Использование регрессионного анализа для создания эффективных малых РНК, играющих важную роль в регуляции работы генов у животных и растений

Матвеева Ольга, Назипова Нафиса, Огурцов Алексей, Шабалина Светлана

Институт математических проблем биологии РАН, г. Пущино, Россия

НЦБИ, Национальная медицинская библиотека, Национальные институты здравоохранения, г. Бетесда, США

Университет штата Юта, г. Солт-Лейк-Сити, США

Frontiers in Genetics, 2012, Vol. 3, Article No. 163

РНК-интерференция (англ. *RNA interference, **RNAi***) — процесс подавления экспрессии гена или деградации мРНК при помощи малых молекул РНК. Процессы РНК-интерференции обнаружены в клетках многих эукариот: у животных, растений и грибов. Система РНК-интерференции играет важную роль в защите клеток от вирусов, паразитирующих генов (транспозонов), а также в регуляции развития, дифференцировки и экспрессии генов организма.

Открытие феномена РНК-интерференции (RNAi)

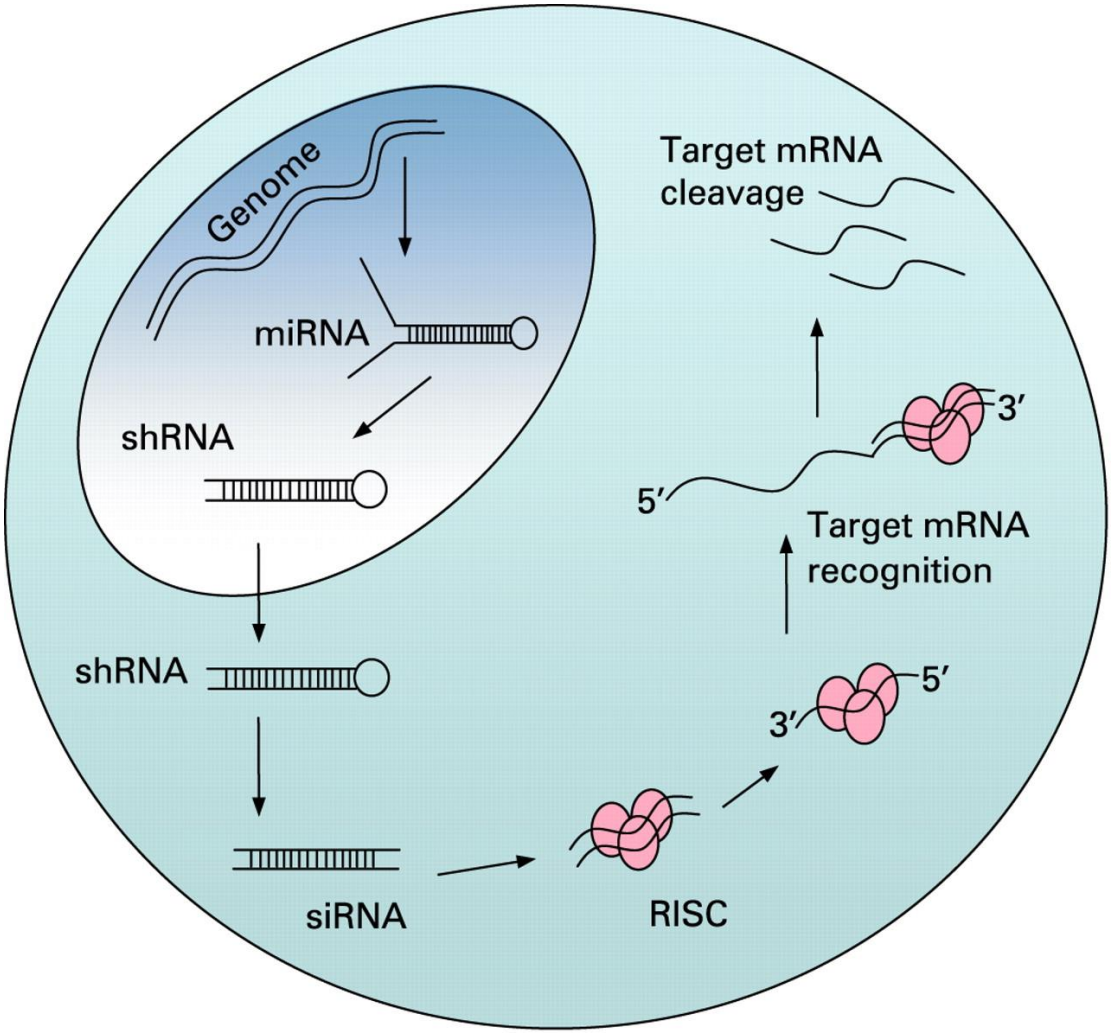
. . .

