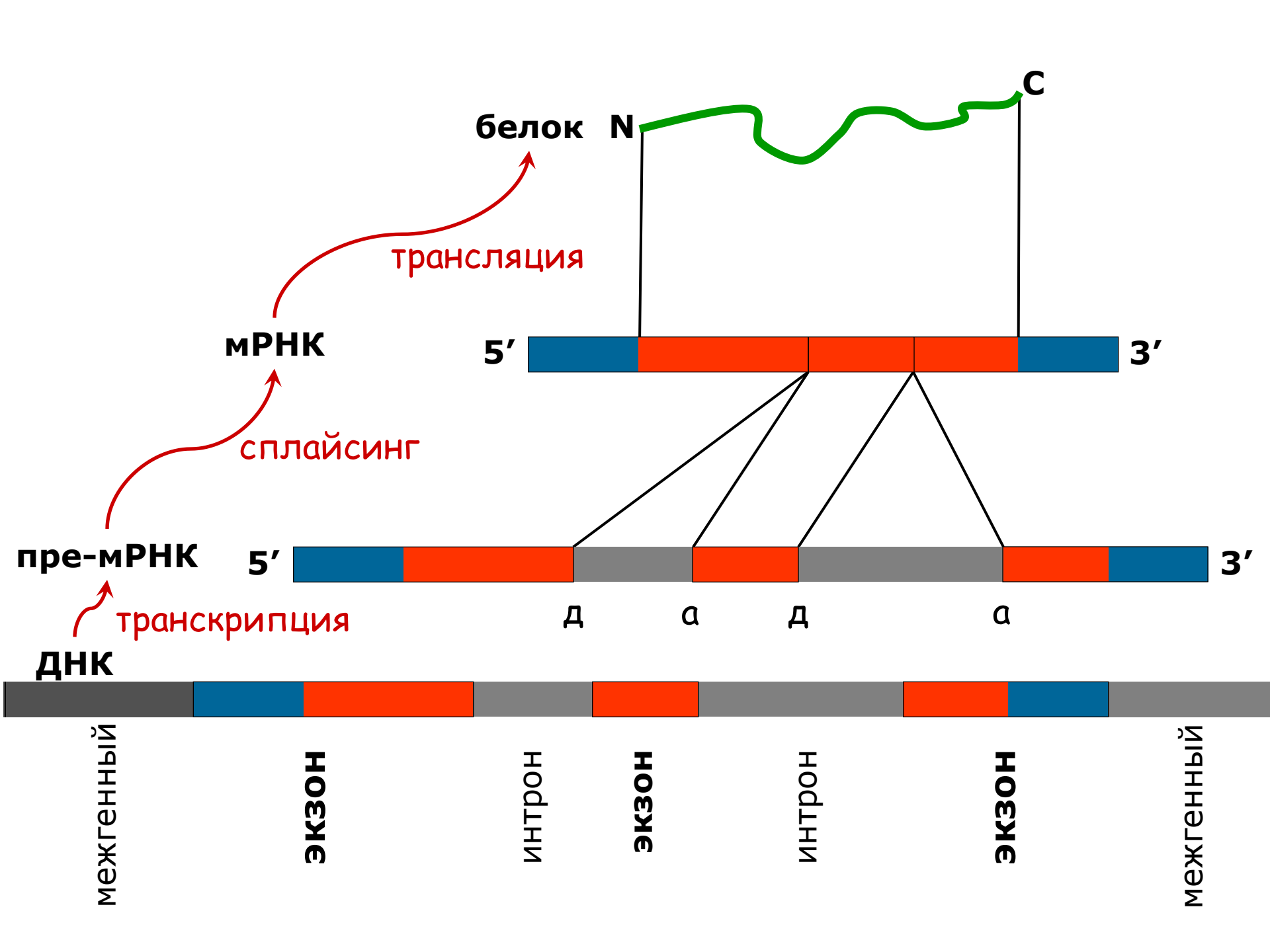
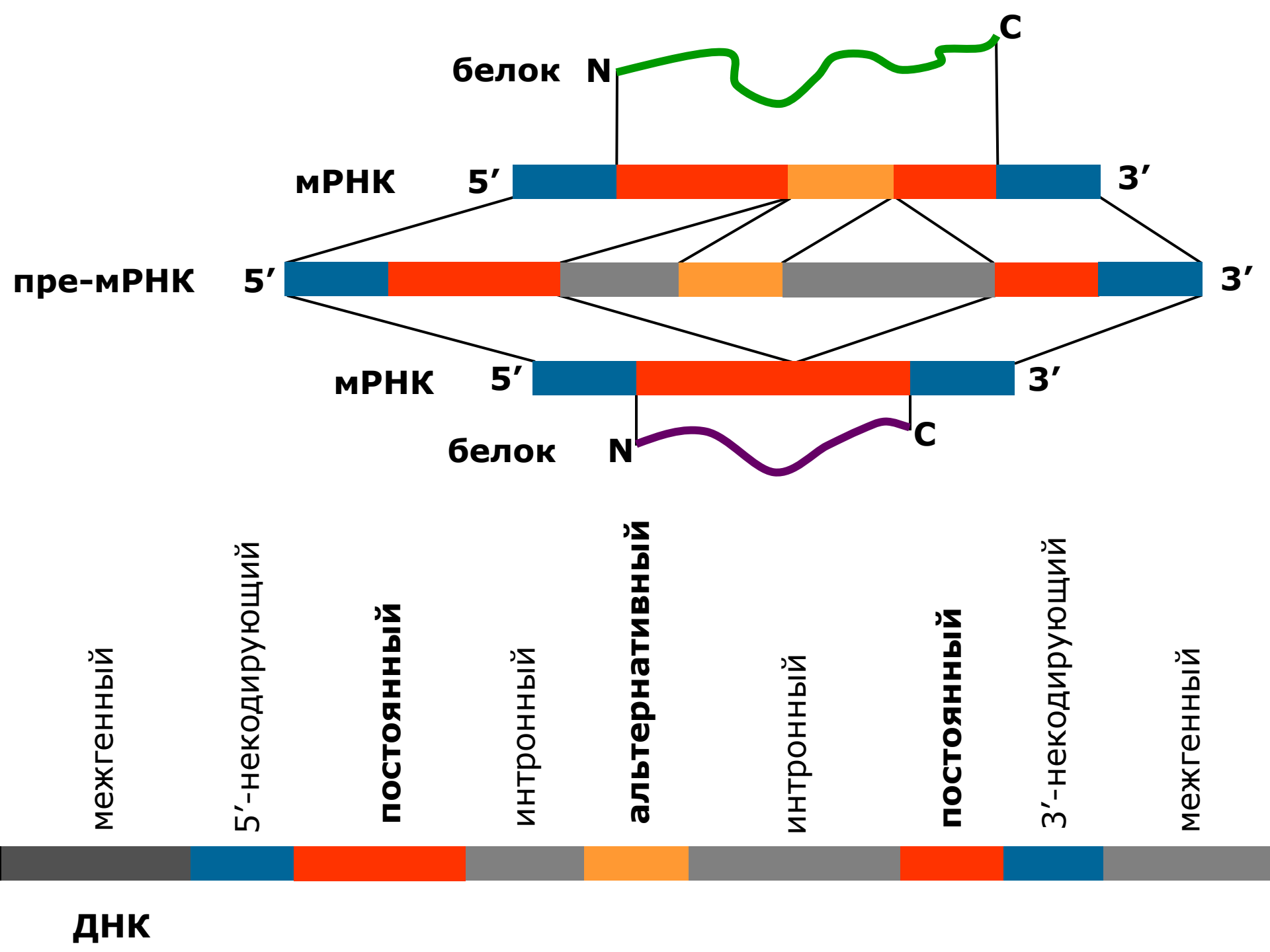
# Распознавание генов

**Факультет биоинженерии и биоинформатики МГУ, второй курс**

М. Гельфанд (лекции)

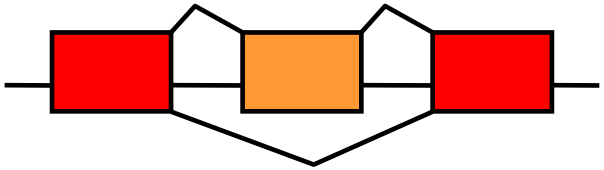
Е.Ермакова (задания, занятия)


белок
N
C
трансляция
мРНК
5′
3′
сплайсинг
пре-мРНК
5′
3′
д
а
д
а
транскрипция
ДНК
межгенный
экзон
интрон
экзон
интрон
экзон
межгенный


белок
N
C
мРНК
5′
3′
пре-мРНК
5′
3′
мРНК
5′
3′
белок
N
C
межгенный
5′-некодирующий
постоянный
интронный
альтернативный
интронный
постоянный
3′-некодирующий
межгенный
ДНК

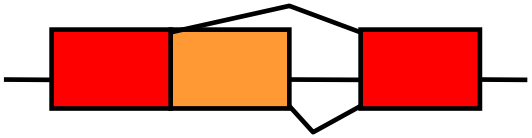
# Типы элементарных альтернатив

кассетный экзон

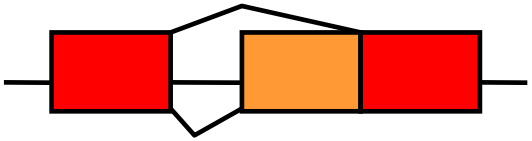


удержанный интрон

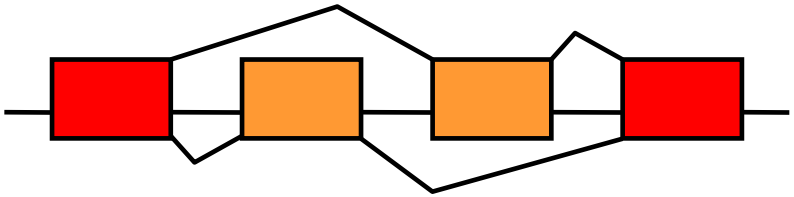
альтернативный донорный сайт



альтернативный акцепторный сайт



чередующиеся экзоны



донорный сайт сплайсинга
акцепторный сайт сплайсинга
5′
GU
AG
3′
экзон
интрон
экзон
-3-2-1 1 2 3 4 5 6

