

Katarzyna MROWIEC, Weronika SOKOŁOWSKA, Jakub PETEJA, Roman JAKSIK,
Anna LALIK, Krzysztof FUJAREWICZ
Politechnika Śląska

ANALIZA WPŁYWU FILTRACJI DANYCH NA DZIAŁANIE MODELU IDENTYFIKACJI MIKRORNA WYWIERAJĄCYCH NAJWIĘKSZY WPŁYW NA ZMIANY TRANSKRYPTOMU W ODPOWIEDZI KOMÓREK NA PROMIENIOWANIE JONIZUJĄCE

Streszczenie. Jednym z istotniejszych mechanizmów regulujących potranskrypcyjnie poziom ekspresji genów jest oddziaływanie mRNA z miRNA. W celu identyfikacji miRNA uczestniczących w odpowiedzi komórek na wykorzystywane w radioterapii promieniowanie jonizujące zmodyfikowaliśmy opisany przez nas model matematyczny [7] przystosowując go do danych otrzymanych z sekwencjonowania mRNA i miRNA. W ramach niniejszej pracy prezentujemy wpływ filtracji danych wejściowych na jakość predykcji zmodyfikowanego modelu.

ANALYSIS OF THE IMPACT OF DATA FILTERING ON THE PERFORMANCE OF A MODEL FOR IDENTIFYING MICRORNAS WITH THE GREATEST INFLUENCE ON TRANSCRIPTOME CHANGES IN CELLULAR RESPONSE TO IONIZING RADIATION

Summary. One of the most important mechanisms for regulating gene expression at the post-transcriptional level is the interaction between mRNA and miRNA. To identify the miRNAs involved in the cellular response to ionizing radiation used in radiation therapy, we modified the mathematical model we previously described [7], adapting it to data obtained from mRNA and miRNA sequencing. In this study, we present the impact of input data filtering on the prediction quality of the modified model.

1. Wstęp

Transkryptom jest to ogół cząsteczek RNA występujących w komórce w danym momencie obejmujący zarówno mRNA (RNA kodujące białka), jak i niekodujące cząsteczki RNA (np. tRNA, rRNA czy miRNA) istotne dla prawidłowego funkcjonowania komórki. Oddziaływanie miRNA z mRNA jest jednym z kluczowych mechanizmów potranskrypcyjnej regulacji ekspresji genów. Czynniki środowiskowe, w tym promieniowanie jonizujące, prowadzą do modyfikacji wpływających na funkcję miRNA [1]. Zmiany w oddziaływaniu miRNA z mRNA wpływają na homeostazę komórkową poprzez modulację komórkowej reakcji na stres i przyczyniają się do

powstawania i progresji chorób, takich jak nowotwory. Zmiany w oddziaływaniu miRNA z mRNA mogą wpływać na radiowrażliwość komórek nowotworowych [13], więc zrozumienie wzajemnego oddziaływania miRNA i mRNA ma zasadnicze znaczenie dla rozwoju metod diagnostycznych i terapeutycznych.

Sieć oddziaływań miRNA z mRNA charakteryzuje się tym, że pojedyncze miRNA może regulować wiele celów molekularnych, a pojedyncza sekwencja może być celem wielu różnych miRNA [3]. Złożoność oddziaływań miRNA z mRNA wymaga zaawansowanej analizy bioinformatycznej, w której istotnym elementem są modele matematyczne, wykorzystywane do interpretacji wyników eksperymentalnych, opisu procesów biologicznych oraz identyfikacji kluczowych cząsteczek [11]. Opisane w literaturze modele matematyczne regulacji ekspresji genów przez miRNA opierają się na danych dotyczących ekspresji miRNA i mRNA oraz na ogólnodostępnych algorytmach przewidywania celów miRNA [7] bądź na cechach strukturalnych lub funkcjonalnych RNA [10]. Skuteczność tych modeli w dużym stopniu zależy jednak od jakości danych wejściowych. Dane transkryptomiczne charakteryzują się znaczną zmiennością biologiczną, obecnością szumu technicznego, wartości brakujących oraz mRNA i miRNA o bardzo niskiej ekspresji, dlatego wymagają odpowiedniej kontroli jakości i filtracji [4, 5, 12]. Odpowiednio dobrana strategia filtracji może zwiększyć stabilność i skuteczność modeli predykcyjnych, natomiast zbyt restrykcyjne kryteria mogą prowadzić do utraty biologicznie istotnych informacji [8]. Pomimo rosnącej liczby badań nadal brakuje jednoznacznych wytycznych dotyczących doboru optymalnych metod i progów filtracji oraz ich wpływu na skuteczność modeli matematycznych. W ramach niniejszej pracy oceniono wpływ filtracji danych transkryptomicznych otrzymanych metodą sekwencjonowania (RNA-seq) na działanie zmodyfikowanej wersji modelu identyfikującego miRNA wywierające największy wpływ na zmiany transkryptomu komórek w odpowiedzi na promieniowanie jonizujące [7].