1990

2000

2010

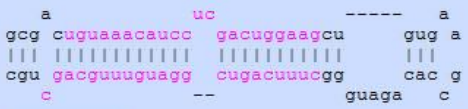
Механизм РНК-интерференции



Lord C J et al. J Clin Pathol 2009;62:195-200

JCP

>hsa-mir-30a MI0000088
GCGACUGUAAACAUCCUCGACUGGAAGCUGUGAAGCCACAGAUGGGCUUUCAGUCGGAUGUUUGC
AGCUGC

Stem-loop	 Get sequence
Deep sequencing	853211 reads, 1.07e+04 reads per million, 79 experiments 
Confidence	Annotation confidence: high Feedback: Do you believe this miRNA is real? Yes (+12) No (-0) Leave comment
Comments	The mature sequences miR-30 [1] and miR-97 [2] appear to originate from the same precursor. Subsequent data confirm that both arms of the precursor appear to give rise to mature miRNA sequences (Pfeffer S, pers. comm.). Landgraf et al. later showed that the 5' product is the predominant one [5]. Related miRNAs are processed from the 5' arms of other precursor loci (mir-30b, MI0000441; mir-30c-1, MI0000736; mir-30c-2, MI0000254; mir-30d, MI0000255; mir-30e, MI0000749).
Genome context	*Coordinates (GRCh38)* chr6: 71403551-71403621 [-]
Database links	EMBL: AF480535; AF480535 EMBL: AF480569; AF480569 EMBL: AJ421752; HSA421752 ENTREZGENE: 407029; MIR30A HGNC: 31624; MIR30A

Mature sequence hsa-miR-30a-5p

Accession	MIMAT0000087
Previous IDs	hsa-miR-30a-5p;hsa-miR-30a
Sequence	6 - uguaaacauccucgacuggaag - 27 Get sequence
Deep sequencing	3520 reads, 61 experiments

Подготовка исходных данных

Исходные данные для анализа - 18719 последовательностей длиной 22 н.п. с известными значениями эффективности сайленсинга (log). Эти данные делились на 2 части: 14972 последовательности для обучения и 3747 для тестирования.