# 0.1% генома E. coli

**Геном бактерии:** несколько миллионов нуклеотидов

От 600 до 9 тысяч генов (примерно 90% генома кодирует белки)

# 0.0001% генома человека

cgtgcacttctgaaggacttcaggtaccggcgtgccccgcgtcctactgtccgcctgctcgcgtcctgggtgccgcctctgagtagggcgggcgaggagg
CAGCCAAGGCGGAGCTGATGGCTGCGCCGAGGGCGGGGCGGGGTGCAGGCTGGAGCCTTCGGGCATGGCGGGCTTTGGGGGGCATTCGCTGGGGGAGGAG
ACCCCGTTTGACCCCTGACCTCCGGGCCCTGCTGACGTCAGGAACTTCTGACCCCCGGGCCCGAGTGACTTATGGGACCCCCAGTCTCTGGGCCCGGTTG
TCTGTTGGGGTCACTGAACCCCGAGCATGCCTGACGTCTGGGACCCCGGGTCCCCGGGCACAACTGACTGCGGTGACCCCAGATACCAGGACCCGGGAGG
CCTCAGAGAACTCTGGAACCCGTTCGCGCGCGTGGCTGGCGGTGGCGCTGGGCGCTGGGGGGGCAGTGCTGTTGTTGTTGTGGGGCGGGGGTCGGGGTCC
TCCGGCCGTCCTCGCCGCCGTCCCTAGCCCGCCGCCCGCTTCTCCCCGGAGTCAGTACAACTTCATCGCAGATGTGGTGGAGAAGACAGCACCTGCCGTG
GTCTATATCGAGATCCTGGACCGgtaatggtgggggtagaccgggaggcactgaagccacaggctggagggcgggcgggtaggaggggtcagagcctcct
cttatctgtgctttccctccatttcagGCACCCTTTCTTGGGCCGCGAGGTCCCTATCTCGAACGGCTCAGGATTCGTGGTGGCTGCCGATGGGCTCATT
GTCACCAACGCCCATGTGGTGGCTGATCGGCGCAGAGTCCGTGTGAGACTGCTAAGCGGCGACACGTATGAGGCCGTGGTCACAGCTGTGGATCCCGTGG
CAGACATCGCAACGCTGAGGATTCAGACTAAGgtgggggctggggtaggccaggtctggttggagctgcttatttgctcgcatcttcagatgacaggtct
cttttacccattctcccttaggagcctctccccacgctgcctctgggacgctcagctgatgtccggcaaggggagtttgttgttgccatgggaagtccct
ttgcactgcagaacacgatcacatccggcattgttagctctgctcagcgtccagccagagacctgggactcccccaaaccaatgtggaatacattcaaac
tgatgcagctattgatgtgcgtcctgataggagagaaatgacaaatgatgggggagggggagaggctgtgtggtacaagcaccaactgatatatggtgg
atgagcctatatagagcttaggctgcaaaaatgtggccacttattcatgggctgagaaagaagagaatttggagaaagtacctacatcctggtatgcccc
cagacttagaatccccagatctctttcatgttttctccttgtcctacagTTTGGAAACTCTGGAGGTCCCCTGGTTAACCTGgtgagtgagacatccttc
cttccaagaatccctgccccaggtcagtgtgggaagggtaggtttcccctaattcaaggatgtttggtcaagtttctgagcagttctttgttggctatct
ctcaatatccaaccagatctccccaacacttgctggtacttttgttcgggtgcccccatcccctactatttgtttaggctagggaactgggggctgtatc
cctgcagGATGGGGAGGTGATTGGAGTGAACACCATGAAGGTCACAGCTGGAATCTCCTTTGCCATCCCTTCTGATCGTCTTCGAGAGTTTCTGCATCGT
GGGGAAAAGAAGAgtgagcctgccttatggggaaacgggttcctttaatgtggtggaaataggggaagggcattcagtgggacttcctggagggtggtct
actgggagaagagggcagggaaggaaggatgtagctgggtggggctcatttgtccctctgtcacagATTCCTCCTCCGGAATCAGTGGGTCCCAGCGGCG
CTACATTGGGGTGATGATGCTGACCCTGAGTCCCAGgtatgagctttagggacagtgacatgtaatgtgaccagtgtaatcagagggggcacctctatt
gagctttgttctcatttctgtctttatctaagatgaactgtgtcacacttgaaataatcacaagagctgtctcccttcatcatcttgactttcttatccc
actccactttgtacacctgtcaccagattgatttcatcctgttactgctttgatttcaagccttcaatccattaacttggcatttaagggccattttcca
tctgtctgtaaatcaactttct tgt
tcttgttttatatggttaatat gct
attgatttaacactgtttgtca cat
tgtattttgtagtacctagccc ttt
ggctaatagggtgatctgtgta aag
tcatcctgggctcccctgcacaccggtgagggagaggctgcagtgtgatatggggatgggcaaggtgtgcatgtgtccttgaactaggctttgtactcct
tcctttctctctgtccatttttctctatagGGCTGGTCTGCGGCCTGGTGATGTGATTTTGGCCATTGGGGAGCAGATGGTACAAAATGCTGAAGATGTT
TATGAAGCTGTTCGAACCCAATCCCAGTTGGCAGTGCAGATCCGGCGGGGACGAGAAACACTGACCTTATATGTGACCCCTGAGGTCACAGAATGAATAG
ATCACCAAGAGTATGAGGCTCCTGCTCTGATTTCCTCCTTGCCTTTCTGGCTGAGGTTCTGAGGGCACCGAGACAGAGGGTTAAATGAACCAGTGGGGGC
AGGTCCCTCCAACCACCAGCACTGACTCCTGGGCTCTGAAgAATCACAGAAACACTTTTTATATAAAATAAAATTATACCTAGCaacatattatagtaaa

**Геном человека:** 3 000 000 000 нуклеотидов

Примерно 20 тысяч генов, < 5% генома кодирует белки

# Генетический код

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TTT | F | TCT | S | TAT | Y | TGT | C |
| TTC | F | TCC | S | TAC | Y | TGC | C |
| TTA | L | TCA | S | TAA | *stop* | TGA | *stop* |
| TTG | L | TCG | S | TAG | *stop* | TGG | W |
| CTT | L | CCT | P | CAT | H | CGT | R |
| CTC | L | CCC | P | CAC | H | CGC | R |
| CTA | L | CCA | P | CAA | Q | CGA | R |
| CTG | L | CCG | P | CAG | Q | CGG | R |
| ATT | I | ACT | T | AAT | N | AGT | S |
| ATC | I | ACC | T | AAC | N | AGC | S |
| ATA | I | ACA | T | AAA | K | AGA | R |
| ATG | M/ *start* | ACG | T | AAG | K | AGG | R |
| GTT | V | GCT | A | GAT | D | GGT | G |
| GTC | V | GCC | A | GAC | D | GGC | G |
| GTA | V | GCA | A | GAA | E | GGA | G |
| GTG | V | GCG | A | GAG | E | GGG | G |

# Поиск генов если известен белок: просто

```
aaacgccctggccttctatagccactattatcagccgctgccgttggtactgcgcggata

tggtgcgggcaatgacgttacagctgccggtgtctttgctgatctgctacgtaccctctc

                          M  V  K  V  Y  A  P  A  S  S  A  N  M
atggaagttaggagtctgacatggttaaagtttatgccccggcttccagtgccaatatga

S  V  G  F  D  V  L  G  A  A  V  T  P  V  D  G  A  L  L  G
gcgtcgggtttgatgtgctcggggcggcggtgacacctgttgatggtgcattgctcggag

D  V  V  T  V  E  A  A  E  T  F  S  L  N  N  L  G  R  F  G
atgtagtcacggttgaggcggcagagacattcagtctcaacaacctcggacgctttgggt

A  D  K  L  P  S  E  P  R  E  N  V  Y  Q  C  W  E  R  F  C
ccgataagctgccgtcagagccacgggaaaatgtttatcagtgctgggagcgtttttgcc

Q  E  L  G  K  Q  I  P  V  A  M  T  L  E  K  N  M  P  I  G
aggagcttggcaagcaaattccagtggcgatgactctggaaaagaatatgccgatcggtt

F  V  H  I  C  R  L  D  T  A  G  A  R  V  L  E  N
ttgttcatatttgccggctggatacggcgggcgcacgagtactggaaaactaaatgaaac

tctacaatctgaaagatcacaatgagcaggtcagctttgcgcaagccgtaacccaggggt

tgggcaaaaatcaggggctgtttttcccgcacgacctgccggaattcagcctgactgaaa
```

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TTT | F | TCT | S | TAT | Y | TGT | C |
| TTC | F | TCC | S | TAC | Y | TGC | C |
| TTA | L | TCA | S | TAA | *stop* | TGA | *stop* |
| TTG | L | TCG | S | TAG | *stop* | TGG | W |
| CTT | L | CCT | P | CAT | H | CGT | R |
| CTC | L | CCC | P | CAC | H | CGC | R |
| CTA | L | CCA | P | CAA | Q | CGA | R |
| CTG | L | CCG | P | CAG | Q | CGG | R |
| ATT | I | ACT | T | AAT | N | AGT | S |
| ATC | I | ACC | T | AAC | N | AGC | S |
| ATA | I | ACA | T | AAA | K | AGA | R |
| ATG | *M/start* | ACG | T | AAG | K | AGG | R |
| GTT | V | GCT | A | GAT | D | GGT | G |
| GTC | V | GCC | A | GAC | D | GGC | G |
| GTA | V | GCA | A | GAA | E | GGA | G |
| GTG | V | GCG | A | GAG | E | GGG | G |

