Модель предсказания сайтов связывания транскрипционных факторов

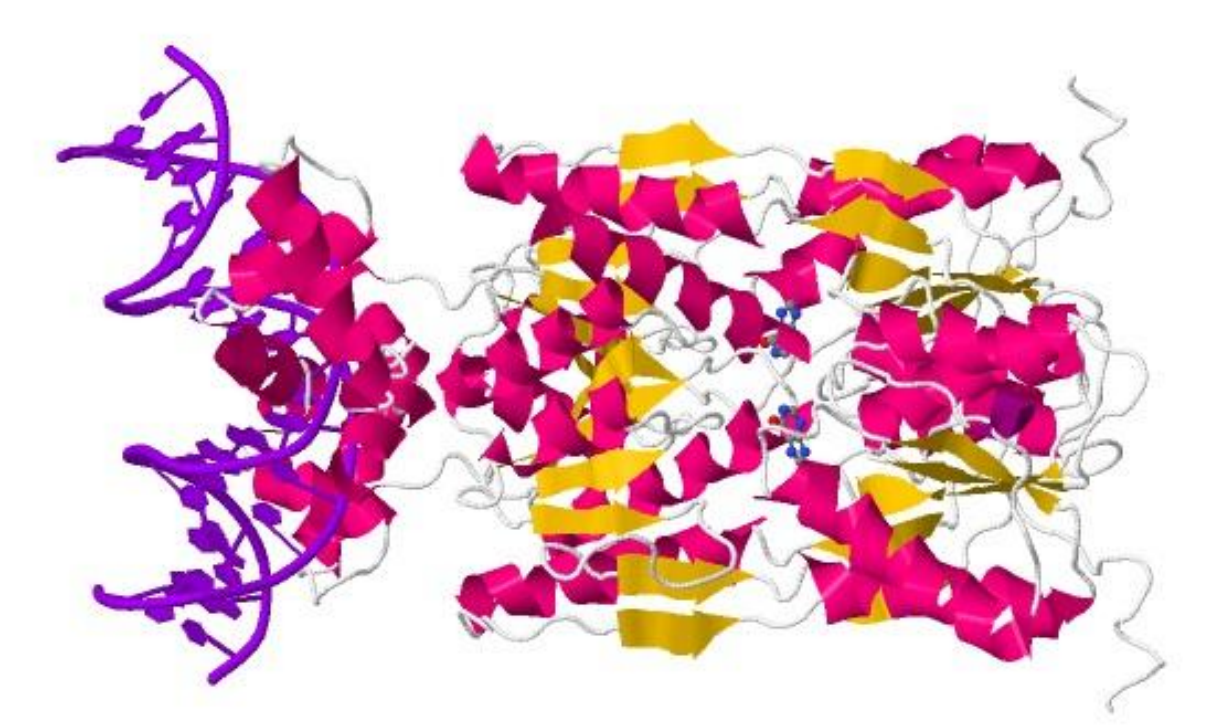
Выполнил: Афанасьев Филипп Александрович

Научный руководитель: Казанов Марат Джамалудинович

План презентации

- Введение
- Методы
- Результаты

Транскрипционные факторы



- ТФ – белок, регулирующий транскрипцию
- Сайт связывания – место контакта ТФ и ДНК

PWM

	Site					
Nucleotide	1	2	3	4	5	6
A	0.73	0.17	0.00	0.00	0.05	0.62
T	0.05	0.26	0.00	1.00	0.49	0.00
C	0.15	0.57	0.00	0.00	0.16	0.22
G	0.07	0.00	1.00	0.00	0.30	0.16

- Вероятность связывания зависит от последовательности нуклеотидов
- Разные позиции в сайте независимы

DWM

Positions	1+2	1+3	1+4
AA	0.12	0.07	0.24
AT	0	0.01	0.03
AG	0.01	0	0
AC	0.02	0	0
TA	0	0	0.01
…			