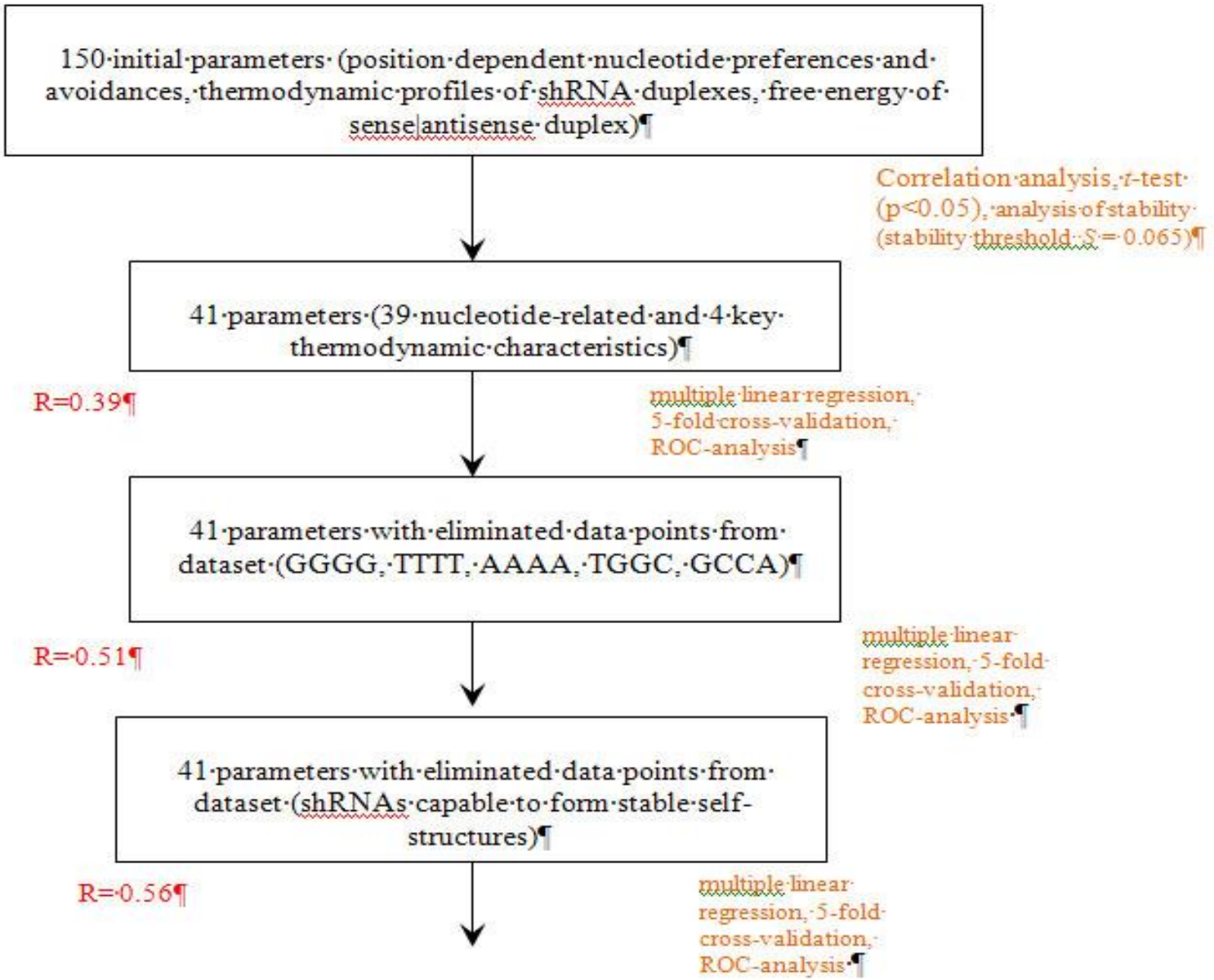
A ⇨ 1,0,0,0; T ⇨ 0,1,0,0; C ⇨ 0,0,1,0; G ⇨ 0,0,0,1.

Около 150 гетерогенных параметров из литературы, нуклеотидного контекста, термодинамических расчетов (OligoHybrid, OligoTherm, Afold). Для выбора наиболее универсальных параметров использовалось 2 критерия:

- значимая корреляция с эффективностью для каждого параметра,
- стабильность корреляции для каждого параметра.

Оба критерия оценивались только для обучающей выборки.