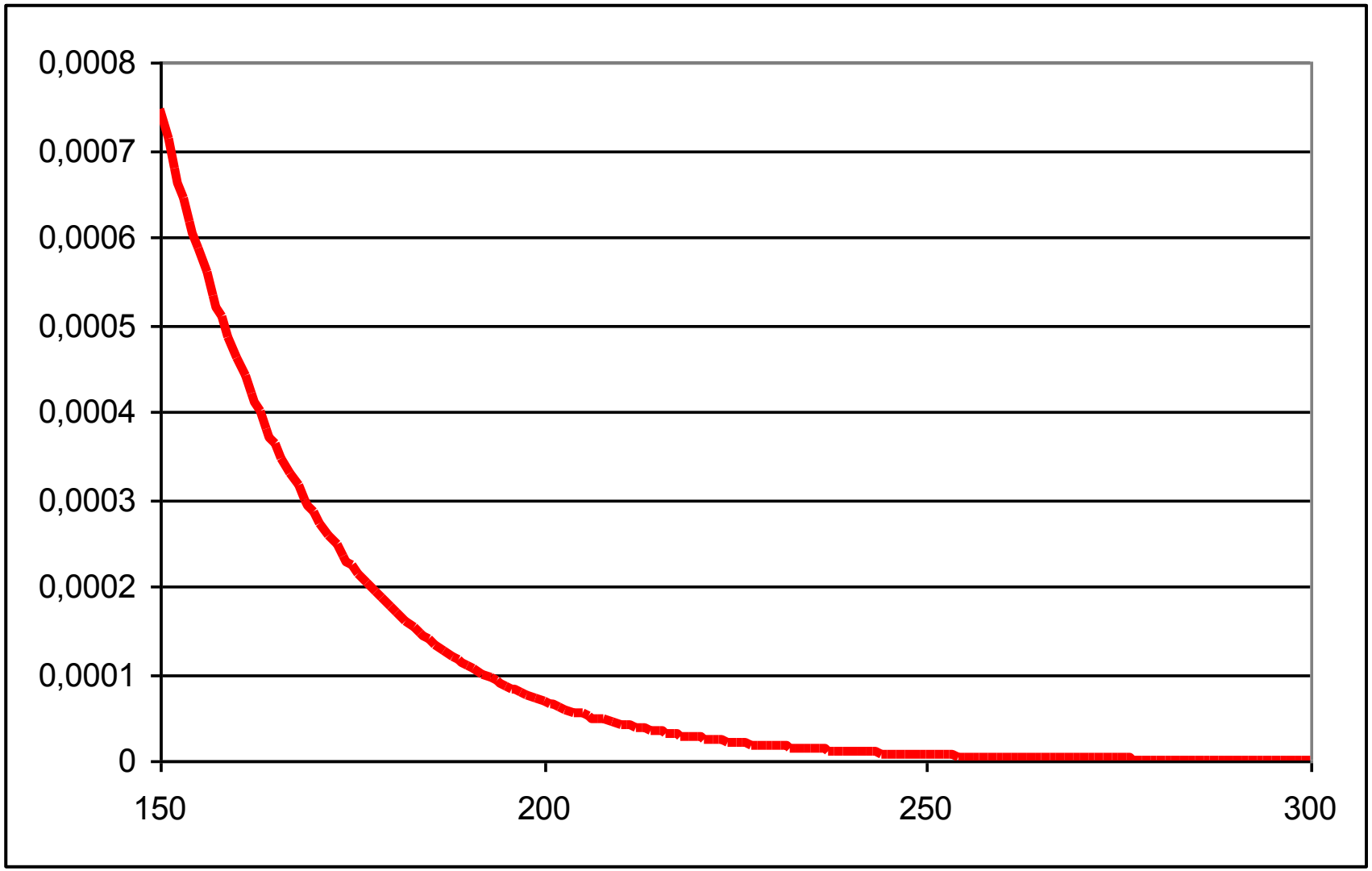
# … или родственный белок: тоже просто

```
aaacgccctggccttctatagccactattatcagccgctgccgttggtactgcgcggata

tggtgcgggcaatgacgttacagctgccggtgtctttgctgatctgctacgtaccctctc

                        M  V  V  V  Y  A  P  A  S  I  G  N  V
atggaagttaggagtctgacatggttaaagtttatgccccggcttccagtgccaatatga

S  V  G  F  D  V  L  G  A  A  V  S  P  I  D  G  S  L  L  G
gcgtcgggtttgatgtgctcggggcggcggtgacacctgttgatggtgcattgctcggag

D  R  V  E  V  A  A  G  E  Q  P  F  T  L  K  C  V  G  D  F
atgtagtcacggttgaggcggcagagaca---ttcagtctcaacaacctcggacgctttg

V  A  K  L  P  V  E  Q  E  E  N  V  Y  H  C  W  L  V  F  A
ccgataagctgccgtcagagccacgggaaaatgtttatcagtgctgggagcgttttgcc

R  E  L  D  K  K  V  P  V  S  M  T  L  E  K  N  M  P  I  G
aggagcttggcaagcaaattccagtggcgatgactctggaaaagaatatgccgatcggtt

F  V  H  V  C  R  L  D  S  T  G  S  K  V  L  E  N
ttgttcatatttgccggctggatacggcgggcgcacgagtactggaaaactaaatgaaac

tctacaatctgaaagatcacaatgagcaggtcagctttgcgcaagccgtaacccaggggt

tgggcaaaaatcaggggctgtttttcccgcacgacctgccggaattcagcctgactgaaa
```

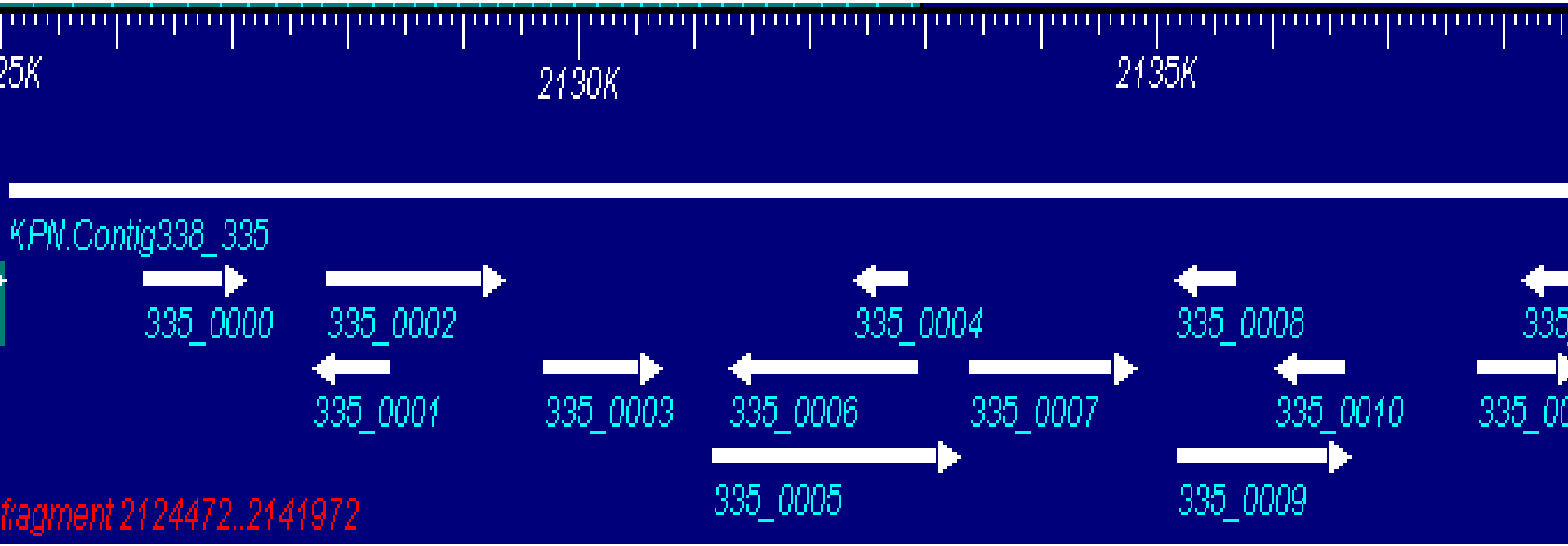
# Генетический код: стоп-кодоны

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TTT | F | TCT | S | TAT | Y | TGT | C |
| TTC | F | TCC | S | TAC | Y | TGC | C |
| TTA | L | TCA | S | TAA | *stop* | TGA | *stop* |
| TTG | L | TCG | S | TAG | *stop* | TGG | W |
| CTT | L | CCT | P | CAT | H | CGT | R |
| CTC | L | CCC | P | CAC | H | CGC | R |
| CTA | L | CCA | P | CAA | Q | CGA | R |
| CTG | L | CCG | P | CAG | Q | CGG | R |
| ATT | I | ACT | T | AAT | N | AGT | S |
| ATC | I | ACC | T | AAC | N | AGC | S |
| ATA | I | ACA | T | AAA | K | AGA | R |
| ATG | M/ *start* | ACG | T | AAG | K | AGG | R |
| GTT | V | GCT | A | GAT | D | GGT | G |
| GTC | V | GCC | A | GAC | D | GGC | G |
| GTA | V | GCA | A | GAA | E | GGA | G |
| GTG | V | GCG | A | GAG | E | GGG | G |