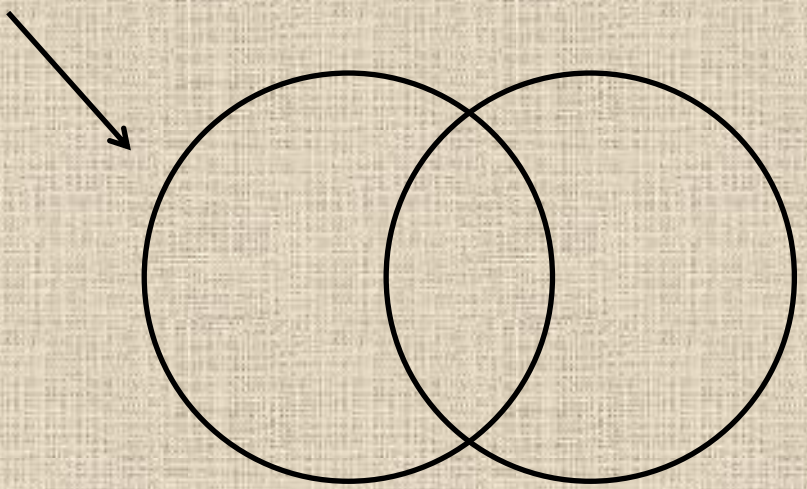
Отбор белков

RegPrecise

RCSB PDB PROTEIN DATA BANK

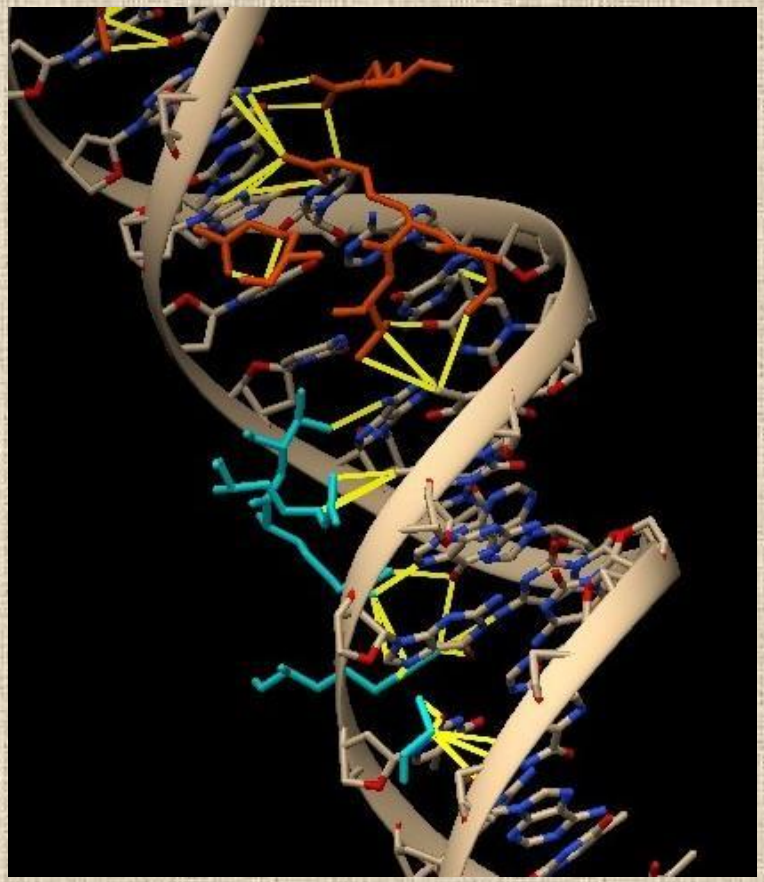
Описанные сайты

Структуры ТФ+ДНК



Пересечение с помощью алгоритма BLAST

Трёхмерные структуры



Контакты между АК и азотистыми основаниями

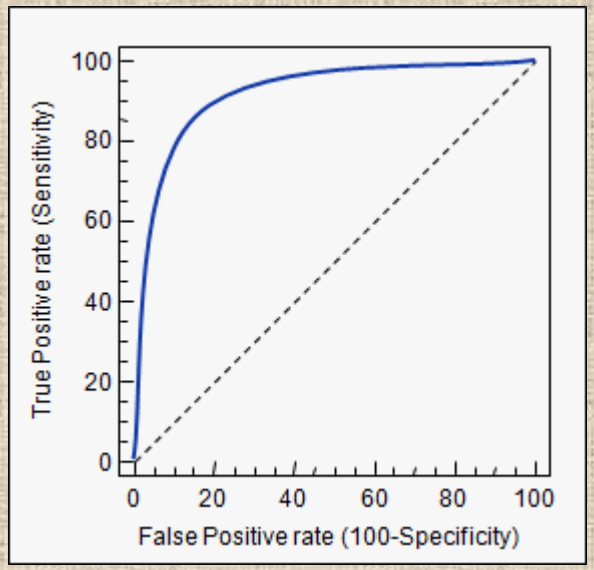
AA	Nuc1	Nuc2	Nuc3
35	4	5	6
44	X9	X10	
45	5	6	
46	X9	X10	
49	7		
65	X4	5	
66	X3	4	