Блок-схема процедуры подбора параметров модели

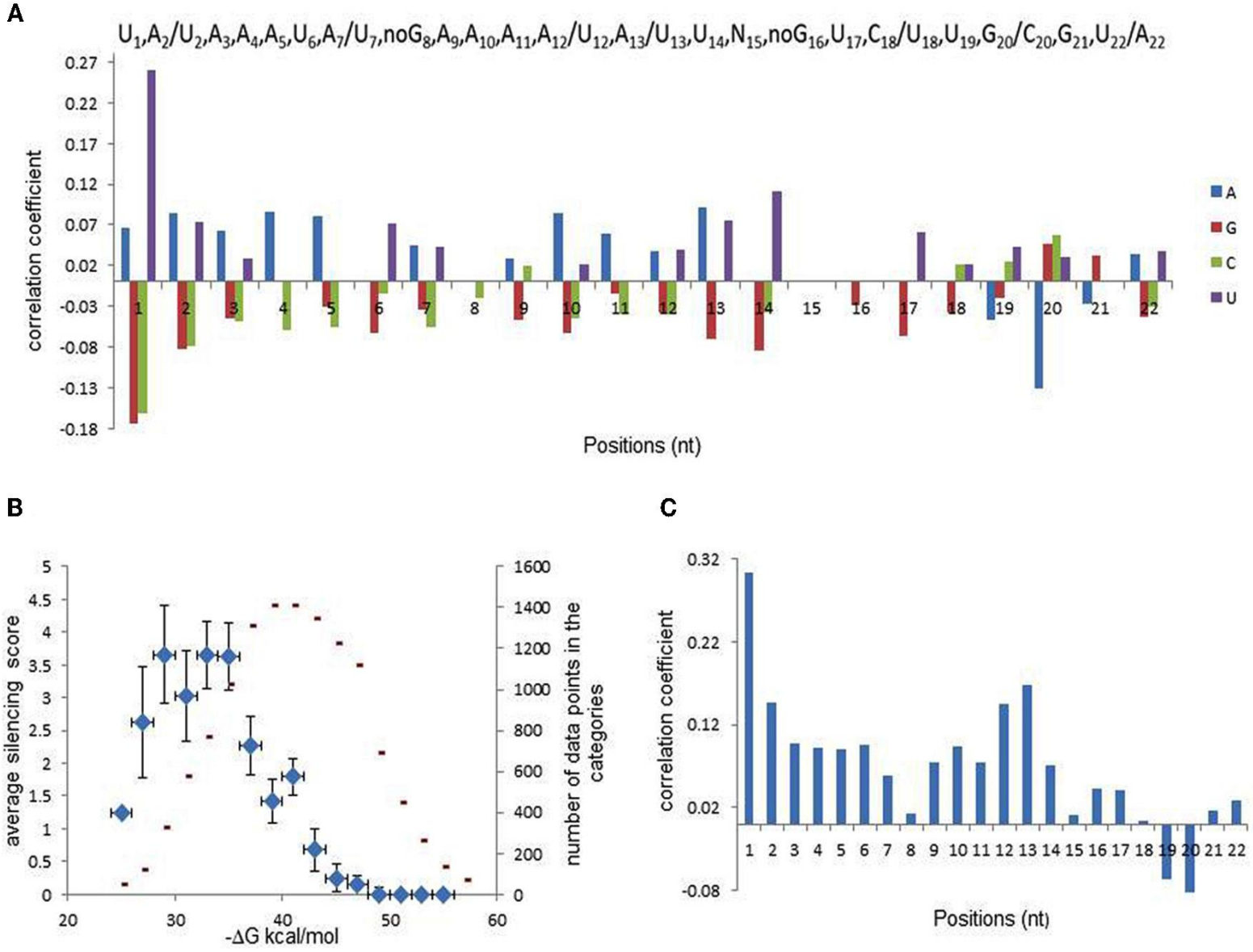


Оценка стабильности корреляции

Оценка стабильности корреляции является индикатором того, насколько предсказательная сила конкретного параметра зависит от выбора подмножества обучающего множества:

$$S_n = \sqrt{\frac{\sum_{i=1}^{m}\left(R_{n_i} - \bar{R}_n\right)^2}{m-1}},\ \bar{R}_n = \frac{\sum_{i=1}^{m} R_{n_i}}{m},\ n = 1,\ldots,1000;\ m = 10$$

Результаты изучения взаимосвязи эффективности и структурных особенностей shRNA



Результат первого этапа уточнения модели. $\Delta G^2, \Delta G^3$ получены программой Afold.

