещенными (AU) оценками. Для точного определения кластера в качестве критерия отсечения были выбраны значения AU > 95, два основных кластера отмечены пунктирными линиями. Белки DEK, HMGB1 и TOP1 (левый кластер) образуют группу, совершенно отличную от остальных белков (правый кластер). *б* – Анализ нечетких кластеров 18 CBPs на основе содержания отдельных аминокислотных остатков из табл. 1. Идентификация трех хорошо различимых кластеров полностью соответствует данным, представленным на дендрограмме (вид *а*). Для анализа нечетких кластеров в качестве метрики использован SqEuclidean (квадрат евклидова расстояния). При использовании подхода с нечеткой кластеризацией точки, близкие к центру кластера, могут принадлежать кластеру в большей степени, чем точки на краю кластера.

нокислотных остатков этого белка (364–393), то заметим, что 20% этой области занимают остатки лизина. Недавно показано, что структура, образованная инвертированными повторами в p53-отвечающих элементах ДНК, определяет трансактивационную активность белка p53 *in vivo* [17]. Локальные структуры ДНК, в том числе крестообразные, чувствительны к сверхспирализации ДНК [44, 45]. Важно отметить, что CBPs могут не только стабилизировать крестообразные структуры в ДНК с инвертированными повторами, но и способствовать их формированию [6]. Интересно, что короткие инвертированные повторы часто встречаются в точках транслокационных разрывов при онкологических заболеваниях человека и стимулируют образование двунитевых разрывов ДНК и делеций в клетках млекопитающих и дрожжей [46]. Таким образом, связывание CBPs с этими последовательностями может быть критически важным в процессах регуляции жизнедеятельности здоровой клетки. Существуют компьютерные средства для поиска палиндромов или инвертированных повторов, которые необходимы для формирования крестообразных ДНК-структур, включая Palindrome analyser [47], emboss palindrome [48] и detectIR [49]. Однако, насколько нам известно, программ, способных предсказывать CBP, пока нет. Полученные нами результаты анализа аминокислотного состава CBPs – это только первый шаг к предсказанию неизвестных пока CBPs человека в семействе ДНК-связывающих белков. Для будущих исследований *in vitro* по поиску новых потенциальных CBPs мы можем порекомендовать несколько “горячих кандидатов”: AIM2 (отсутствующий в меланоме 2), MLH1 (MutL гомолог 1) и ARI4A (белок 4A, содержащий AT-богатый интерактивный домен). Белок AIM2 – важный компонент имфламмасомы, который чувствителен к потенциально опасной цитоплазматической ДНК [50]; MLH1–MLH3 играет важную роль в репарации неспаренных оснований ДНК [51], а ARID4A взаимодействует с AT-богатой ДНК (которая часто содержит инвертированные повторы и благодаря этому может образовывать крестообразные формы) и его транскрипция может быть специфически активирована изофор-

