Биомедицинская химия, 2017 том 63, вып. 5, с. 423-427.

ИСПОЛЬЗОВАНИЕ МОЛЕКУЛЯРНЫХ ДЕСКРИПТОРОВ ДЛЯ РАСПОЗНАВАНИЯ САЙТОВ ФОСФОРИЛИРОВАНИЯ В АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ

Д.А. Карасев[1,2]*, П.И. Савосина[1,2], Б.Н. Соболев[1], Д.А. Филимонов[1], А.А. Лагунин[1,2]

[1]Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича, 119121, Москва, ул. Погодинская, 10; эл. почта: w.dmitrykarasev@gmail.com
[2]Российский национальный исследовательский медицинский университет имени Н.И. Пирогова, Москва

Распознавание сайтов фосфорилирования в белках необходимо для реконструкции регуляторных процессов в живых системах. Эта задача осложняется тем, что мотивы фосфорилирования в аминокислотных последовательностях вырождены. Для повышения эффективности предсказания часто используют дополнительные дескрипторы, которые должны отражать физико-химические свойства сайт-содержащих участков. Мы оценили целесообразность такого подхода, применив структурное описание пептидных сегментов с помощью молекулярных дескрипторов MNA. Было проведено сравнительное тестирование с использованием прогностического метода PASS и двух типов входных данных – наборов MNA-дескрипторов, описывающих пептиды как химические структуры, и буквенных символов, характеризующих аминокислотные последовательности этих же пептидов. Обучающие выборки были классифицированы в соответствии с установленным типом модифицирующего фермента (протекинкиназы). Полученные при валидации прогноза оценки точности для разных классов субстратов существенно различались при использовании как буквенных, так и молекулярных дескрипторов. В случае буквенного описания точность прогноза в меньшей степени зависела от длины анализируемых пептидных сегментов в обучающей выборке, тогда как при структурном описании точность определялась размерами пептидов и характеристикой (уровнем) MNA-дескрипторов. Наибольшая точность прогноза специфичности к различным семействам протеинкиназ достигалась при разных размерах молекулярных фрагментов, покрываемых дескрипторами соответствующих уровней. Это, по-видимому, отражало структурные различия в окружении сайтов, модифицируемых протеинкиназами того или иного типа. Применение молекулярных дескрипторов обеспечило результаты прогноза, сопоставимые по точности с результатами, полученными при традиционном буквенном описании. Точность прогноза при высоких значениях меньше зависела от способа описания сайт-содержащих пептидов. В то же время, использование MNA-дескрипторов позволило достигнуть большей точности там, где буквенное описание не обеспечивало приемлемой точности.

Ключевые слова: фосфорилирование белков, молекулярные дескрипторы, аминокислотные последовательности, мотивы фосфорилирования, предсказание сайтов

DOI: 10.18097/PBMC20176305423

ВВЕДЕНИЕ

Посттрансляционное фосфорилирование белков – важное звено таких процессов, как передача клеточных сигналов, модуляция активности ферементов, белок-белковые взаимодействия и др. Нарушения фосфорилирования белков связаны с целым рядом патологических состояний. Сегодня, благодаря развитию высокопроизводительных методов масс-спектрометрии, доступны большие массивы данных по фосфопротеомике. Однако и они недостаточны для реконструкции регуляторных путей, поскольку фосфорилируемые участки могут не распознаваться в силу обратимого и стохастического характера данной реакции. Кроме того, для большей части сайтов фосфорилирования не установлен тип модифицирующего фермента – протеинкиназы, без чего их функциональная интерпретация существенно ограничена. Предсказание сайтов фосфорилирования в аминокислотных последовательностях с учётом специфики модифицирующих ферментов остаётся необходимым этапом в целенаправленном исследовании биологических процессов как в норме, так и при патологии.

Фосфорилирование производится обширным суперсемейством протеинкиназ, которое у человека представлено более чем 500 белками [1]. Каждая из киназ модифицирует свой круг белков-субстратов. Исследование трёхмерных структур показало, что отдельные остатки модифицируемого пептида размещаются в различных участках субстрат-связывающего кармана киназы. Поэтому короткие сегменты последовательностей, включающие фосфосайты, представляют основной тип данных, используемых для обучения в прогностических методах.