Parameter	Coefficient	*P*-value
Intercept	−43.85	2.046×10^{-178}
ΔG	−2.694	4.459×10^{-111}
ΔG^2	−0.06917	1.716×10^{-105}
ΔG^3	−0.0005959	3.017×10^{-106}
ΔG_1	0.3705	0.00048315
A_1	0.8201	2.61×10^{-06}
A_2	0.8225	1.4939×10^{-11}
A_3	0.6747	0
A_4	0.4644	0
A_5	0.5027	0
A_{11}	0.3433	1.0172×10^{-10}
A_{12}	0.4986	0
A_{13}	0.9504	0
A_{19}	−0.4455	2.1798×10^{-17}
A_{20}	−1.067	2.6228×10^{-98}
A_{22}	0.3039	2.0603×10^{-08}
G_6	−0.3063	1.8561×10^{-08}
G_7	−0.3133	2.2034×10^{-08}
G_9	−0.2439	2.8698×10^{-06}
G_{10}	−0.6907	7.4738×10^{-28}
G_{14}	−0.4776	2.4605×10^{-18}
G_{17}	−0.2477	4.2719×10^{-06}
G_{18}	−0.2378	3.9472×10^{-06}
G_{19}	−0.132	0.0140694
G_{21}	0.3678	1.1793×10^{-12}
C_4	−0.239	9.663×10^{-06}
C_5	−0.3201	3.7375×10^{-09}
C_7	−0.4681	3.9052×10^{-17}
C_8	−0.1913	0.00023711
C_{10}	−0.6128	2.4693×10^{-22}
C_{11}	−0.2256	3.1081×10^{-05}
C_{13}	−0.2326	0.00018904
U_1	1.989	0
U_2	0.6581	4.3741×10^{-07}
U_3	0.5757	0
U_6	0.5343	0
U_{10}	−0.3281	3.3946×10^{-08}
U_{12}	0.5828	0
U_{13}	0.7992	0
U_{14}	0.6866	0
U_{17}	0.3141	1.7814×10^{-09}
U_{22}	0.3012	2.7612×10^{-08}

Множественный коэффициент корреляции

$$R = \sqrt{1 - \frac{\sum_i \left(x_i - y_i\right)^2}{\sum_i \left(x_i - \overline{x}\right)^2}}$$, где

x_i - экспериментальные эффективности олигов,

y_i - предсказанные с помощью модели эффективности олигов,

$\overline{x} = \frac{\sum_i x_i}{n}$ - среднее экспериментальных эффективностей,

n - объем выборки,

R^2 - коэффициент детерминации.