2. Metody

2.1. Hodowla komórkowa

Komórki Me45 (linia komórek ludzkiego czerniaka uzyskana w Narodowym Instytucie Onkologii w Gliwicach) oraz HCT116 (linia komórek ludzkiego raka jelita grubego, ATCC) hodowano w warunkach standardowych. Komórki napromieniano dawką 4 Gy i zbierano próbki po 1 i 12 godzinach od napromienienia. Kontrolę stanowiły komórki nienapromienione inkubowane w takich samych warunkach jak komórki traktowane promieniowaniem IR.

2.2. Sekwencjonowanie RNA

Całkowite RNA wyizolowano przy użyciu zestawu RNeasy Mini Kit (Qiagen) z zastosowaniem etapu trawienia DNazą I (RNase-Free DNase set, Qiagen). Sekwencjonowanie wykonano w firmie Macrogen wykorzystując do przygotowania bibliotek zestaw odczynników TruSeq Stranded mRNA (dla mRNA) oraz NEBNext Small RNA (dla miRNA) oraz platformę NovaSeqX.

W przypadku danych mRNA-seq sekwencje adapterowe i zasady o jakości Phred < 20 usunięto za pomocą programu Trim Galore v0.6.10, odrzucając odczyty krótsze niż 20 nt. Przetworzone odczyty, średnio 130 mln na próbkę, zmapowano do genomu referencyjnego GRCh38 przy użyciu programu STAR v2.7.11b [2] oraz adnotacji GENCODE v49 dla głównego zestawu chromosomów. Średni odsetek zmapowanych odczytów wyniósł 97%. Liczbę odczytów przypisanych do poszczególnych genów wyznaczono za pomocą programu featureCounts z pakietu Subread v1.6.3 [6].

Dane miRNA-seq analizowano za pomocą programu miRge3.0 [9]. Po usunięciu sekwencji adaptorowych odczyty mapowano za pomocą programu Bowtie v1.3.1 do referencyjnych bibliotek ludzkich małych RNA dostarczanych wraz z miRge3.0. Do identyfikacji i ilościowego oznaczania znanych miRNA wykorzystano adnotację miRBase.

2.3. Wstępna selekcja i filtracja danych pochodzących z RNA-seq

W pierwszym etapie wyznaczono wspólny zbiór cząsteczek obejmujący zarówno geny, jak i miRNA, zidentyfikowane na podstawie analizy RNA-seq oraz występujące w bazie interakcji miRNA–mRNA DIANA-microT-CDS. Spośród początkowo zidentyfikowanych 78899 genów oraz 3152 cząsteczek miRNA, do dalszych analiz zakwalifikowano 19519 genów oraz 3152 miRNA. W celu wyznaczenia progów ekspresji genów, dla każdego genu obliczono średnią wartość ekspresji, a następnie wyznaczono wartości percentylowe, które wykorzystano jako progi filtracji. W kolejnych krokach usuwano geny, których poziom ekspresji, wyrażony w jednostkach TPM (ang. *Transcripts per million*), był niższy od przyjętego progu percentylowego. W przypadku danych dotyczących miRNA w pierwszej kolejności usunięto cząsteczki, dla których wartość ekspresji wynosiła 0 (dla Me45 odrzucono 907, a dla HCT116 865 miRNA). Następnie przeprowadzono filtrację polegającą na odrzuceniu określonego odsetka (percentyla) cząsteczek o najniższym poziomie ekspresji.