Алгоритмы, определяющие фосфосайты в аминокислотных последовательностях, должны выявлять обобщенные паттерны ближнего окружения сайтов – линейные мотивы фосфорилирования. Простейшие способы описания таких мотивов основаны на регулярных выражениях или весовых позиционных матрицах. Используется также машинное обучение – искусственные нейронные сети, метод опорных векторов и др [2-4]. Однако основная проблема заключается в существенной вырожденности коротких функциональных мотивов

* - адресат для переписки

в аминокислотных последовательностях, включая и мотивы фосфорилирования [5]. Это приводит к большому числу ложноположительных результатов при приемлемых уровнях чувствительности прогноза.

Для повышения точности прогноза привлекают сведения по функциональным связям между тестируемыми белками и протеинкиназами для отбора первичных результатов прогноза. Один из наиболее популярных подходов состоит в использовании дополнительных признаков, которые отражают физико-химические свойства аминокислотных остатков, такие как гидрофобность, конформационная подвижность и др. В конечном счёте сайт-содержащий участок последовательности представляется вектором в многомерном пространстве признаков, отражающих положение остатка относительно фосфосайта и упомянутые характеристики [6, 7].

В данной работе мы оценили насколько улучшается распознавание фосфосайтов при использовании вместо буквенного описания аминокислотных остатков их структурных дескрипторов. Дескрипторы напрямую расчитываются из химической структуры фрагментов полипептидной цепи. При этом мы исходили из положения, что химическая структура однозначно определяет физико-химические свойства такого молекулярного сегмента. В нашем подходе структурная формула пептидного участка, включающего фосфосайт, представляется в виде соответствующей структурной формулы, которая описывается с помощью дескрипторов атомных окрестностей (Multilevel Neighborhoods of Atoms, MNA), разработанных нами ранее [8]. Этот способ описания успешно используется при прогнозе биологической активности низкомолекулярных химических соединений на протяжении ряда лет. При этом оценка принадлежности вещества к тому или иному виду биологической активности производится с помощью алгоритма PASS [9-11]. Одна из главных областей применения этой методики – поиск лигандов, специфичных к различным белкам-мишеням, что близко по смыслу к распознаванию пептидов, связываемых модифицирующим ферментом.

МЕТОДИКА

Исследуемая выборка

Протеинкиназы одной таксономической группы (в пределах суперсемейства) характеризуются значительным структурным подобием субстрат-связывающих карманов. Это используется при поиске паттернов общих для семейства или подсемейства ферментов. Исследователи прибегают к такому объединению в связи с небольшим количеством субстратов, установленных для отдельных киназ. Мы также исследовали возможность прогнозирования фосфосайтов, которые модифицируются киназами, отнесенными к определенному семейству.

В работе рассматривались фрагменты белков-субстратов протеинкиназ – пептиды длиной по 5, 7, 9, 11, 13 и-15 остатков, в центральной позиции которых расположены фосфосайты, то есть остатки треонина (Thr), серина (Ser) и тирозина (Tyr), которые подвергаются фосфорилированию.

Информация о фосфосайтах была получена из базы данных PhosphoSitePlus (PSP) (www.phosphosite.org). В исследуемую выборку включались сайты с установленным типом протеинкиназы при условии, что их модификация показана как *in vitro*, так и *in vivo*. Прогноз осуществляли для ряда семейств протеинкиназ (таблица). Таксон (семейство) модифицирующей киназы определяли с помощью сервиса KinomeRender (http://biophys.umontreal.ca/nrg/NRG/KinomeRender.html).

Пептиды, содержащие сайты с установленным типом модифицирующей протеинкиназы, рассматривали как положительные примеры. Набор сайтов с определенной киназной специфичностью составлял не менее 30 сайт-содержащих пептидов. Для формирования выборки “отрицательных примеров” использовали последовательности тех же фосфорилированных белков с той разницей, что отрицательные примеры представляли собой сегменты с остатками серина (Ser), треонина (Thr) и тирозина (Tyr) в центральной позиции, для которых не был установлен факт модифицирования какой-либо протеинкиназой.