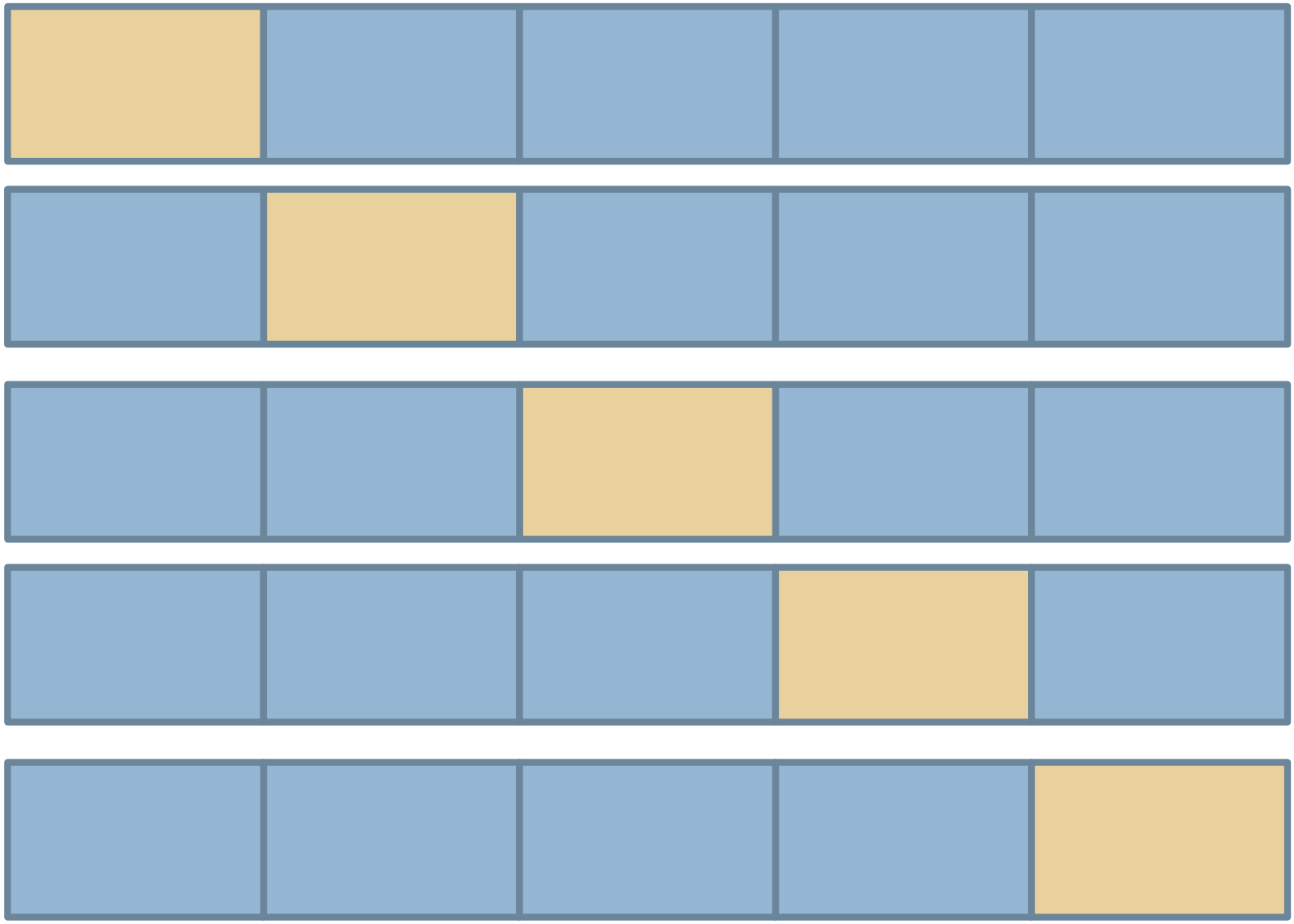
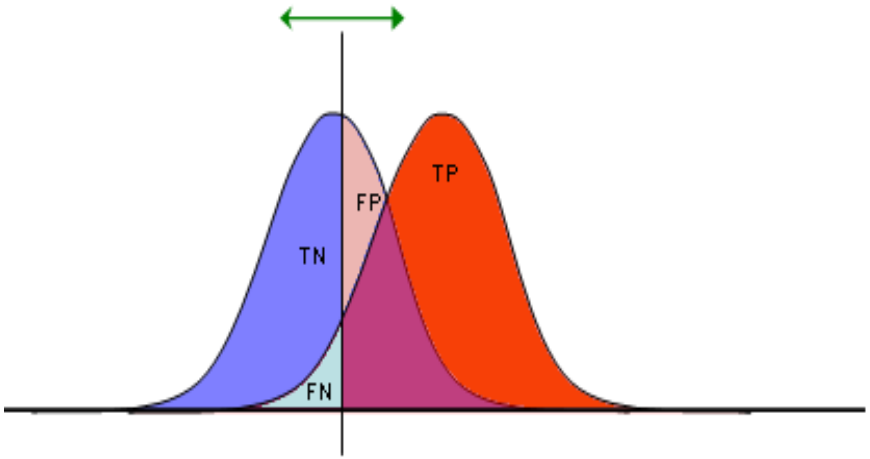
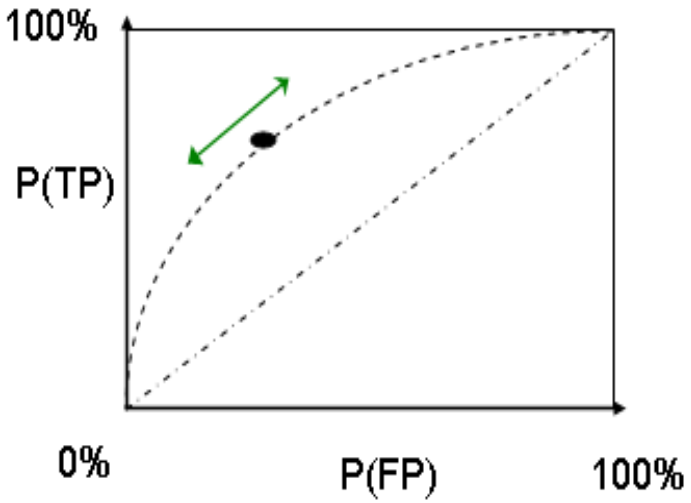