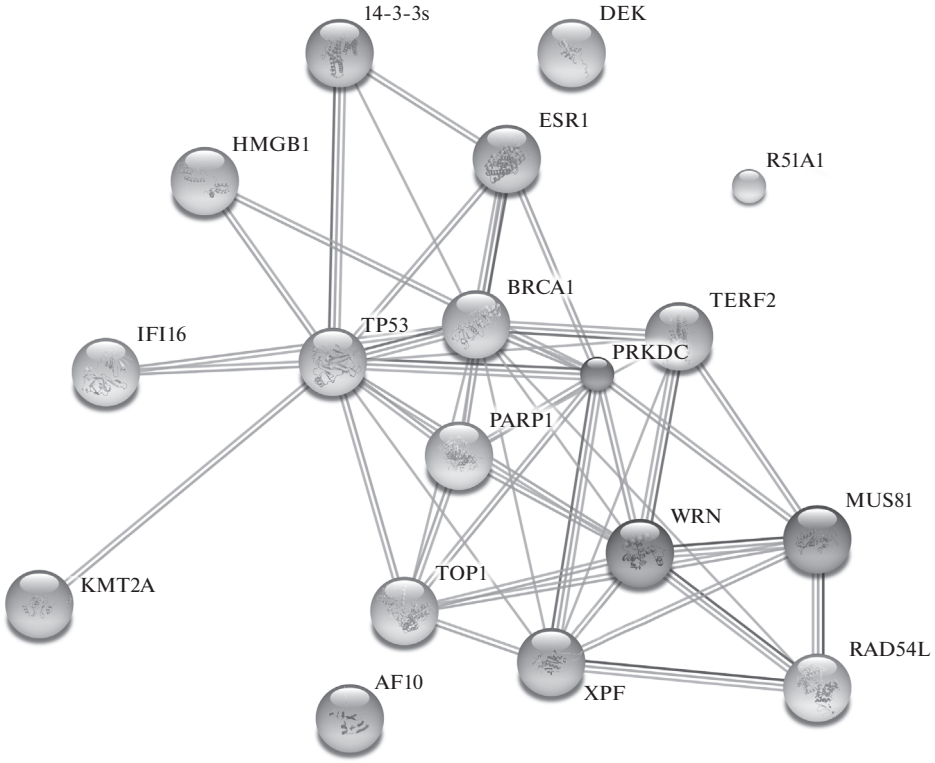

Рис. 4. Сеть взаимодействий 18 CBPs, построенная в STRING. Число линий между отдельными CBPs соответствует разным способам доказательства конкретного взаимодействия и коррелирует с его вероятностью. Линии взаимодействия отражают скорее функциональные, чем прямые физические взаимодействия, хотя некоторые CBPs взаимодействуют и физически (например, p53 с WRN [36] или BRCA1 с ESR1 [37]).

мами рецепторов эстрогена [52]. Аминокислотный состава всех этих "горячих кандидатах" имеет те же особенности, что и проанализированные нами CBPs. Заслуживает внимания обнаруженный нами факт, что белки CBPs содержат протяженные неупорядоченные области. Ранее сообщалось, что неупорядоченные области могут играть важную роль в адаптации белка и могут изменять его укладку [53]. При комплексном анализе неупорядоченных областей выявлено их повышенное содержание в некоторых мотивах аминокислотной последовательности [54]. По-видимому, гибкость неупорядоченного региона играет важную роль в распознавании ДНК. В дальнейшем для предсказания кандидатных CBPs можно использовать так называемый анализ изменений подвижности крестообразных ДНК (Cruciform DNA Mobility Shift Assay) для интерпретации результатов экспериментальных исследований [55]. При проведения этого анализа Stefanovsky & Moss использовали 4 коротких стандартизированных олигонуклеотида, которые при отжиге формировали межмолекулярную крестообразную структуру. Верификация метода проведена на основе ранее известного предпочтительного связывания HMGB-боксов фактора транскрипции UBF с описанной выше крестообразной ДНК при использовании одноцепочечной ДНК в качестве конкурента. Возникает очевидный вопрос, может ли этот метод использоваться для всех CBPs? Ответить на этот вопрос однозначно трудно – нельзя исключить, что некоторые CBPs предпочитают другие типы крестообразных структур (межмолекулярные или внутримолекулярные, внутримолекулярные с большими или малыми боковыми цепями и т.д.).

Большинство известных CBPs, исследованных в этой работе, вовлечены в ключевые биологические процессы. Мутации в CBPs (особенно в белках BRCA1 и p53) ассоциированы с различными онкологическими заболеваниями человека. Эти мутации могут приводить к изменениям в сродстве CBP к регуляторным элементам, содержащим крестообразные структуры. И еще один важный факт: митохондриальная ДНК человека содержит много локусов, потенциально способных образовывать крестообразную ДНК [56], и некоторые из проанализированных нами CBPs (например, p53 [57], TOP1mt [58], HMGB1 [59]) встречаются в митохондриях, а, следовательно, могут играть важную роль в регулировании целостности генома митохондрий или экспрессии генов благодаря их способности связываться с крестообразными структурами. Таким образом,

информация по CBPs может быть чрезвычайно важна для понимания патогенетических процессов при различных заболеваниях человека. Чем больше мы будем знать о CBPs и механизмах их связывания с крестообразными структурами ДНК, тем быстрее поймем роль этих интереснейших белков в различных патологических процессах, в том числе опухолевых. Представленные здесь данные могут быть востребованы для предсказания новых CBPs по первичной структуре.

Работа выполнена при финансовой поддержке Министерства образования, молодежи и спорта Чешской Республики в рамках "Национальной технической программы I" (Ministry of Education, Youth and Sports of the Czech Republic in the "National Feasibility Program I", project LO1208 TEWEP); Оперативной программы структурного финансирования ЕС "Исследования и разработки в области инноваций" (EU structural funding Operational Programme Research and Development for innovation, project No. CZ.1.05/2.1.00/19.0388); Чешского научного фонда (The Czech Science Foundation, 18-15548S) и Остравского университета (University of Ostrava, projects SGS17/PrF/2016 & SGS/17/PrF/2017).