# Вероятность открытой рамки считывания длины не меньше данной



# ORFы в геноме *K. pneumoniae*



# Перепред-сказание (*E. coli*)

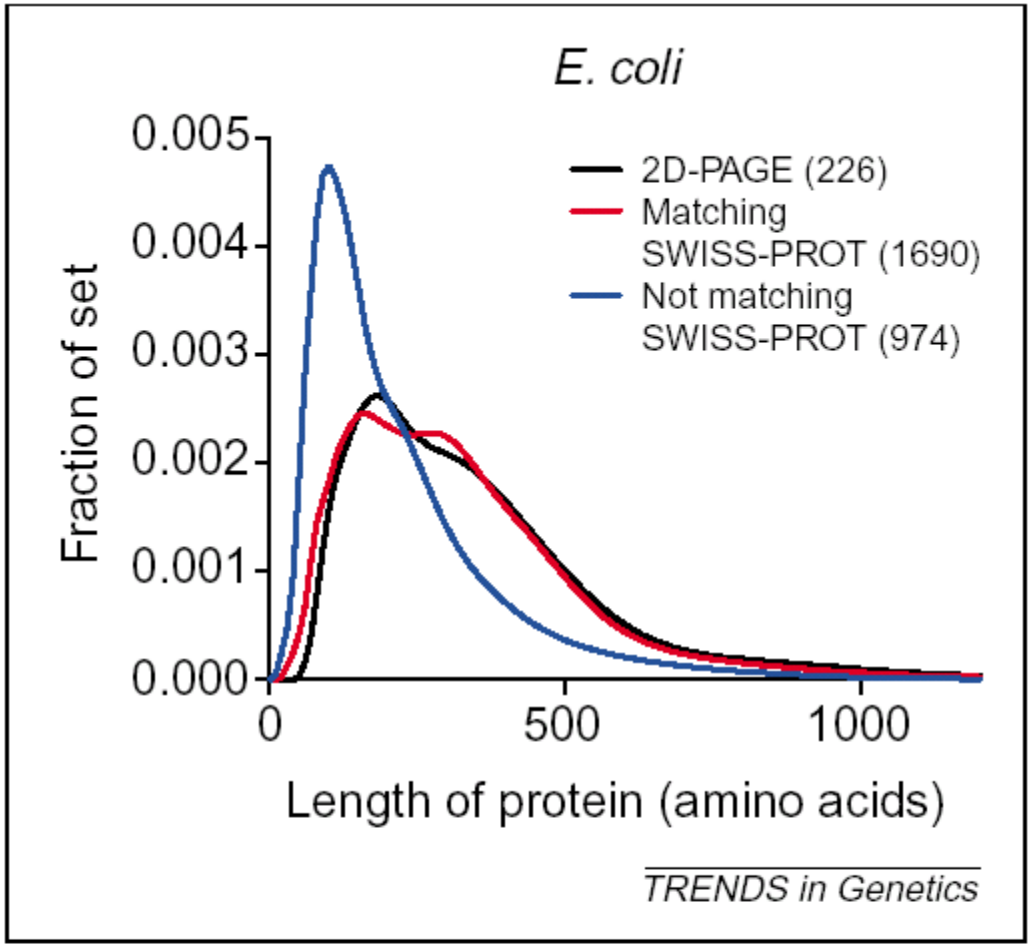



**Fig. 3.** Protein length distributions for *Escherichia coli*. The red and blue lines show the length distributions of unique sequences matching and not matching entries from SWISS-PROT, respectively. The black line shows the distribution of unique proteins from SWISS-2D-PAGE. All length distributions are normalized and have been smoothed by Gaussian kernel density estimation.

# Сильное перепред-сказание (*Aeropyrum pernix*)

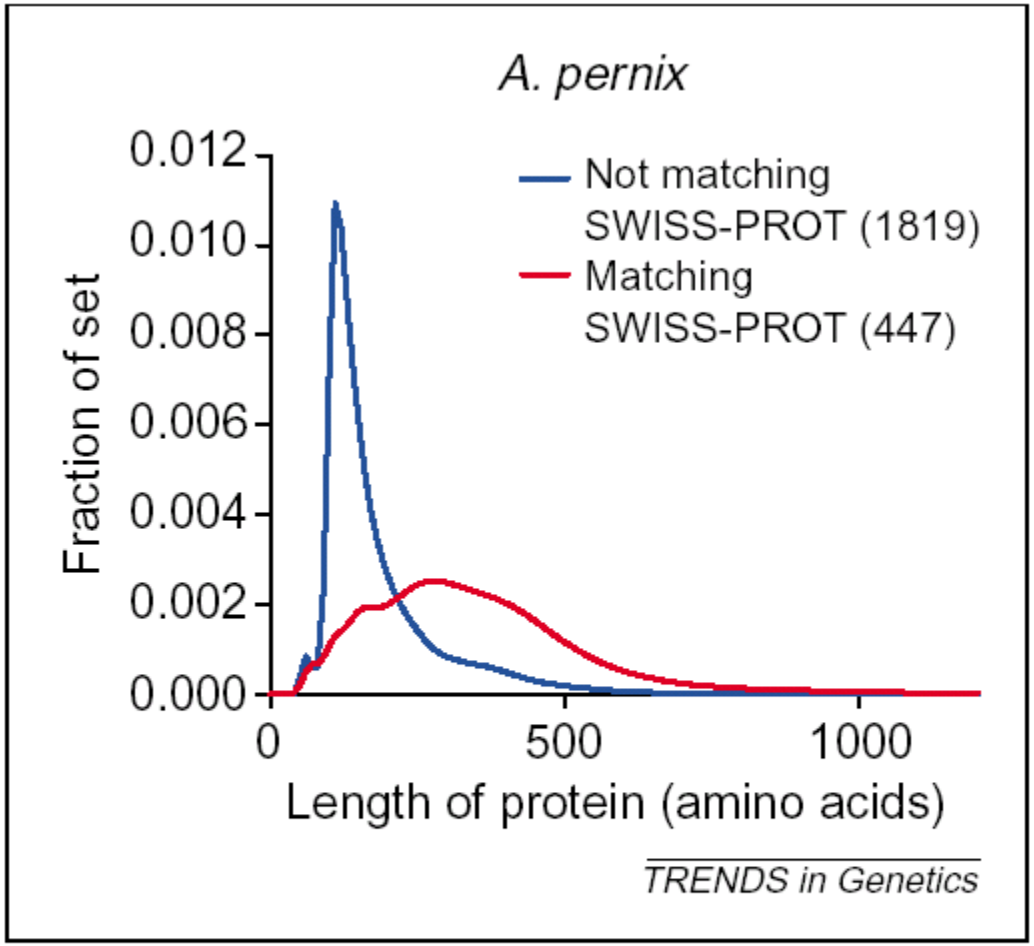
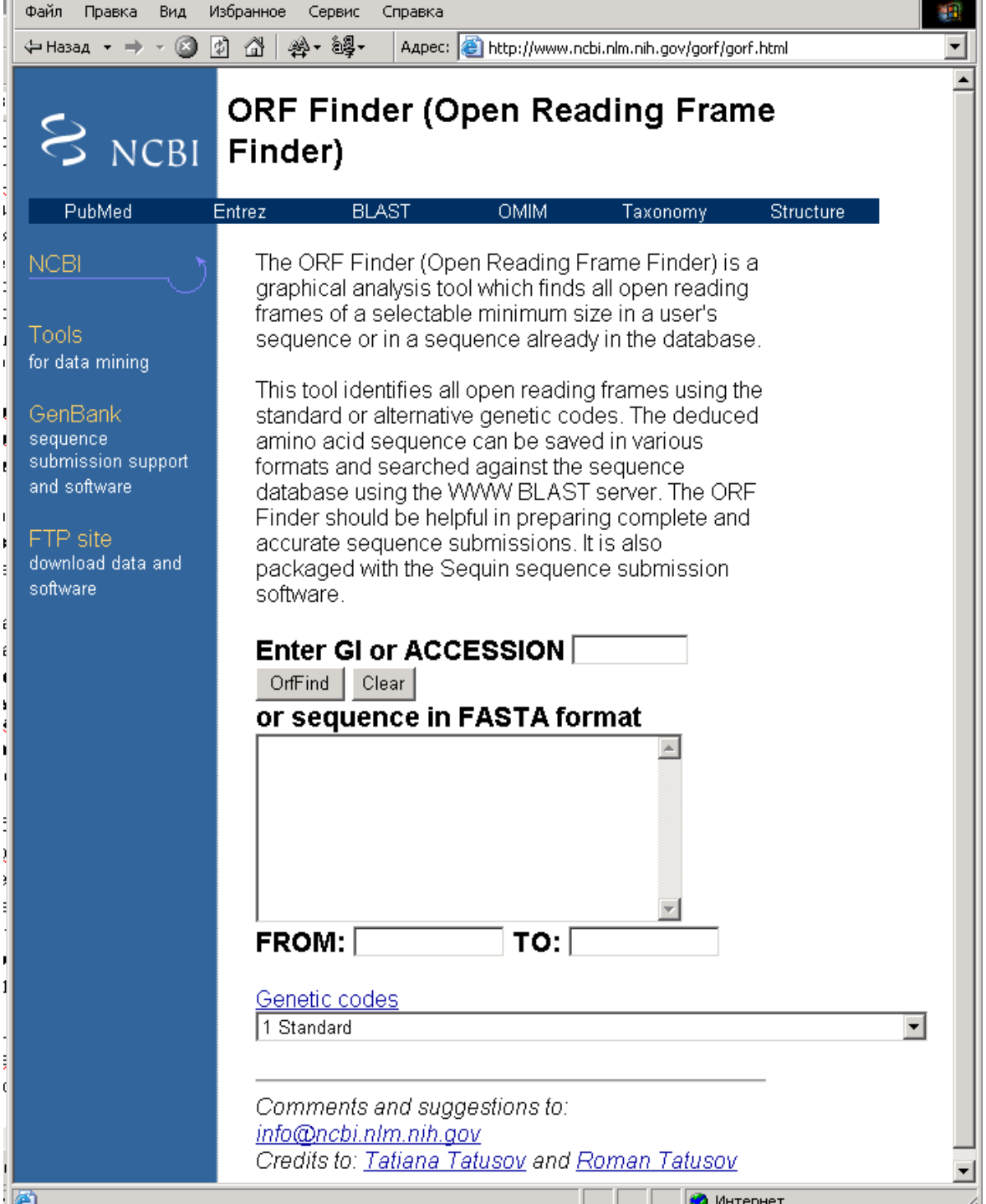



**Fig. 4.** Protein length distributions for *Aeropyrum pernix*. The curves are similar to those in Fig. 3, except that there is no 2D-PAGE data.