Прогноз и валидация

Для предсказания типа модифицирующего фермента использовали алгоритм PASS, основанный на байесовском классификаторе. Структуру пептидов представляли в виде MNA-дескрипторов от 2 до 7 уровней. MNA-дескрипторы основаны на таком представлении структурной формулы, в котором, согласно валентностям и зарядам атомов, явно указаны все атомы водорода и не учитываются типы связей. MNA-дескрипторы для каждого атома молекулы строятся рекурсивно следующим образом:

- MNA-дескриптор 0-го уровня – метка A самого атома;
- MNA-дескриптор любого следующего уровня – условное обозначение структурного фрагмента $A(D_1D_2...D_i...)$, где D_i – MNA-дескриптор предыдущего уровня для i-го непосредственного соседа данного атома с меткой A. Дескрипторы соседей $D_1D_2...D_i...$ записываются в однозначном лексикографическом порядке.

Химические структуры восстанавливались из буквенного представления аминокислотных последовательностей. Дескриптор отображает молекулярный фрагмент, включающий атом и соседние атомы, отделенные от него таким числом химических связей, которое не превышает заданного уровня (рис. 1). Таким образом, все дескрипторы более низких уровней входят в дескриптор данного уровня [12].

Мы использовали также описание сайт-содержащих пептидов с помощью набора пар, каждая из которых включала букву (код аминокислотного остатка) и число (номер позиции в пептиде). Обработка буквенных и структурных дескрипторов с помощью алгоритма PASS позволила сопоставить эффективность их применения.

Таблица. Перекрёстная валидация прогноза сайтов фосфорилирования с исключением по одному. Показаны наибольшие значения инвариантной точности прогноза (IAP), полученные при использовании указанных параметров

Семейство киназ	Тип остатка	MNA		Буквы		IAP-дельта	Число сайтов
		IAP	уровень / рамка	IAP	рамка		
PKC	Thr	0,74	6/9, 7/7, 7/9	0,65	11, 13	0,09	38
IKK	Ser	0,79	7/13	0,72	11	0,07	42
Abl	Tyr	0,72	3/7, 3/9, 3/13, 4/7, 4/9	0,71	7	0,01	49
CDK	Ser	0,95	3/5, 4/5, 5/7	0,95	7	0	141
GSK	Ser	0,88	4/9, 5/9	0,88	11, 13	0	57
MAPK	Ser	0,95	3/5, 4/5	0,95	5	0	171
Src	Tyr	0,75	4/7	0,75	7, 9	0	126
CDK	Thr	0,95	5/5, 5/7, 6/5, 6/7, 7/5, 7/7	0,96	7	-0,01	81
CK2	Ser	0,94	4/7, 5/7, 6/7, 6/11	0,95	7, 9, 11, 13	-0,01	89
STE20	Ser	0,81	2/7, 3/7	0,82	7	-0,01	52
AUR	Ser	0,85	2/5, 5/5, 6/5	0,87	7, 9, 11	-0,02	60
MAPK	Thr	0,9	4/5, 5/5, 6/5, 6/9, 7/5	0,92	13	-0,02	61
PKA	Ser	0,93	5/7, 6/7, 6/9, 7/9	0,95	7, 9, 11, 13, 15	-0,02	112
PLK	Ser	0,73	5/13, 6/13, 6/15	0,75	15	-0,02	68
CAMKL	Ser	0,78	6/11, 7/11	0,81	11, 13	-0,03	66
AKT	Ser	0,93	4/7, 5/7	0,97	11, 13, 15	-0,04	94
PIKK	Ser	0,87	5/5, 6/5	0,91	5, 9	-0,04	93
PKC	Ser	0,8	2/9, 3/9, 4/9, 5/7, 5/9, 6/9	0,85	9, 11	-0,05	90

Примечание. Рамка - длина пептида, включающего модифицируемый остаток в центральной позиции. “IAP-дельта” - разница между величинами IAP, полученных с помощью MNA и буквенных дескрипторов. Приводится число сайтов (положительных примеров), фосфорилируемых киназами указанной группы.