СПИСОК ЛИТЕРАТУРЫ

1. Bochman M.L., Paeschke K., Zakian V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet*. **13**, 770–780.
2. Siddiqui-Jain A., Grand C.L., Bearss D.J., Hurley L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. USA*. **99**, 11593–11598.
3. Wells R.D. (2007) Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci*. **32**, 271–278.
4. Zhao J., Bacolla A., Wang G., Vasquez K.M. (2010) Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci*. **67**, 43–62.
5. Mizuuchi K., Mizuuchi M., Gellert M. (1982) Cruciform structures in palindromic DNA are favored by DNA supercoiling. *J. Mol. Biol*. **156**, 229–243.
6. Chasovskikh S., Dimtchev A., Smulson M., Dritschilo A. (2005) DNA transitions induced by binding of PARP-1 to cruciform structures in supercoiled plasmids. *Cytometry A*. **68**, 21–27.
7. Limanskaya O.Y. (2009) Bioinformatic analysis of inverted repeats of coronaviruses genome. *Biopolymers Cell*. **25**, 307–314.
8. Werbowy K., Cieśliński H., Kur J. (2009) Characterization of a cryptic plasmid pSFKW33 from *Shewanella* sp. 33B. *Plasmid*. **62**, 44–49.
9. Pearson C.E., Zorbas H., Price G.B., Zannis-Hadjopoulos M. (1996) Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *J. Cell. Biochem*. **63**, 1–22.
10. van Holde K., Zlatanova J. (1994) Unusual DNA structures, chromatin and transcription. *Bioessays*. **16**, 59–68.
11. Zannis-Hadjopoulos M., Frappier L., Khoury M., Price G.B. (1988) Effect of anti-cruciform DNA monoclonal antibodies on DNA replication. *EMBO J.* **7**, 1837.
12. Waga S., Mizuno S., Yoshida M. (1990) Chromosomal protein HMG1 removes the transcriptional block caused by the cruciform in supercoiled DNA. *J. Biol. Chem*. **265**, 19424–19428.
13. Alvarez D., Novac O., Callejo M., Ruiz M.T., Price G.B., Zannis-Hadjopoulos M. (2002) 14-3-3σ is a cruciform DNA binding protein and associates in vivo with origins of DNA replication. *J. Cell. Biochem*. **87**, 194–207.
14. Brázda V., Laister R.C., Jagelská E.B., Arrowsmith C. (2011) Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol*. **12**, 33.
15. Bianchi M.E., Beltrame M., Paonessa G. (1989) Specific recognition of cruciform DNA by nuclear protein HMG1. *Science*. **243**, 1056.
16. Waldmann T., Baack M., Richter N., Gruss C. (2003) Structure-specific binding of the proto-oncogene protein DEK to DNA. *Nucleic Acids Res*. **31**, 7003–7010.
17. Brázda V., Čechová J., Battistin M., Coufal J., Jagelská E.B., Raimondi I., Inga A. (2017) The structure formed by inverted repeats in p53 response elements determines the transactivation activity of p53 protein. *Biochem. Biophys. Res. Commun*. **483**, 516–521.
18. Jagelská E.B., Pivoňková H., Fojta M., Brázda V. (2010) The potential of the cruciform structure formation as an important factor influencing p53 sequence-specific binding to natural DNA targets. *Biochem. Biophys. Res. Commun*. **391**, 1409–1414.
19. Cobb A.M., Jackson B.R., Kim E., Bond P.L., Bowater R.P. (2013) Sequence-specific and DNA structure-dependent interactions of Escherichia coli MutS and human p53 with DNA. *Anal. Biochem*. **442**, 51–61.
20. Pane K., Durante L., Crescenzi O., Cafaro V., Pizzo E., Varcamonti M., Zanfardino A., Izzo V., Di Donato A., Notomista E. (2017) Antimicrobial potency of cationic antimicrobial peptides can be predicted from their amino acid composition: Application to the detection of "cryptic" antimicrobial peptides. *J. Theor. Biol*. **419**, 254–265.
21. Settanni G., Zhou J., Suo T., Schöttler S., Landfester K., Schmid F., Mailänder V. (2017) Protein corona composition of poly (ethylene glycol)-and poly (phosphoester)-coated nanoparticles correlates strongly with the amino acid composition of the protein surface. *Nanoscale*. **9**, 2138–2144.
22. Minhas F., Ross E.D., Ben-Hur A. (2017) Amino acid composition predicts prion activity. *PLoS Comp. Biol*. **13**, e1005465.
23. The UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. **45**, D158–D169.
24. Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. (2005) Protein identification and analysis tools on the ExPASy server.

In: *The Proteomics Protocols Handbook*. Ed. Walker J.M. Humana Press, pp. 571–607.