# Поиск открытых рамок в заданной последова-тельности



# Генетический код: синонимы

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TTT | F | TCT | S | TAT | Y | TGT | C |
| TTC | F | TCC | S | TAC | Y | TGC | C |
| TTA | L | TCA | S | TAA | *stop* | TGA | *stop* |
| TTG | L | TCG | S | TAG | *stop* | TGG | W |
| CTT | L | CCT | P | CAT | H | CGT | R |
| CTC | L | CCC | P | CAC | H | CGC | R |
| CTA | L | CCA | P | CAA | Q | CGA | R |
| CTG | L | CCG | P | CAG | Q | CGG | R |
| ATT | I | ACT | T | AAT | N | AGT | S |
| ATC | I | ACC | T | AAC | N | AGC | S |
| ATA | I | ACA | T | AAA | K | AGA | R |
| ATG | M/ *start* | ACG | T | AAG | K | AGG | R |
| GTT | V | GCT | A | GAT | D | GGT | G |
| GTC | V | GCC | A | GAC | D | GGC | G |
| GTA | V | GCA | A | GAA | E | GGA | G |
| GTG | V | GCG | A | GAG | E | GGG | G |

# Codon usage
# (статистика употребления кодонов)

- частоты кодонов отличаются от частот триплетов в некодирующих областях
  - различия в частотах аминокислот в белках
  - различия в частотах синонимичных кодонов
- частоты синонимичных кодонов
  - специфичны для генома
  - коррелируют с концентрациями тРНК

# Ещё про codon usage

- различается у высоко- и низко-экспрессируемых генов (у высокоэкспрессируемых генов больше доля «оптимальных» кодонов) – прокариоты, дрожжи
- нестандартный у горизонтально перенесенных генов
- у фага Т4 – близок к хозяйскому (*E. coli*) у ранних генов, специфический (соответствует своему набору тРНК) – у поздних

# Кодирующий потенциал

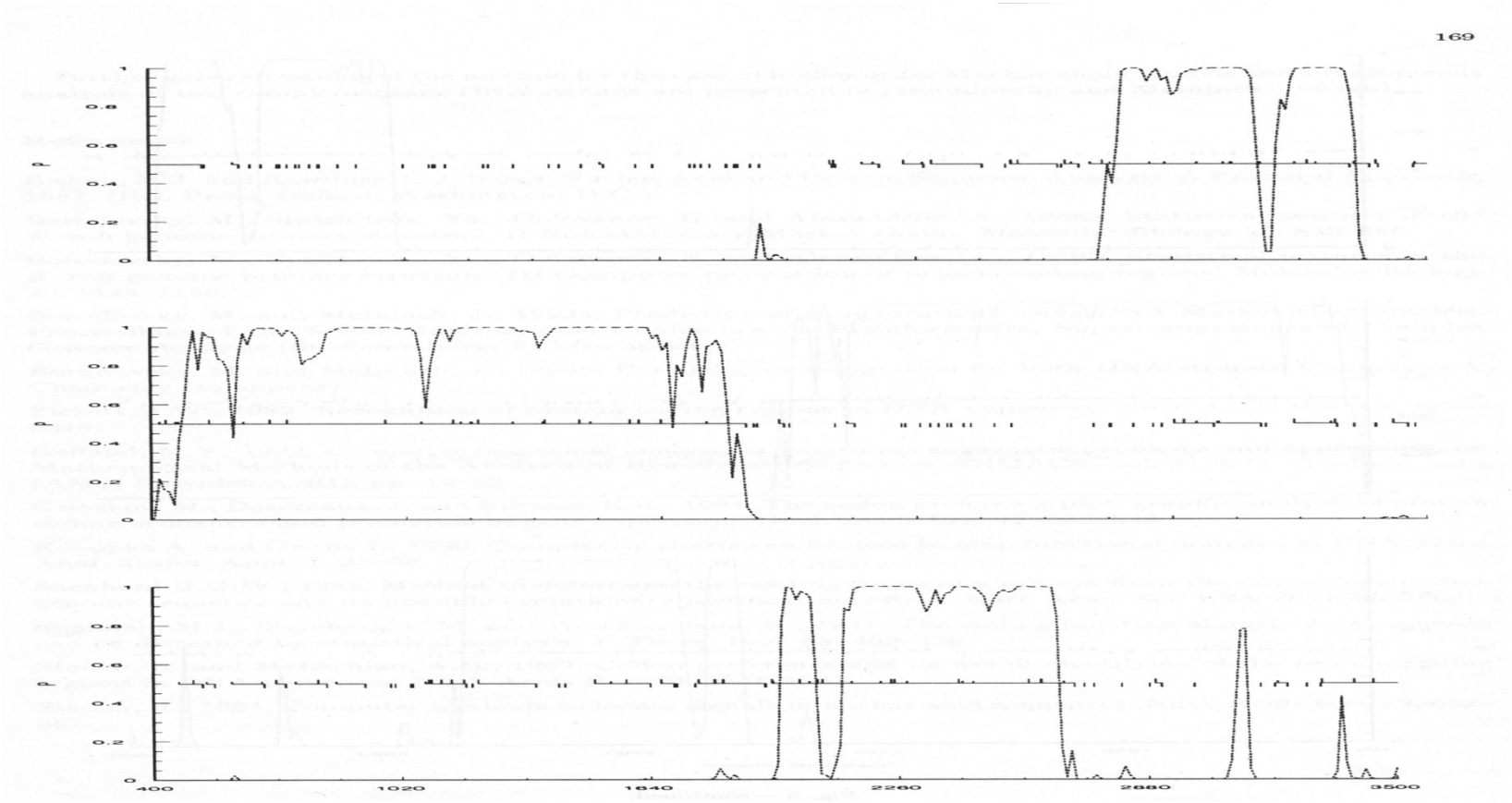
Функция, измеряющая, насколько участок генома похож на белок-кодирующий (и отличается от некодирующего) с точки зрения статистики

Можно вычислять кодирующий потенциал

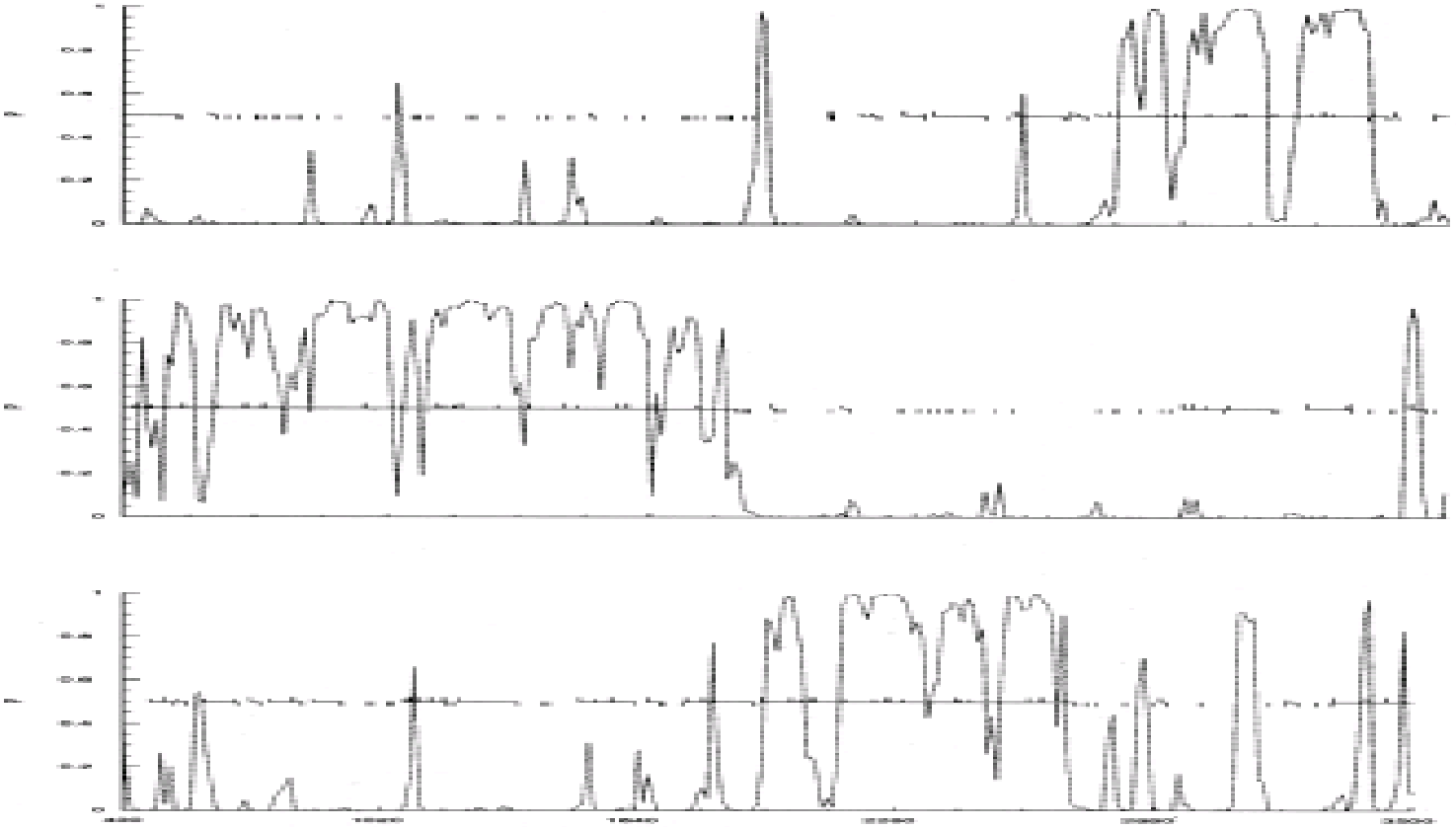
- скользящего окна (не слишком маленького!)
- открытой рамки считывания