Карта контактов

Качество предсказания

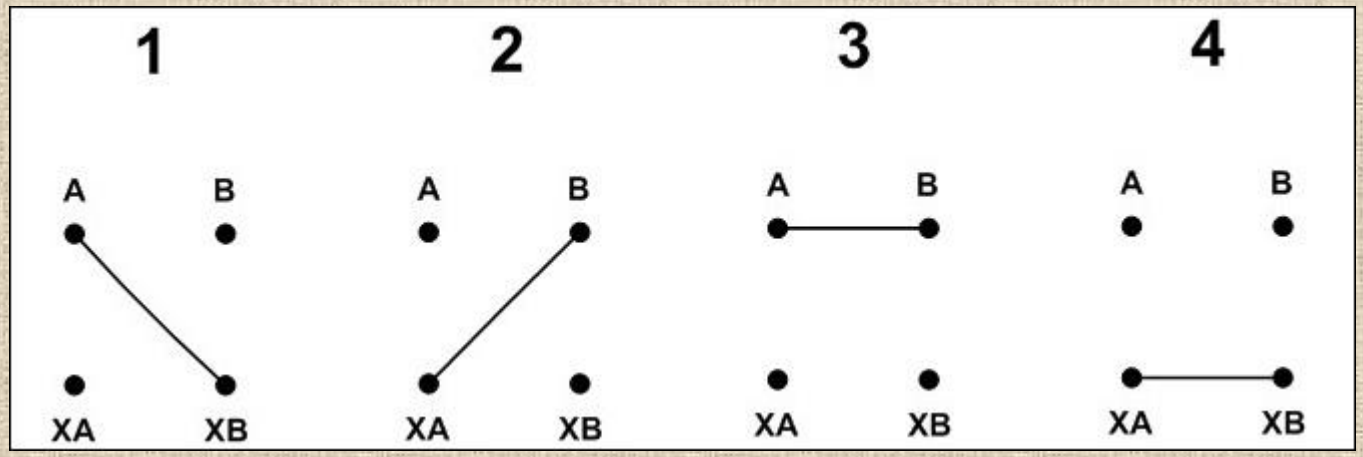
qtcreator_process_stub

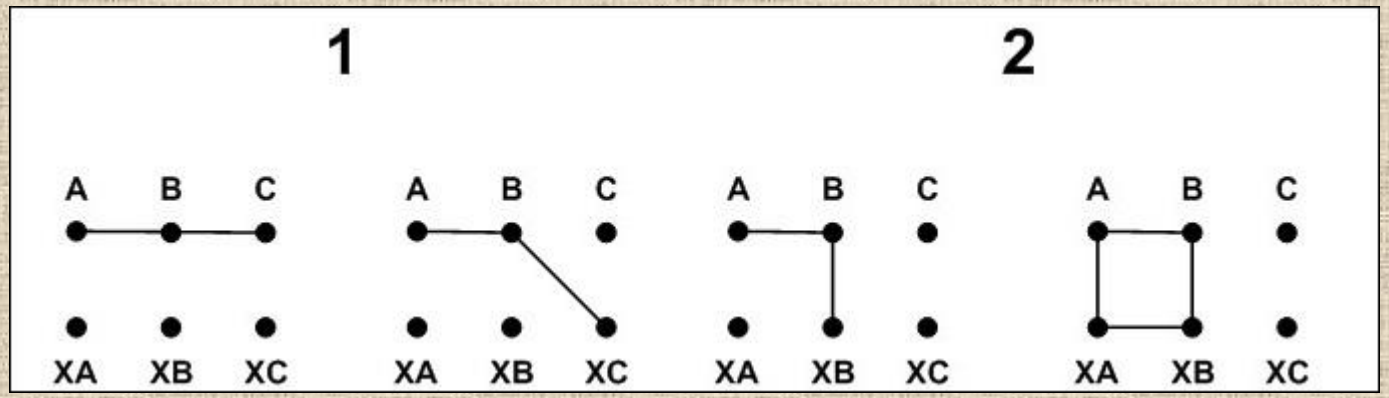
```
125
1348102 ATCTGGTCTGACCTGTT       0.00136678      0
3417399 AACTGGTCTGATTTGTT       0.000251246     0
3672013 AACTCATCCGACCACAT       7.67715e-05     1
2742456 AAGTGGTCTGACCTGAT       4.83949e-05     0
3520188 AACTCGTCTGATTTCTT       4.36918e-05     0
4752450 AACAGATCCGATGAGTT       1.90985e-05     0
2706757 CACAGGTCAGCCCTCTT       9.51919e-06     0
538269  AAGAGGTCATACCAGTT        4.70002e-06     1
2743163 CACTGGTCTGATTTCTA       4.68813e-06     1
185450  AACTGATACTCCCTCTT        3.60911e-06     0
2041490 AACTCATCTTACCAGCT       2.00493e-06     0
1472202 CTCTGGTATGATGAGTC       1.6531e-06      1
3519301 ATGTGGTATTACCTGAT       1.57761e-06     0
1960253 CACTGGTATGCCGACTA       1.19757e-06     0
2131480 CACTGGTCATACGTGAA       9.52138e-07     0
4052205 AACTGATCCGCCCACCC       9.12641e-07     0