2.4. Model matematyczny

W pracy wykorzystano zaproponowany przez nas wcześniej matematyczny model interferencji RNA [7] przy czym, w przeciwieństwie do pierwotnej wersji modelu opartej na danych uzyskanych techniką mikromacierzową, do modelu wprowadzono dane o poziomie mRNA (12h po napromienieniu) i miRNA (dla kontroli 1h po napromienieniu) uzyskane metodą RNA-seq. Dodatkowo w zastosowanej przez nas wersji modelu interferencji RNA macierz oddziaływań miRNA–mRNA (macierz c_{ji}) została wyznaczona wyłącznie na podstawie bazy DIANA-microT-CDS (v5, Sept 2013), z której dane zostały pobrane poprzez pakiet multiMiR (v.2.4.0). Baza ta dostarcza zweryfikowanych adnotacji dotyczących interakcji miRNA–mRNA i integruje informacje o miejscach wiązania uwzględniając zarówno dopasowanie sekwencji, jak i cechy strukturalne RNA.

W skrócie: zastosowany przez nas model interferencji RNA opisuje wpływ (negatywną regulację) konkretnego (i-tego) miRNA na poziom konkretnego (j-tego)

mRNA jako złożenie trzech czynników: liczby miejsc wiązania miRNA do mRNA (c_{ji}), poziomu miRNA w komórce ($miRNA_i$) oraz własnej aktywności miRNA (k_i). Wpływ ten można wyrazić wzorem:

$$\Delta mRNA_{j,i} = -c_{ji} k_i miRNA_i \quad (1)$$

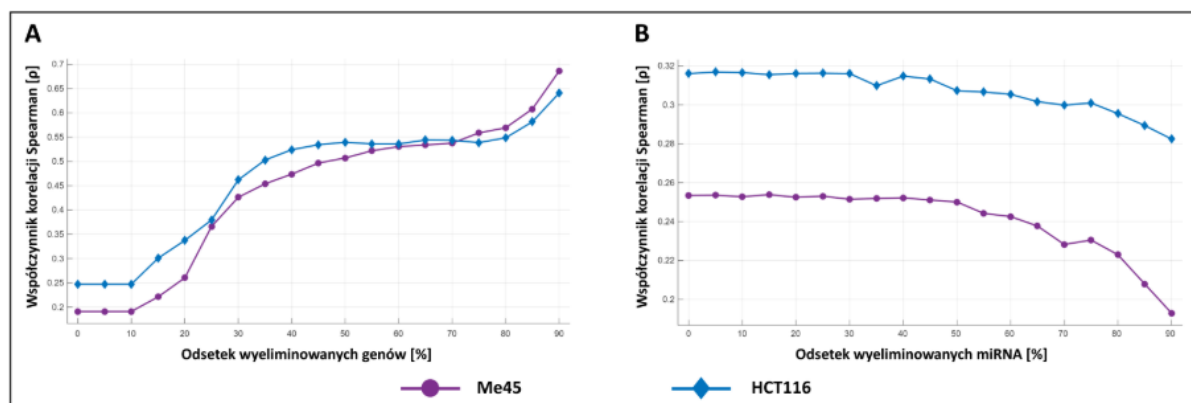
W odpowiedzi komórki na promieniowanie jonizujące zmianie ulega aktywność miRNA, podczas gdy pozostałe parametry traktowane są jako stałe. Różnica aktywności Δk_i prowadzi do zmiany poziomu transkryptu wyrażonej jako fold change (FC_{mRNA}):

$$FC_{mRNA_{j,i}} = c_{ji} \Delta k_i miRNA_i \quad (2)$$

Dopasowanie modelu do danych eksperymentalnych polega na wyznaczeniu współczynników Δk_i , poprzez rozwiązanie nadokreślonego układu równań metodą najmniejszych kwadratów (z wyjątkiem wariantu z najbardziej restrykcyjną filtracją). Wartości tych współczynników umożliwiają identyfikację miRNA o największym znaczeniu w odpowiedzi na promieniowanie.