Нужна обучающая выборка генов (и межгенных промежутков) из данного организма

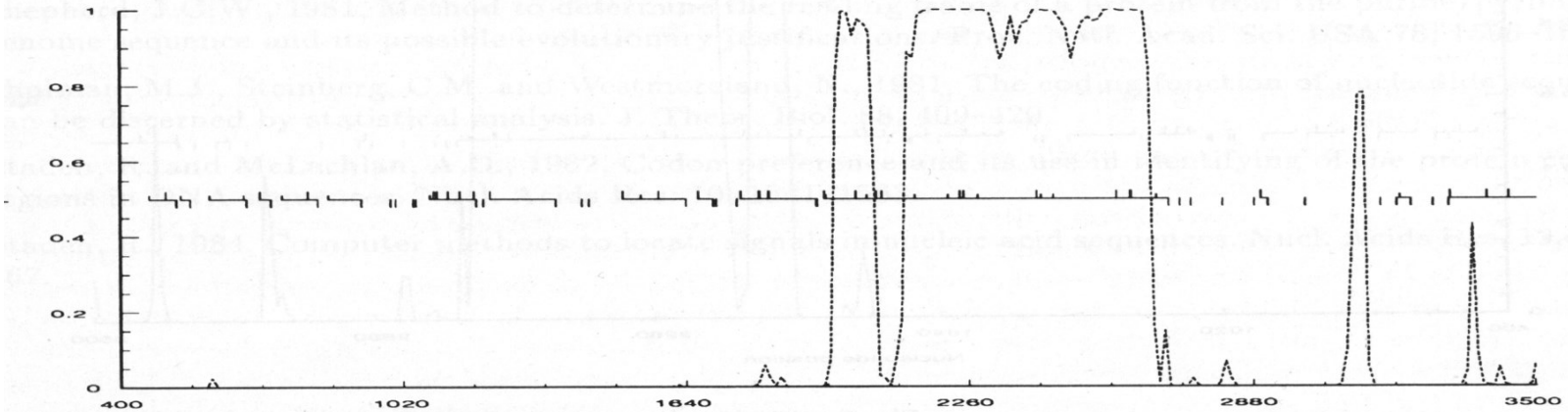
# *E. coli*. Окно 96 нт

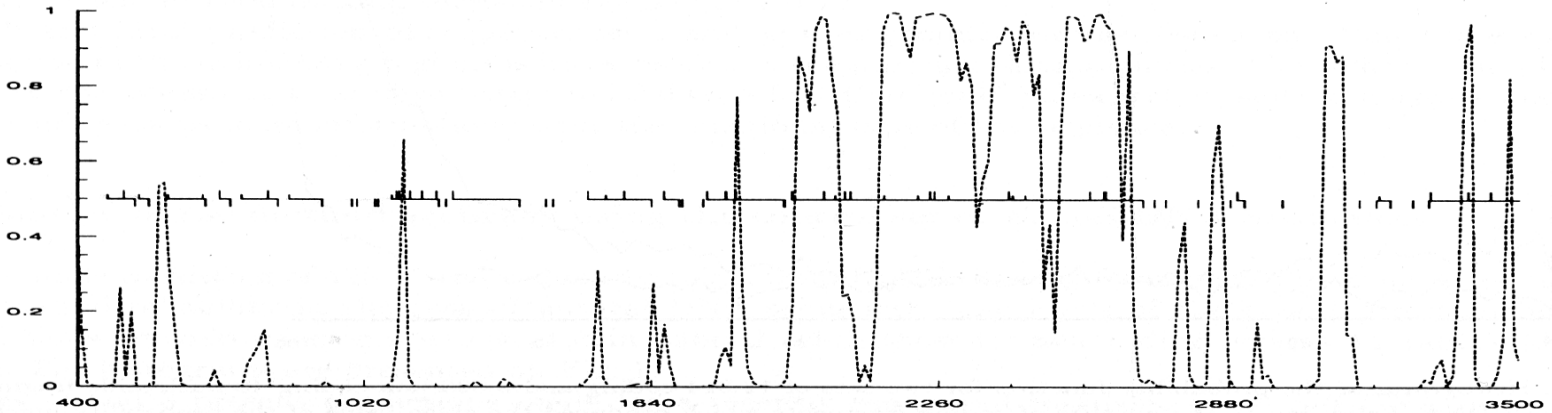


# *E. coli*. Окно 48 нт



# Сравнение предсказаний при разной длине окон



# Генетический код: старт-кодоны

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| TTT | F | TCT | S | TAT | Y | TGT | C |
| TTC | F | TCC | S | TAC | Y | TGC | C |
| TTA | L | TCA | S | TAA | *stop* | TGA | *stop* |
| TTG | L | TCG | S | TAG | *stop* | TGG | W |
| CTT | L | CCT | P | CAT | H | CGT | R |
| CTC | L | CCC | P | CAC | H | CGC | R |
| CTA | L | CCA | P | CAA | Q | CGA | R |
| CTG | L | CCG | P | CAG | Q | CGG | R |
| ATT | I | ACT | T | AAT | N | AGT | S |
| ATC | I | ACC | T | AAC | N | AGC | S |
| ATA | I | ACA | T | AAA | K | AGA | R |
| ATG | M/ *start* | ACG | T | AAG | K | AGG | R |
| GTT | V | GCT | A | GAT | D | GGT | G |
| GTC | V | GCC | A | GAC | D | GGC | G |
| GTA | V | GCA | A | GAA | E | GGA | G |
| GTG | V | GCG | A | GAG | E | GGG | G |

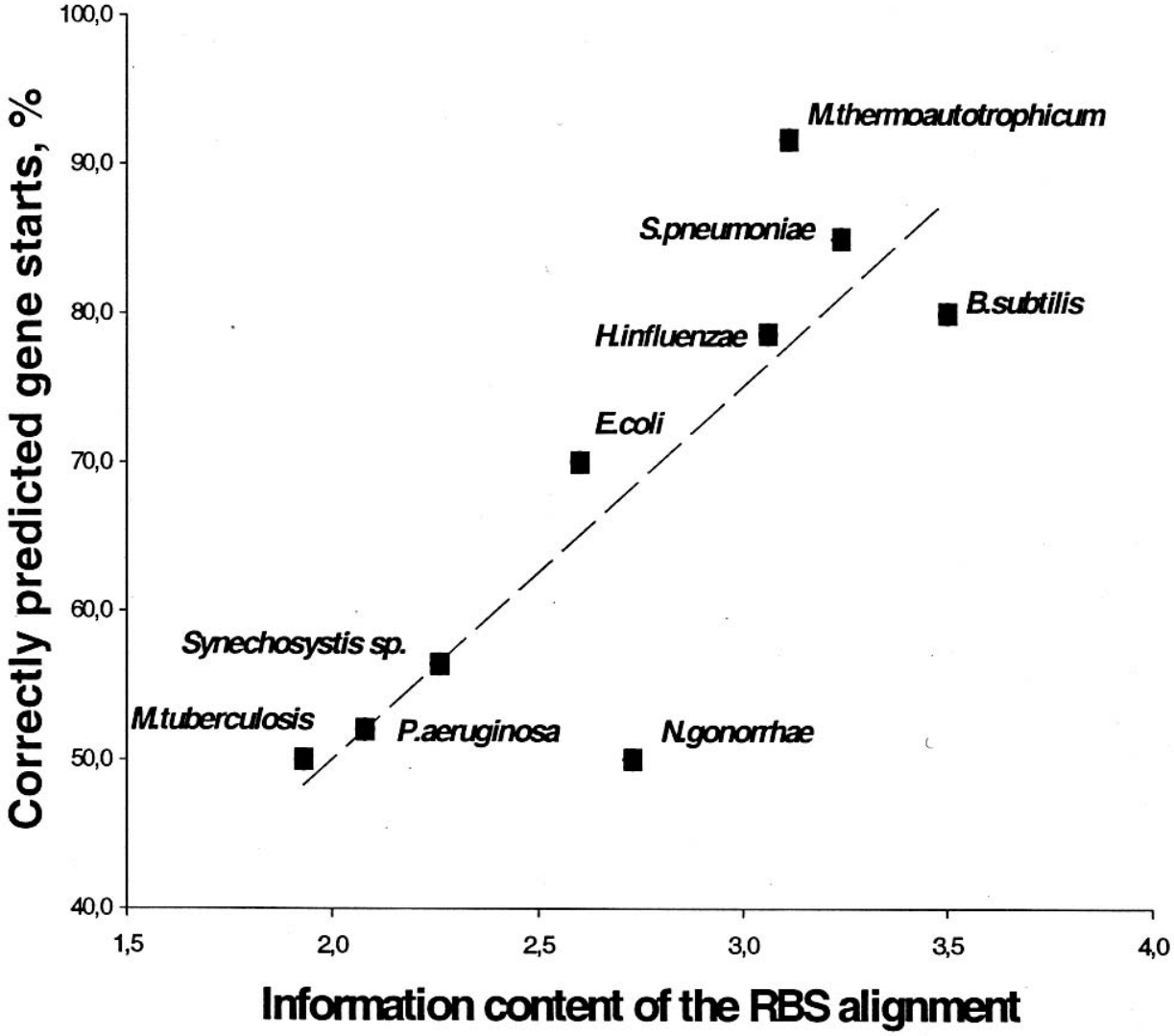
# Сигналы на границах генов

| | |
|---|---|
| *dnaN* | ACATTATCCGTTAGGAGGATAAAA**ATG** |
| *gyrA* | GTGATACTTCAGGGAGGTTTTTTA**ATG** |
| *serS* | TCAATAAAAAAAGGAGTGTTTCGC**ATG** |
| *bofA* | CAAGCGAAGGAGATGAGAAGATTC**ATG** |
| *csfB* | GCTAACTGTACGGAGGTGGAGAAG**ATG** |
| *xpaC* | ATAGACACAGGAGTCGATTATCTC**ATG** |
| *metS* | ACATTCTGATTAGGAGGTTTCAAG**ATG** |
| *gcaD* | AAAAGGGATATTGGAGGCCAATAA**ATG** |
| *spoVC* | TATGTGACTAAGGGAGGATTCGCC**ATG** |
| *ftsH* | GCTTACTGTGGGAGGAGGTAAGGA**ATG** |
| *pabB* | AAAGAAAATAGAGGAATGATACAA**ATG** |
| *rplJ* | CAAGAATCTACAGGAGGTGTAACC**ATG** |
| *tufA* | AAAGCTCTTAAGGAGGATTTTAGA**ATG** |
| *rpsJ* | TGTAGGCGAAAAGGAGGGAAAATA**ATG** |
| *rpoA* | CGTTTTGAAGGAGGGTTTTAAGTA**ATG** |
| *rplM* | AGATCATTTAGGAGGGGAAATTCA**ATG** |

# … после выравнивания

```
dnaN     ACATTATCCGTTAGGAGGATAAAAATG
gyrA      GTGATACTTCAGGGAGGTTTTTTAATG
serS      TCAATAAAAAAAGGAGTGTTTCGCATG
bofA          CAAGCGAAGGAGATGAGAAGATTCATG
csfB       GCTAACTGTACGGAGGTGGAGAAGATG
xpaC         ATAGACACAGGAGTCGATTATCTCATG
metS      ACATTCTGATTAGGAGGTTTCAAGATG
gcaD      AAAAGGGATATTGGAGGCCAATAAATG
spoVC     TATGTGACTAAGGGAGGATTCGCCATG
ftsH     GCTTACTGTGGGAGGAGGTAAGGAATG
pabB         AAAGAAAATAGAGGAATGATACAAATG
rplJ      CAAGAATCTACAGGAGGTGTAACCATG
tufA       AAAGCTCTTAAGGAGGATTTTAGAATG
rpsJ      TGTAGGCGAAAAGGAGGGAAAATAATG
rpoA         CGTTTTGAAGGAGGGTTTTAAGTAATG
rplM        AGATCATTTAGGAGGGGAAATTCAATG
cons.         tacataaaggaggtttaaaaat
num.          0000000111111000000001
              5755779156663678679890
```