496959  ATCAGATCTTCCCTGAT        6.97268e-07     0
630596  AACACGTCAGACGAATA       5.39014e-07     0
3335440 CACAGGTAATATCAGTA       4.21017e-07     0
3035139 ATGTCGTCCGCTTTCTT       3.19963e-07     0
5008194 ATCTCATACGATGAAAT       2.72396e-07     0
1541746 AAGTGATCTGATCAGCA       2.38891e-07     0
4725648 AACTGGTCTTCCCAGCG       2.19483e-07     0
```



AUC = Area under curve

Зависимости





Получение результатов

- Зависимость ⇔ С двумя нуклеотидами связана одна аминокислота
- Перебор всех зависимых пар
- Модификация PWM только для одной пары

Модификация PWM	AUC
—	0.9219
1+2	0.875
2+3	0.9219
3+4	0.75
4+5	0.75
5+6	0.9688
7+8	0.9375
14+15	0.9219
16+17	0.9063
17+18	0.9375
18+19	0.9219
19+20	0.8594
20+21	0.9219

Три положительных правила

Выводы

- ✓ > 50% белков – лучший результат
- ✓ Всю работу может делать машина
- ✓ Быстро
- ✘ Есть значительные ухудшения
- ✘ Мало данных
- ✘ Нужны структуры

Спасибо за внимание!