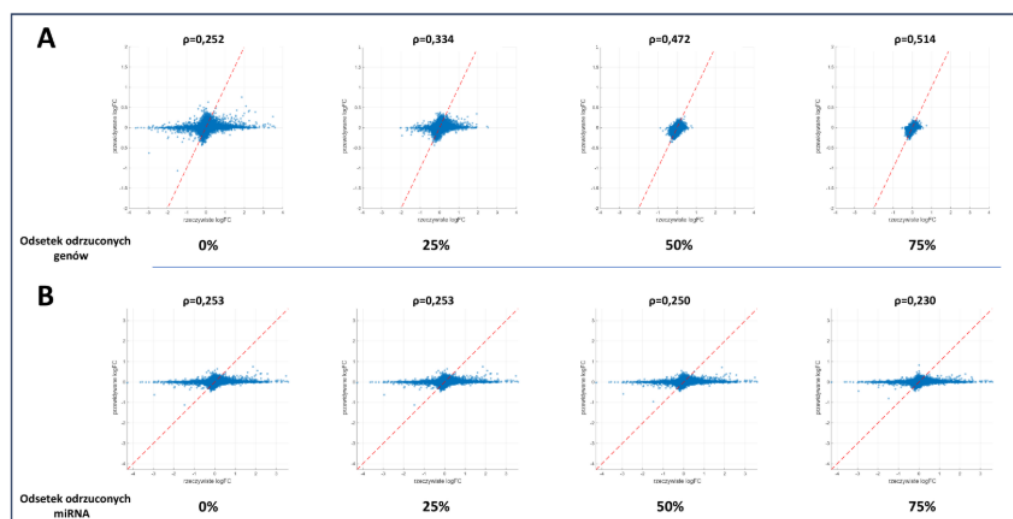
3. Wyniki

Na początkowych etapach filtracji genów (0–10%) korelacja pomiędzy wartościami przewidywanymi przez model a wartościami rzeczywistymi była stosunkowo niska ($\rho \approx 0,19$ dla Me45 oraz $\rho \approx 0,25$ dla HCT116), wskazując na ograniczoną zgodność predykcji z danymi referencyjnymi (rys 1A).

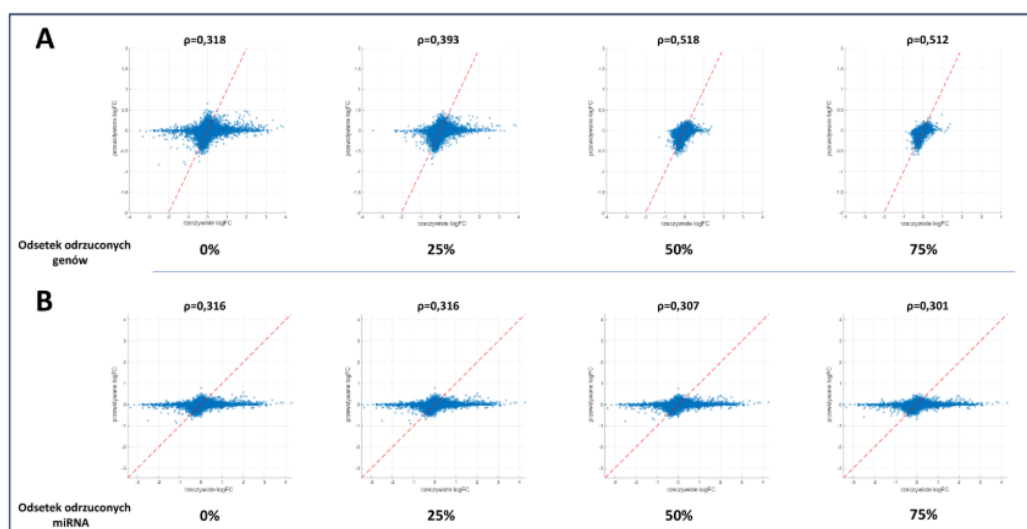


Rys 1. Wpływ filtracji danych uzyskanych dla dwóch linii komórkowych Me45 i HCT116 na działanie modelu. Przedstawiono wykresy zależności: (A) współczynnika korelacji Spearmana w zależności od odsetka wyeliminowanych genów, (B) współczynnika korelacji Spearmana w zależności od odsetka wyeliminowanych miRNA.

Analiza wykresów rozrzutu wykazała znaczny rozrzut pomiędzy wartościami przewidywanymi i rzeczywistymi oraz zawężony zakres przewidywanych wartości logFC względem wartości referencyjnych (rys. 2A, rys. 3A).