# Участки связывания рибосом



# Сравнение генов в родственных геномах

Гены консервативнее, чем межгенные области (точнее, особенности эволюции другие)

- преимущественно синонимичные замены – по третьим позициям кодонов
- вставки и удаления имеют длину, кратную 3

# Паттерн нуклеотидных замен в белок-кодирующих областях: *pdxB* в энтеробактериях

```
Sty  TCGCTCG--CAGCGGAAAGAGGATTACGCCCTTCGCCTGGAGGCTGTGCAGGGGC---GCCGGAGATGGGATGCATAATT
Stm  TCGCTCG--CAGCGGAAAGAGGATTACGCCCTTCGCCTGGAGGCTGTGCAGGGGC---GCCGGAGATGGGATGCATAATT
Sen  TCGCTCG--CAGCGGAAAGAGGATTACGCCCTTCGCCTGGAGGCTGTGCAGGGGC---GCCGGAGATGGGATGCATAATT
Eco  TTGCCCG--TGCCAGACGGCAGATTATCTCCCTGACCTGGTGGTTGCCCAGGAGGAGGGCCGGAAATAGGTTGTATCATT
Kpn  ----CGG--TGGCGCAGTGCCTGATGGG-CCTCGCCCTGGAGGACGGTCTGGCAT---ATCAGCAAGGGGGTGCGTCATG
Ype  TTGTTAGAACAGGGGAAAACGGTAAACAGTGTGGCATTAGATGTCGGTTATAGCT-----CCGCCTCTGCTTTTATCGCC
           *        *                      * *  *  *              * *     *  *   *

Sty  AATTATCCTTTAAC----------CATAAATCTGAGCAATA-TATGCTTGGCGGCCAGATTATGGC--ACACTTGTCCGG
Stm  AATTATCCTTTAAC----------CATAAATCTGAGCAATA-TATGCCTGGCGGCCAGATTATGGC--ACACTTGTCCGG
Sen  AATTATCCTTTAAC----------CATAAATCTGAGCAATA-TATGCCTGGCGGCCAGATTATGGC--ACACTTGTCCGG
Eco  ACGTATCCTTATAC----------CTGAAATCTTCGCAAG--TATGCCTGGCCGCGAGATTATGGC--ACACTTGTCCGG
Kpn  ATTCATCCTTTCGATATCGCGGTGCTGGAACCAGGTGATGAGTATGCCTGGCGGCCAGATTATGGC--ACACTTCCCCAG
Ype  ATGTTTCAGCAAATAT--------CGGGTACCA-CGCCTGAGCGTTTCCGGCGGGGCAATAGTGGCTTATACTAAGCCCC
     *    **                 *    * *            *    *** *    **  ****  * ***   **

Sty  TTAACTCTCGTT-CTCAAACAG------GTACGACAGTC--GTGAAAATTCTCGTTGATGAAAATATGCCTTACGCCCGC
Stm  TTAACTCTCGTT-CTCAAACAG------GTACGACAGTC--GTGAAAATTCTCGTTGATGAAAATATGCCTTACGCCCGC
Sen  TTAACTCTCGTT-CTCAAACAG------GTACGACAGTC--GTGAAAATTCTCGTTGATGAAAATATGCCTTACGCCCGC
Eco  TTAACTCTCGT--CTCATACAG------GTAACACAAAC--GTGAAAATCCTTGTTGATGAAAATATGCCTTATGCCCGC
Kpn  TTAACTCTCGTT-CTCAGACAG------GTACTGAACT---GTGAAAATCCTCGTTGATGAAAATATGCCCTATGCCCGT
Ype  CTGTTTTTCATCTGTATGGCAGTTCGCTGTCGGAGAGTAAAGTGAAAATTCTGGTTGATGAAAATATGCCGTACGCTGAG
      *   * ** *   *    ***      **     *     ******** ** ***************** ** **
                                             123123123123123123123123123123123123123
```

# *rbsD* в энтеробактериях

```
Sty AGGGTTACACTGCGGC-CAGCGAAACGTTTCGCTAGTGGAGCAGAAAAATGAAGAAAGGC
Sen AGGGTTACACTGCGGC-CAGCGAAACGTTTCGCTAGTGGAGCAGAAAAATGAAGAAAGGC
Stm GGGGTTACACTGCGGC-CAGCGAAACGTTTCGCTAGTGGAGCAGAAAAATGAAGAAAGGC
Eco AGGATTAAACTGTGGGTCAGCGAAACGTTTCGCTGATGGAGAA-AAAAATGAAAAAAGGC
Ype TTTTCTAAACTCCTTGTTAGCGAAACGTTTCGCTCTTGGAGTA-GATCATGAAAAAAGGT
         ** ***       ****************  ***** *  *  ***** *****

Sty ACCGTACTCAACTCTGAAATCTCGTCGGTCATTTCCCGTCTGGGGCATACTGATACTCTG
Sen ACCGTACTCAACTCTGAAATCTCGTCGGTCATTTCCCGTCTGGGGCATACTGATACTCTG
Stm ACCGTACTCAACTCTGAAATCTCGTCGGTCATTTCCCGTCTGGGGCATACTGATACTCTG
Eco ACCGTTCTTAATTCTGATATTTCATCGGTGATCTCCCGTCTGGGACATACCGATACGCTG
Ype GTATTACTGAACGCTGATATTTCCGCGGTTATCTCCCGTCTGGGCCATACCGATCAGATT
        * ** **  **** ** **  **** ** *********** ***** ***    *
```

# *rbsD* в энтеробактериях: ответ

```
Sty AGGGTTACACTGCGGC-CAGCGAAACGTTTCGCTAGTGGAGCAGAAAAATGAAGAAAGGC
Sen AGGGTTACACTGCGGC-CAGCGAAACGTTTCGCTAGTGGAGCAGAAAAATGAAGAAAGGC
Stm GGGGTTACACTGCGGC-CAGCGAAACGTTTCGCTAGTGGAGCAGAAAAATGAAGAAAGGC
Eco AGGATTAAACTGTGGGTCAGCGAAACGTTTCGCTGATGGAGAA-AAAAATGAAAAAAGGC
Ype TTTTCTAAACTCCTTGTTAGCGAAACGTTTCGCTCTTGGAGTA-GATCATGAAAAAAGGT
         ** ***       ****************  ***** *  *  ***** *****

Sty ACCGTACTCAACTCTGAAATCTCGTCGGTCATTTCCCGTCTGGGGCATACTGATACTCTG
Sen ACCGTACTCAACTCTGAAATCTCGTCGGTCATTTCCCGTCTGGGGCATACTGATACTCTG
Stm ACCGTACTCAACTCTGAAATCTCGTCGGTCATTTCCCGTCTGGGGCATACTGATACTCTG
Eco ACCGTTCTTAATTCTGATATTTCATCGGTGATCTCCCGTCTGGGACATACCGATACGCTG
Ype GTATTACTGAACGCTGATATTTCCGCGGTTATCTCCCGTCTGGGCCATACCGATCAGATT
        * ** **  **** ** **  **** ** *********** ***** ***    *
```

# Существующая аннотация (была) неправильна

```
Sty AGGGTTACACTGCGGC-CAGCGAAACGTTTCGCTAGTGGAGCAGAAAAATGAAGAAAGGC
Sen AGGGTTACACTGCGGC-CAGCGAAACGTTTCGCTAGTGGAGCAGAAAAATGAAGAAAGGC
Stm GGGGTTACACTGCGGC-CAGCGAAACGTTTCGCTAGTGGAGCAGAAAAATGAAGAAAGGC
Eco AGGATTAAACTGTGGGTCAGCGAAACGTTTCGCTGATGGAGAA-AAAAATGAAAAAAGGC
Ype TTTTCTAAACTCCTTGTTAGCGAAACGTTTCGCTCTTGGAGTA-GATCATGAAAAAAGGT
         ** ***       ***************  ***** *   *  ***** *****
```

```
Sty ACCGTACTCAACTCTGAAATCTCGTCGGTCATTTCCCGTCTGGGGCATACTGATACTCTG
Sen ACCGTACTCAACTCTGAAATCTCGTCGGTCATTTCCCGTCTGGGGCATACTGATACTCTG
Stm ACCGTACTCAACTCTGAAATCTCGTCGGTCATTTCCCGTCTGGGGCATACTGATACTCTG
Eco ACCGTTCTTAATTCTGATATTTCATCGGTGATCTCCCGTCTGGGACATACCGATACGCTG
Ype GTATTACTGAACGCTGATATTTCCGCGGTTATCTCCCGTCTGGGCCATACCGATCAGATT
        * ** **  **** ** **  **** ** *********** ***** ***    *
```

# Уточнение стартов – белковое выравнивание (*ribD*)

```
Eco V_____QDEYYMARALKLAQRGRFTTHPNPNVGCVIVKDGEIVGEGYHQRAGEPHAEVHA
          QD  +M RAL LA +G +TT PNP VGCV VK+GEIVGEG+H +AG+PHAE  A
Hin MLEFSSQDCVFMQRALDLAAKGQYTTTPNPSVGCVLVKNGEIVGEGFHFKAGQPHAERVA
```

```
Eco   GCGCGCCTGGAGGACTAA----G----------CCGTGCAGGAC-GAGTATTACATGGCGCGGGCGCTAA
      *      *  **** ***    *          **   ** **  * *****  ***   ** **  **
Hin   GAAAAATTAAAGGATTAATTATGCTTGAATTTTCCTCACAAGATTGCGTATTT-ATGCAACGTGCCTTAG
```