25. Tekaia F., Yeramian E., Dujon B. (2002) Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene*. **297**, 51–60.
26. Vacic V., Uversky V.N., Dunker A.K., Lonardi S. (2007) Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinform*. **8**, 211.
27. Liu B., Liu F., Wang X., Chen J., Fang L., Chou K.-C. (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res*. **43**, W65–W71.
28. Lobanov M.Y., Sokolovskiy I.V., Galzitskaya O.V. (2014) HRaP: database of occurrence of HomoRepeats and patterns in proteomes. *Nucleic Acids Res.* **42**, D273–D278.
29. Wei T., Wei M.T. (2016) Package 'corrplot.' *Statistician*. **56**, 316–324.
30. Ishida T., Kinoshita K. (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res*. **35**, W460–W464.
31. Lobanov M.Y., Sokolovskiy I.V., Galzitskaya O.V. (2013) IsUnstruct: prediction of the residue status to be ordered or disordered in the protein chain by a method based on the Ising model. *J. Biomol. Struct. Dynamics*. **31**, 1034–1043.
32. Suzuki R., Shimodaira H. (2013) Hierarchical clustering with P-values via multiscale bootstrap resampling. *R package.* https://cran.r-project.org/web/packages/pvclust/index.html.
33. Maechler M., Rousseeuw P., Struyf A. (2014) Package 'cluster'. *R package.* https://cran.r-project.org/web/packages/cluster/index.html.
34. Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A., Tsafou K.P., Kuhn M., Bork P., Jensen L.J., von Mering C. (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. **43**, D447–D452.
35. Tompa P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci*. **27**, 527–533.
36. Blander G., Kipnis J., Leal J.F.M., Yu C.-E., Schellenberg G.D., Oren M. (1999) Physical and functional interaction between p53 and the Werner's syndrome protein. *J. Biol. Chemistry*. **274**, 29463–29469.
37. Kawai H., Li H., Chun P., Avraham S., Avraham H.K. (2002) Direct interaction between BRCA1 and the estrogen receptor regulates vascular endothelial growth factor (VEGF) transcription and secretion in breast cancer cells. *Oncogene*. **21**, 7730.
38. Ross E.D., Ben-Hur A. (2017) Amino acid composition predicts prion activity. *PLoS Comput. Biol*. **13**, e1005465.
39. Sukackaite R., Jensen M.R., Mas P.J., Blackledge M., Buonomo S.B., Hart D.J. (2014) Structural and biophysical characterization of murine rif1 C terminus reveals high specificity for DNA cruciform structures. *J. Biol. Chem*. **289**, 13903–13911.
40. Adhikari U.K., Rahman M.M. (2016) *In silico* identification and comparative analyses of active sites of copper containing nitrite reductase (CuNiR) in fungal and bacterial spp. *J. Biol. Eng. Res. Rev.* **3**, 08–18.
41. Wang L., Huang C., Yang M.Q., Yang J.Y. (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol*. **4**, S3.
42. Keller H., Kiosze K., Sachsenweger J., Haumann S., Ohlenschläger O., Nuutinen T., Syväoja J.E., Görlach M., Grosse F., Pospiech H. (2014) The intrinsically disordered amino-terminal region of human RecQL4: multiple DNA-binding domains confer annealing, strand exchange and G4 DNA binding. *Nucleic Acids Res*. **42**, 12614–12627.
43. Laptenko O., Tong D.R., Manfredi J., Prives C. (2016) The tail that wags the dog: how the disordered C-terminal domain controls the transcriptional activities of the p53 tumor-suppressor protein. *Trends Biochem. Sci*. **41**, 1022–1034.
44. Benham C.J., Savitt A.G., Bauer W.R. (2002) Extrusion of an imperfect palindrome to a cruciform in superhelical DNA: complete determination of energetics using a statistical mechanical model. *J. Mol. Biol*. **316**, 563–581.
45. Reddy K., Tam M., Bowater R.P., Barber M., Tomlinson M., Nichol Edamura K., Wang Y.-H., Pearson C.E. (2011) Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. *Nucleic Acids Res*. **39**, 1749–1762.
46. Lu S., Wang G., Bacolla A., Zhao J., Spitser S., Vasquez K.M. (2015) Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. *Cell Rep*. **10**, 1674–1680.
47. Brázda V., Kolomazník J., Lỳsek J., Hároníková L., Coufal J., Šťastnỳ J. (2016) Palindrome analyser–A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem. Biophys. Res. Commun*. **478**, 1739–1745.
48. Faller M. (1999) *Emboss-palindrome*. Online tool. http://www.bioinformatics.nl/cgi-bin/emboss/help/palindrome.
49. Ye C., Ji G., Li L., Liang C. (2014) DetectIR: A novel program for detecting perfect and imperfect inverted repeats using complex numbers and vector calculation. *PloS One*. **9**, e113349.
50. Fernandes-Alnemri T., Yu J.-W., Wu J., Datta P., Alnemri E.S. (2009) AIM2 activates the inflammasome and cell death in response to cytoplasmic DNA. *Nature*. **458**, 509–513.
51. Rogacheva M.V., Manhart C.M., Chen C., Guarne A., Surtees J., Alani E. (2014) Mlh1-Mlh3, a meiotic crossover and DNA mismatch repair factor, is a Msh2-Msh3-stimulated endonuclease. *J. Biol. Chemistry*. **289**, 5664–5673.
52. Monroe D.G., Secreto F.J., Hawse J.R., Subramaniam M., Khosla S., Spelsberg T.C. (2006) Estrogen receptor isoform-specific regulation of the retinoblastoma-binding protein 1 (RBBP1) gene: roles of AF1 and enhancer elements. *J. Biol. Chemistry*. **281**, 28596–28604.
53. Pietrosemoli N., García-Martín J.A., Solano R., Pazos F. (2013) Genome-wide analysis of protein disorder in *Arabidopsis thaliana*: implications for plant environmental adaptation. *PLoS One*. **8**, e55524.