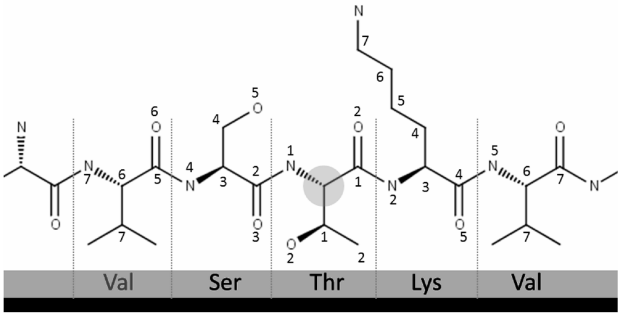

Рисунок 1. MNA-дескрипторы, рассчитанные для Cα-атома остатка Thr в составе центрального остатка сегмента полипептидной цепи. Цифрами отмечены наиболее удаленные атомы, входящие в дескриптор данного уровня. Набор дескрипторов всех уровней генерируется для каждого атома, что позволяет полностью описать структуру пептида.

Программа PASS рассчитывает вероятности принадлежности тестируемого объекта к классу (p_a) и к остальной части обучающей выборки (p_i). С помощью процедуры скользящего контроля с исключением по одному, рассчитывалась инвариантная точность прогноза (IAP) [9], которая совпадает по величине со значением площади под кривой (AUC), вычисляемой при ROC-анализе.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Оценки точности прогноза, полученные при скользящем контроле с исключением по одному, приведены в таблице. В большинстве случаев применение структурных дескрипторов не привело к какому-либо повышению эффективности прогноза. Причём в тех случаях, когда величина IAP составляла не менее 0,95 при использовании буквенных дескрипторов (CDK-Thr, CDK-Ser, MAPK-Ser, PKA-Ser, CK2), он была практически такой же, как и при использовании MNA-дескрипторов, за исключением результатов для выборки AKT-Ser. В последнем случае наблюдалось некоторое падение точности при структурном описании по сравнению с буквенным, но при этом значение IAP сохранялось достаточно высоким. Специфичность сайтов других групп предсказывалась с меньшей точностью, и по большей части была несколько ниже при использовании MNA-дескрипторов. Однако для групп PKC-Thr и IKK-S наблюдалось некоторое улучшение прогноза в случае применения структурных дескрипторов. Таким образом, точность прогноза при высоких значениях меньше зависит от способа описания сайт-содержащих пептидов. В то же время, наибольшее изменение точности прогноза отмечается

именно в тех случаях (PKC-Thr и IKK-Ser), когда точность возрастает от незначительных величин до более приемлемых значений.

Параметры, обеспечивающие максимальную точность, были неодинаковы для различных выборок (таблица). При более детальном рассмотрении результатов эти различия выражены в большей степени.

При буквенном описании точность прогноза (величина IAP) незначительно зависит от размера рамки (длины сайт-содержащих пептидов). Это видно на примере трёх выборок (рис. 2). При определённых значениях параметров (различных для разных семейств), MNA-дескрипторы обеспечивают точность, сопоставимую или более высокую, чем при буквенном описании (таблица, рис. 2). Причем в последнем случае зависимость точности от величины рамки и уровня MNA-дескрипторов носит более выраженный характер. Так для выборки CDK-Ser (рис. 2А) IAP при значениях рамки 5 и 7 и уровнях MNA 3, 4 и 5 совпадает с оценкой, полученной при буквенном описании. При увеличении рамки точность падает независимо от уровня MNA-дескрипторов. Вероятно, в случае серин-содержащих сайтов, аминокислотные остатки, расположенные в непосредственной близости от сайта, наиболее важны для формирования специфичности к киназам семейства CDK.

Для выборки PKC-Thr (рис. 2Б) точность возрастает при увеличении уровня MNA-дескриптора при всех значениях рамки, за исключением рамки 5. Максимальная точность достигается при рамке 7 и 9 и уровне MNA 7, превосходя точность, полученную при буквенном описании. В данном случае учёт вклада более отдаленных фрагментов полипептидной цепи приводит к увеличению точности. Эта тенденция более выражена для выборки IKK-Ser (рис. 2В). Наибольшие значения IAP также, полученные для рамок 11, 13 и 15 и уровня MNA 5, 6 и 7, также превышают показатели точности, достигнутые при буквенном описании.