Rys 2. Wpływ stopnia filtracji genów i miRNA na zgodność pomiędzy przewidywanymi i rzeczywistymi wartościami logFC dla linii Me45. (A) Wykresy rozrzutu zależności pomiędzy rzeczywistymi i przewidywanymi wartościami logFC dla danych uśrednionych po stopniowym odrzucaniu genów (0%, 25%, 50% i 75%). (B) Wykresy rozrzutu zależności pomiędzy rzeczywistymi i przewidywanymi wartościami logFC dla stopniowej filtracji miRNA (0%, 25%, 50% i 75%). Czerwona, przerywana linia przedstawia linię regresji liniowej, natomiast ρ oznacza współczynnik korelacji Spearmana.



Rys 3. Wpływ stopnia filtracji genów i miRNA na zgodność pomiędzy przewidywanymi i rzeczywistymi wartościami logFC dla linii HCT116. (A) Wykresy rozrzutu zależności pomiędzy rzeczywistymi i przewidywanymi wartościami logFC dla danych uśrednionych po stopniowym odrzucaniu genów (0%, 25%, 50% i 75%). (B) Wykresy rozrzutu zależności pomiędzy rzeczywistymi i przewidywanymi wartościami logFC dla stopniowej filtracji miRNA (0%, 25%, 50% i 75%). Czerwona, przerywana linia przedstawia linię regresji liniowej, natomiast ρ oznacza współczynnik korelacji Spearmana.

Wraz ze zwiększeniem odsetka usuniętych genów do 15–35% zaobserwowano wyraźny wzrost wartości współczynnika korelacji Spearmana dla obu linii komórkowych. Wzrost ten był bardziej dynamiczny dla linii HCT116, która przy eliminacji około 30–35% genów osiągnęła wartość $\rho \approx 0,50$, podczas gdy dla linii Me45 współczynnik korelacji wyniósł około $\rho \approx 0,45$ (rys. 1A). Zwiększeniu wartości współczynnika korelacji towarzyszyło stopniowe zmniejszenie rozrzutu punktów względem linii idealnego dopasowania oraz większa koncentracja obserwacji wokół tej zależności (rys. 2A, rys. 3A). W zakresie 40–75% wyeliminowanych genów wzrost współczynnika korelacji uległ wyhamowaniu, stabilizując się na poziomie około $\rho = 0,53$ – $0,55$. Różnice pomiędzy liniami komórkowymi w tym przedziale były niewielkie (rys. 1A). Reprezentatywne wykresy rozrzutu dla poziomów filtracji 50% i 75% wskazują na wyraźnie lepsze dopasowanie wartości przewidywanych do rzeczywistych w porównaniu z danymi niefiltrowanymi, przy jednoczesnym niewielkim wpływie dalszego zwiększania poziomu filtracji na rozkład punktów (rys. 2A, rys. 3A). Po usunięciu ponad 80% genów odnotowano ponowny wzrost współczynnika korelacji. Najwyższe wartości uzyskano przy progu 90%, osiągając około $\rho = 0,69$ dla linii Me45 oraz około $\rho = 0,64$ dla linii HCT116. Wyniki te wskazują, że pozostawienie jedynie niewielkiej części genów wiązało się z najwyższą zgodnością pomiędzy wartościami przewidywanymi przez model i wartościami referencyjnymi (rys. 1A). W przeciwieństwie do filtracji genów, eliminacja miRNA nie prowadziła do poprawy jakości predykcji modelu. Wartości współczynnika korelacji pozostawały względnie stabilne do około 40–50% usuniętych miRNA, po czym obserwowano ich stopniowy spadek (rys. 1B). Analiza wykresów rozrzutu wykazała jedynie nieznaczne zmiany w rozmieszczeniu punktów wraz ze wzrostem poziomu filtracji, bez wyraźnej poprawy dopasowania do linii idealnej zgodności (rys. 2B, rys. 3B). Ostatecznie, przy 90% eliminacji miRNA, współczynnik korelacji zmniejszył się z około 0,317 do 0,283 dla HCT116 oraz z około 0,253 do 0,193 dla Me45.