Схема пятикратной перекрестной проверки модели

ROC кривая



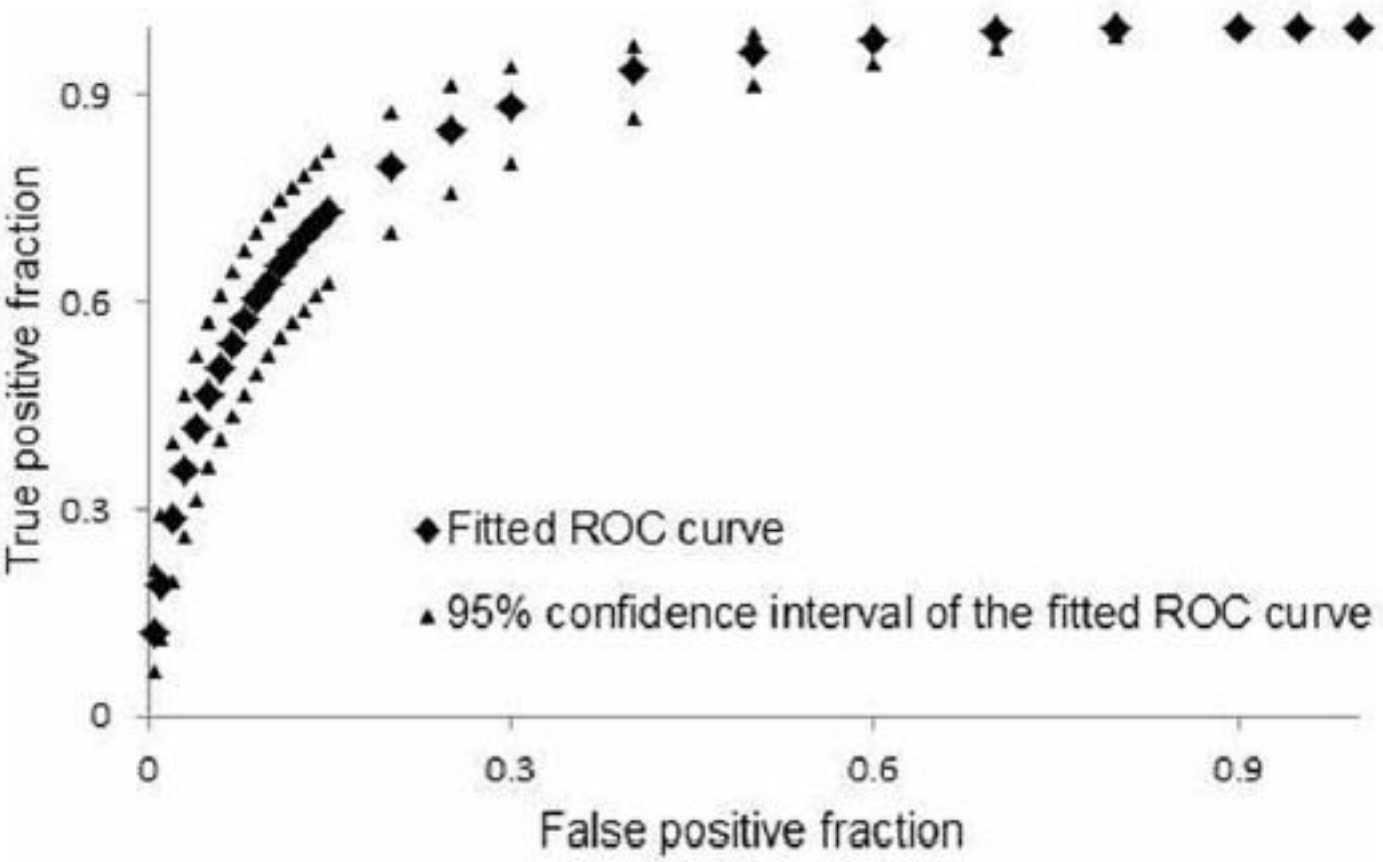
TP	FP
FN	TN
1	1



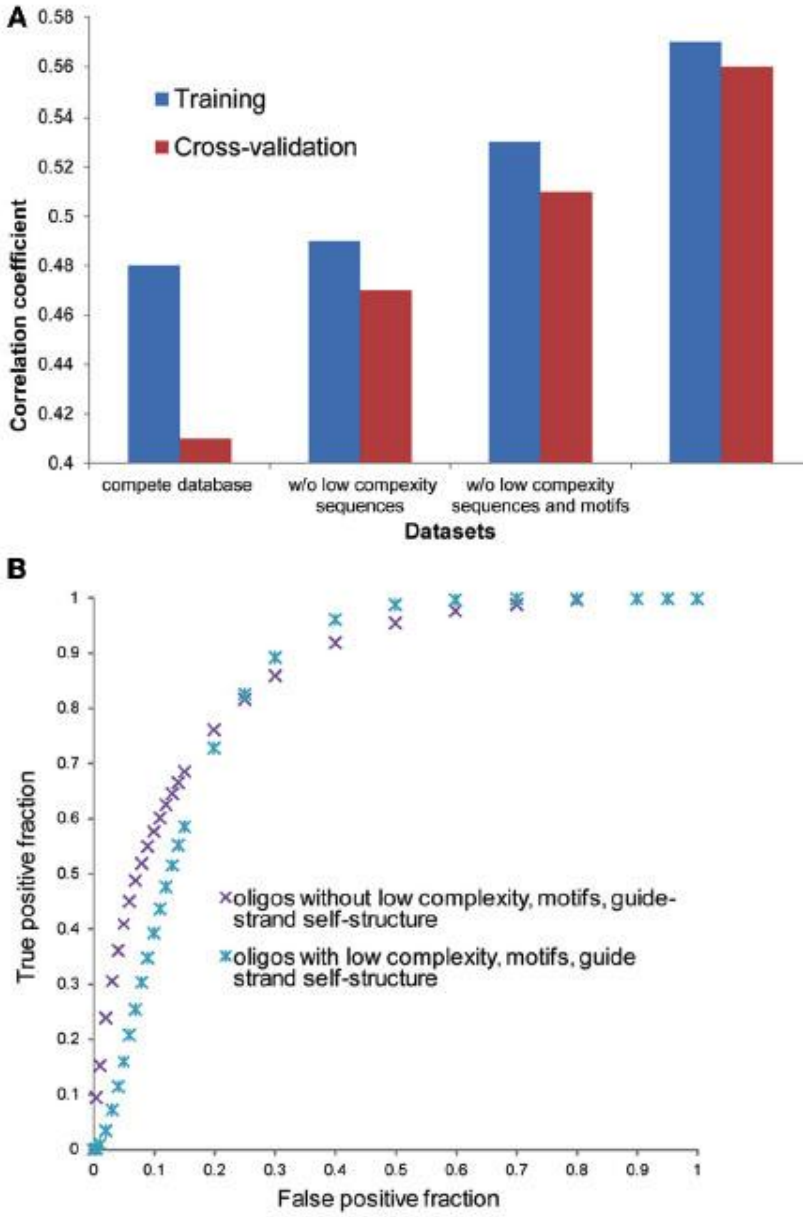
$$Sensitivity = \frac{TP}{TP + FN},$$

$$Specificity = \frac{TN}{TN + FP}$$

ROC-кривая для результатов предсказания эффективности сайленсинга по 41 параметру (0.882)



Результат улучшения предсказательной силы модели после удаления из набора данных простых последовательностей



miR_Scan

www.ncbi.nlm.nih.gov/staff/ogurtsov/projects/mi30/

Работа частично поддержана средствами гранта РФФИ №12-07-00530а