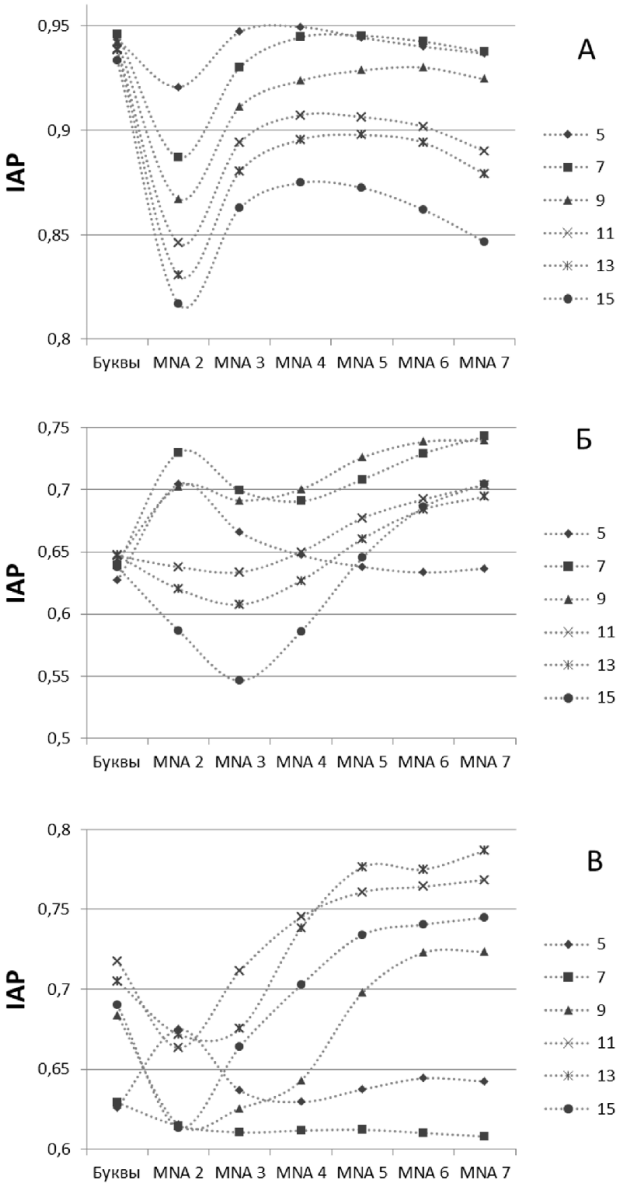

Рисунок 2. Валидация результатов прогноза. Представлены значения инвариантной точности прогноза (IAP) для выборок CDK-Ser (А), PKC-Thr (Б) и IKK-Ser (В) в зависимости от способа описания и уровня MNA-дескрипторов. Значения при разных величинах рамки (длины сайт-содержащих пептидов) показаны отдельными линиями.

ЗАКЛЮЧЕНИЕ

Наши результаты показывают, что применение молекулярных структурных дескрипторов при предсказании сайтов фосфорилирования в последовательностях белков может обеспечить результаты прогноза, сопоставимые по точности с результатами, полученными при традиционном буквенном описанием.

MNA-дескрипторы описывают фрагменты полипептидной цепи, которые не обязательно совпадают с подстроками аминокислотных последовательностей и могут включать уникальные подструктуры на границах аминокислотных остатков. Это может обуславливать то, что точность прогноза в большей мере зависит от длины пептидных сегментов в обучающей выборке, чем в случае буквенного описания. Таким образом, наибольшая точность прогноза специфичности к различным семействам протеинкиназ достигается при разных размерах молекулярных фрагментов, что отражает структурные различия в окружении сайтов, модифицируемых протеинкиназами того или иного типа.

Применение MNA-дескрипторов позволяет достичь большей точности прогноза там, где буквенное описание не может обеспечивать приемлемой точности.

БЛАГОДАРНОСТИ

Работа выполнена в рамках Программы фундаментальных научных исследований государственных академий наук на 2013–2020 гг.