4. Podsumowanie

Podsumowując, stopniowa eliminacja genów prowadziła do wzrostu wartości współczynnika korelacji Spearmana, co wskazuje na poprawę jakości predykcyjnej modelu. Sugeruje to, że usuwane geny mogły wnosić do modelu nadmiarowe lub zakłócające informacje. Odmienny efekt zaobserwowano w przypadku eliminacji miRNA, której towarzyszył systematyczny spadek wartości współczynnika korelacji, świadczący o pogorszeniu jakości predykcyjnej modelu. Wyniki te wskazują, że miRNA stanowią istotny element wpływający na jakość predykcji. Ponadto linia HCT116 wykazywała większą stabilność względem eliminacji miRNA niż linia Me45, podczas gdy w przypadku eliminacji genów obie linie charakteryzowały się podobnym trendem wzrostu korelacji, choć ostatecznie wyższe wartości współczynnika osiągnięto dla linii Me45. Należy jednak zachować ostrożność przy interpretacji wyników uzyskanych przy najwyższych poziomach filtracji. Usunięcie nawet 90% cząsteczek prowadzi do znacznej utraty informacji biologicznej. W konsekwencji uzyskana poprawa jakości predykcyjnej modelu po eliminacji dużej liczby cząsteczek nie musi oznaczać, że model lepiej odzwierciedla rzeczywiste zależności biologiczne, ponieważ

usunięte cząsteczki mogły mieć istotne znaczenie w regulacji ekspresji genów, które nie są odzwierciedlone wyłącznie przez wartość współczynnika korelacji.

Praca powstawała w wyniku realizacji projektu badawczego o nr 2022/45/B/NZ2/03599 finansowanego przez Narodowe Centrum Nauki oraz 02/040/BK_26/1080.

LITERATURA

1. Chaiwangyen W., Khantamat O., Kangwan N., Tipsuwan W., de Sousa F. L. P.: MicroRNA expression in response to environmental hazards: Implications for health. *Ecotoxicology and environmental safety*, vol. 300, 2025, s. 118420.
2. Dobin A., Davis C. A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T. R.: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, vol. 29(1), 2013, s.15-21.
3. Hendrickson D., Hogan D., McCullough H., Myers J., Herschlag D., Ferrell J., Brown P.: Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS biology*, vol. 7(11), 2009, e1000238.
4. Jiang P.: Quality control of single-cell RNA-seq. *Computational Methods for Single-Cell Data Analysis*. New York, NY: Springer New York, vol. 1935, 2019, s. 1-9.
5. Leek J. T., Scharpf R. B., Bravo H. C., Simcha D., Langmead B., Johnson W. E., Geman D., Baggerly K., Irizarry R. A.: Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, vol. 11(10), 2010, s. 733-739.
6. Liao Y., Smyth G.K., Shi W.: featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, vol. 30(7), 2014, s. 923-30.
7. Mura M., Jaksik R., Lalik A., Biernacki K., Kimmel M., Rzeszowska-Wolny J., Fajarewicz K.: A mathematical model as a tool to identify microRNAs with highest impact on transcriptome changes. *BMC Genomics*, vol. 20(1), 2019, s. 114.
8. Parkinson E., Liberatore F., Watkins W. J., Andrews R., Edkins S., Hibbert J., Strunk T., Currie A., Ghazal P.: Gene filtering strategies for machine learning guided biomarker discovery using neonatal sepsis RNA-seq data." *Frontiers in genetics*, vol. 14, 2023, 1158352.
9. Patil A. H., Halushka M. K.: Mirge3. 0: A comprehensive microRNA and tRF sequencing analysis pipeline. *NAR Genom Bioinform*, vol.3, 202, lqab068.
10. Shakyawar S., Southekal S., Guda C.: mintRULS: Prediction of miRNA-mRNA Target Site Interactions Using Regularized Least Square Method. *Genes*, 13(9), 2022, s. 1528.
11. Tomlin C., Axelrod J.: Biology by numbers: mathematical modelling in developmental biology. *Nature reviews genetics*, vol. 8(5), 2007, s. 331-340.

12. Wang B., Zhang S. G., Wang X. F., Tan M., Xi Y.: Testing for differentially-expressed microRNAs with errors-in-variables nonparametric regression. *PLoS One*, vol. 7(5), 2012, e37537.
13. Wei W., Ren Y., Lan J., Yi J., Wang M., Zhang Y., Wang S., Xu Y., Han G., Fu Y., You L., Xue J., Ji S., Li, J.: Ionizing radiation: molecular mechanisms, biological effects, and therapeutic targets. *Molecular Biomedicine*, vol. 7(1), 2026, s. 3.