54. Lobanov M.Y., Galzitskaya O.V. (2015) How common is disorder? Occurrence of disordered residues in four domains of life. *Int. J. Mol. Sci*. **16**, 19490–19507.

55. Stefanovsky V.Y., Moss T. (2015) The cruciform DNA mobility shift assay: a tool to study proteins that recognize bent DNA. In: *DNA-Protein Interactions*. Eds Leblanc B.P., Rodrigue S., New York: Springer, pp. 195–203.

56. Čechová J., Lýsek J., Bartas M., Brázda V. (2018) Complex analyses of inverted repeats in mitochondrial genomes revealed their importance and variability. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx729.

57. Yoshida Y., Izumi H., Torigoe T., Ishiguchi H., Itoh H., Kang D., Kohno K. (2003) P53 physically interacts with mitochondrial transcription factor A and differentially regulates binding to damaged DNA. *Cancer Res.* **63**, 3729–3734.

58. Zhang H., Meng L.-H., Pommier Y. (2007) Mitochondrial topoisomerases and alternative splicing of the human TOP1mt gene. *Biochimie*. **89**, 474–481.

59. Ito H., Fujita K., Tagawa K., Chen X., Homma H., Sasabe T., Shimizu J., Shimizu S., Tamura T., Muramatsu S. (2015) HMGB1 facilitates repair of mitochondrial DNA damage and extends the lifespan of mutant ataxin-1 knock-in mice. *EMBO Mol. Med.* **7**, 78–101.

IDENTIFICATION OF DISTINCT AMINO ACID COMPOSITION OF HUMAN CRUCIFORM BINDING PROTEINS

M. Bartas[1], P. Bažantová[1], V. Brázda[2], J. C. Liao[2, 3], J. Červeň[1], P. Pečinka[1, *]

[1]*Department of Biology and Ecology/Institute of Environmental Technologies, Faculty of Science, University of Ostrava, Ostrava, 71000 Czech Republic*

[2]*Institute of Biophysics, Academy of Sciences of the Czech Republic v.v.i., Brno, 61265 Czech Republic*

[3]*School of Medicine, the University of Queensland, Greenslopes Private Hospital, Greenslopes, 4120 Australia*

**e-mail: petr.pecinka@osu.cz*

Cruciform structures are preferential targets for many architectural and regulatory proteins, as well as a number of DNA binding proteins with weak sequence specificity. Some of these proteins are also capable of inducing the formation of cruciform structures upon DNA binding. In this paper we analyzed the amino acid composition of eighteen cruciform binding proteins of *Homo sapiens*. Comparison with general amino acid frequencies in all human proteins revealed unique differences, with notable enrichment for lysine and serine and/or depletion for alanine, glycine, glutamine, arginine, tyrosine and tryptophan residues. Based on bootstrap resampling and fuzzy cluster analysis, multiple molecular mechanisms of interaction with cruciform DNA structures could be suggested, including those involved in DNA repair, transcription and chromatin regulation. The proteins DEK, HMGB1 and TOP1 in particular formed a very distinctive group. Nonetheless, a strong interaction network connecting nearly all the cruciform binding proteins studied was demonstrated. Data reported here will be very useful for future prediction of new cruciform binding proteins or even construction of predictive tool/web-based application.

Keywords: cruciform structures, DNA-protein binding, cluster analysis, lysine, tryptophan