ЛИТЕРАТУРА

1. *Manning G., Whyte D.B., Martinez R., Hunter T., Sudarsanam S.* (2002) Science, **298**(5600), 1912-1934.
2. *Eisenhaber B., Eisenhaber F.* (2010) Methods Mol. Biol., **609**, 365-384.
3. *Palmeri A., Ferrè F., Helmer-Citterich M.* (2014) Front. Genet., **5**, 315.
4. *Trost B., Kusalik A.* (2011) Bioinformatics, **27**, 2927-2935.
5. *Kelil A., Dubreuil B., Levy E.D., Michnick S.W.* (2014) PLoS One, **9**, e106081.
6. *Plewczynski D., Basu S., Saha I.* (2012) Amino Acids, **43**, 573-582.
7. *Song J., Wang H., Wang J., Leier A., Marquez-Lago T., Yang B., Zhang Z., Akutsu T., Webb G.I., Daly R.J.* (2017) Sci. Rep., **7**, 6862.
8. *Filimonov D., Poroikov V., Borodina Yu., Gloriozova T.* (1999) J. Chem. Inf. Comput. Sci., **39**, 666-670.
9. *Filimonov D.A., Poroikov V.V.* (2008) in: Chemoinformatics Approaches to Virtual Screening (Varnek A., Tropsha A., eds.) RSC Publishing: Cambridge, pp. 182-216.
10. *Filimonov D.A., Lagunin A.A., Gloriozova, T.A., Rudik A.V., Druzhilovskii D.S., Pogodin P.V., Poroikov V.V.* (2014) Chem. Heterocycl. Compnds., **50**, 444-457.
11. *Иванов С.М., Лагунин А.А., Захаров А.В., Филимонов Д.А., Поройков В.В.* (2014) Биомед. химия, **60**, 7-17. DOI: 10.18097/PBMC20146001007.
12. *Филимонов Д.А., Поройков В.В.* (2006) Росс. хим. журнал, **50**, 66-75.

Поступила: 31. 08. 2017.
Принята к печати: 14. 09. 2017.

APPLICATION OF MOLECULAR DESCRIPTORS FOR RECOGNITION OF PHOSPHORYLATION SITES IN AMINO ACID SEQUENCES

D.A. Karasev[1,2], P.I. Savosina[1,2], B.N. Sobolev[1], D.A. Filimonov[1], A.A. Lagunin[1,2]

[1]Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; e-mail: w.dmitrykarasev@gmail.com
[2]Pirogov Russian National Research Medical University, Moscow, Russia

Recognition of the phosphorylation sites in proteins is required for reconstruction of regulatory processes in living systems. This task is complicated because the phosphorylation motifs in amino acid sequences are considerably degenerated. To improve the prediction efficacy researchers often use additional descriptors, which should reflect physicochemical features of site-surrounding regions. We have evaluated the reasonability of this approach by applying molecular descriptors (MNA) for structural presentation of the peptide segments. Comparative testing was performed using the prognostic method PASS and two input data types: sets of the MNA descriptors represented peptides as chemical structures and amino acid sequences written using a one-letter code. Training sets were classified in accordance with the established types of the enzymes (protein kinases), modifying corresponding phosphorylation sites. The accuracy estimates obtained by prognosis validation for various classes of substrates were significantly different with both the letters and molecular descriptors. In case of the letter description, the prognosis accuracy demonstrated less dependence on the length of peptides in the training set, while in the case of structural descriptors the accuracy level was determined by the peptide size and descriptor characteristics (MNA levels). The maximal prognosis accuracy related to various kinase families was achieved at different sizes of molecular fragments covered by the MNA descriptors of corresponding levels. This obviously reflected structural differences in surroundings of phosphorylation sites modified by various protein kinases. The use of molecular descriptors provided the prognostic results comparable with the results obtained using traditional letter representation. The prognosis accuracy demonstrated less dependence on the method describing site-surrounding peptides at higher accuracy rates. Applying the MNA descriptors it is possible to achieve better accuracy in the cases when the letter description cannot provide acceptable accuracy.

Key words: protein phosphorylation, molecular descriptors, amino acid sequences, phosphorylation motifs, site prediction