# Множественное выравнивание

```
REC06584       109 tttttatttcaggcaatcggggtgaat---------gtggcgcaggcggaagtgttgaat
RECO04717      109 tttttatttcaggcaatcggggtgaat---------gtggcgcaggcggaagtgttgaat
RECS04752      109 tttttatttcaggcaatcggggtgaat---------gtggcgcaggcggaagtgttgaat
RTY01088        51 tagcgcctgttttgatttatggtgaacggggttaatgtggcgcaggcggaagtgttgaat
RSY05814        51 tagcgcctgttttgatttatggtgaacggggttaatgtggcgcaggcggaagtgttgaat
REO01497        66 atagcgcctgtttgatttcattgaattggggaaggcgtgtctacggcggaagtattgaat
RYPK00397       45 gccggcctgtgcagatctaatagttgggggaaaagtgtgtcgaccgcagcagtgataaac
RYP04048        45 gccggcctgtgcagatctaatagttgggggaaaagtgtgtcgaccgcagcagtgataaac
RYE04903        44 aaccggcctgtgcagatctcatagttggggaatagtgtgtcaaccgcagcagtgataaat
RVFI01204        0 .......tattattgatgagtttttatgtccagcatgatcgcagagcaaccaatggaa
REC06584            f  l  f  q  a  i  g  v  n  =  =  =  V  A  Q  A  E  V  L  N
RECO04717           f  l  f  q  a  i  g  v  n  =  =  =  V  A  Q  A  E  V  L  N
RECS04752           f  l  f  q  a  i  g  v  n  =  =  =  V  A  Q  A  E  V  L  N
RTY01088            *  r  l  f  *  f  m  v  n  g  v  n  V  A  Q  A  E  V  L  N
RSY05814            *  r  l  f  *  f  m  v  n  g  v  n  V  A  Q  A  E  V  L  N
REO01497            i  a  p  v  *  f  h  *  i  g  e  g  V  S  T  A  E  V  L  N
RYPK00397           a  g  l  c  r  s  n  s  w  g  k  s  V  S  T  A  A  V  I  N
RYP04048            a  g  l  c  r  s  n  s  w  g  k  s  V  S  T  A  A  V  I  N
RYE04903            n  r  p  v  q  i  s  *  l  g  n  s  V  S  T  A  A  V  I  N
RVFI01204           .  .  .  i  i  d  e  f  f  m  s  s  M  I  A  E  Q  P  M  E
```

# Распознавание генов в отсутствие обучающей выборки

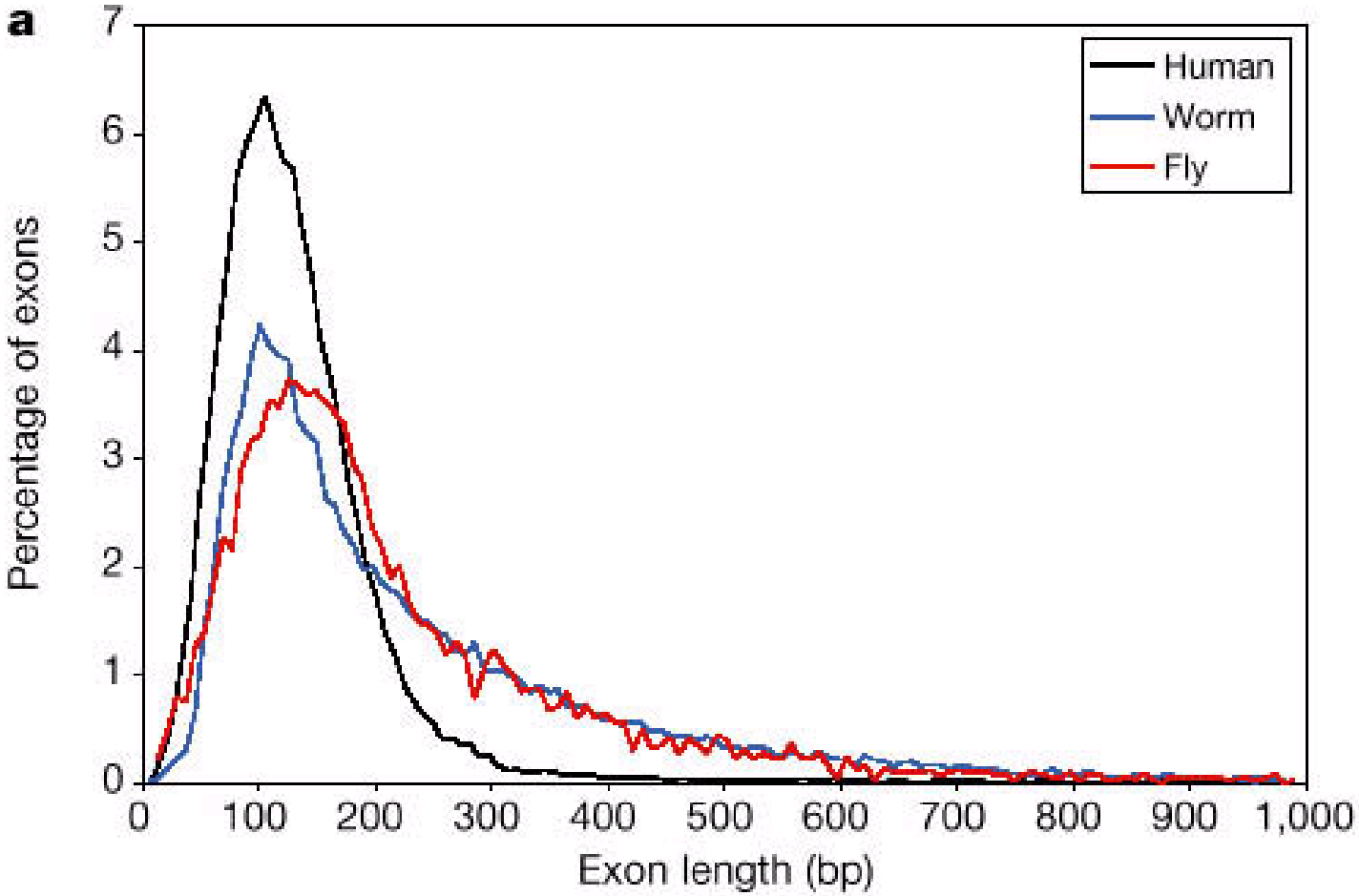
«псевдообучающая выборка»:

- протяженные рамки считывания
- гены, предсказанные по сходству

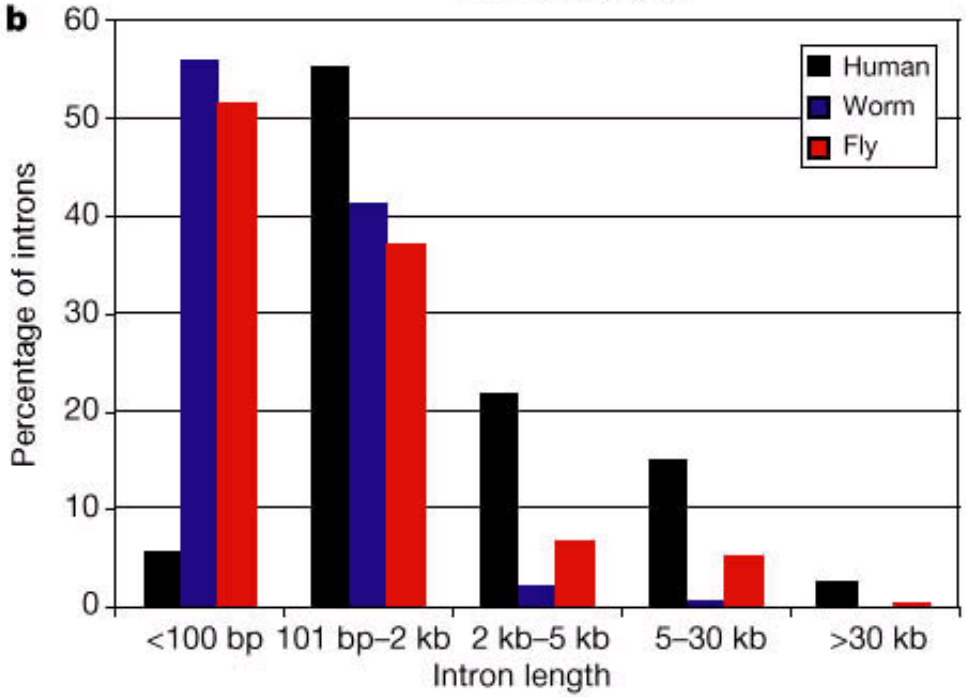
# Эукариоты (человек)

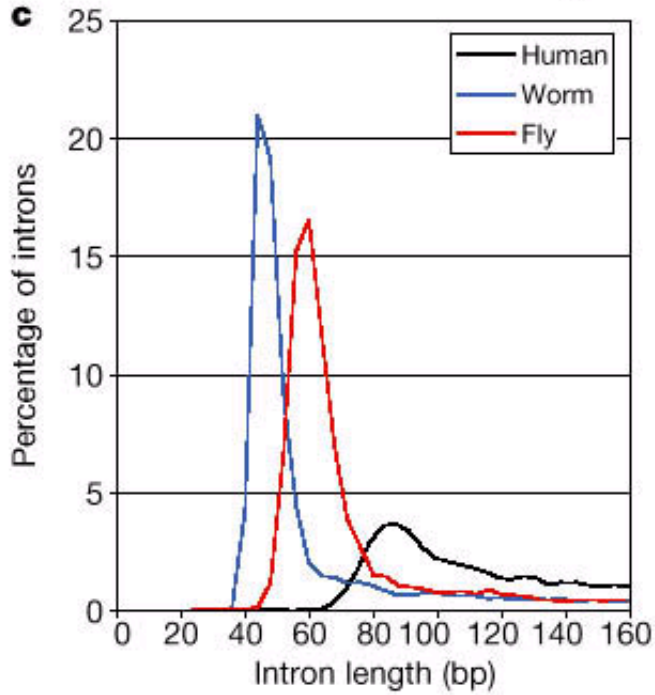
- В среднем 9-10 экзонов (кодирующих) на ген
- Средняя длина (внутреннего) экзона 120-130 нуклеотидов
- Часто очень длинные интроны

# Длины экзонов:
# человек, нематода *C. elegans*, дрозофила

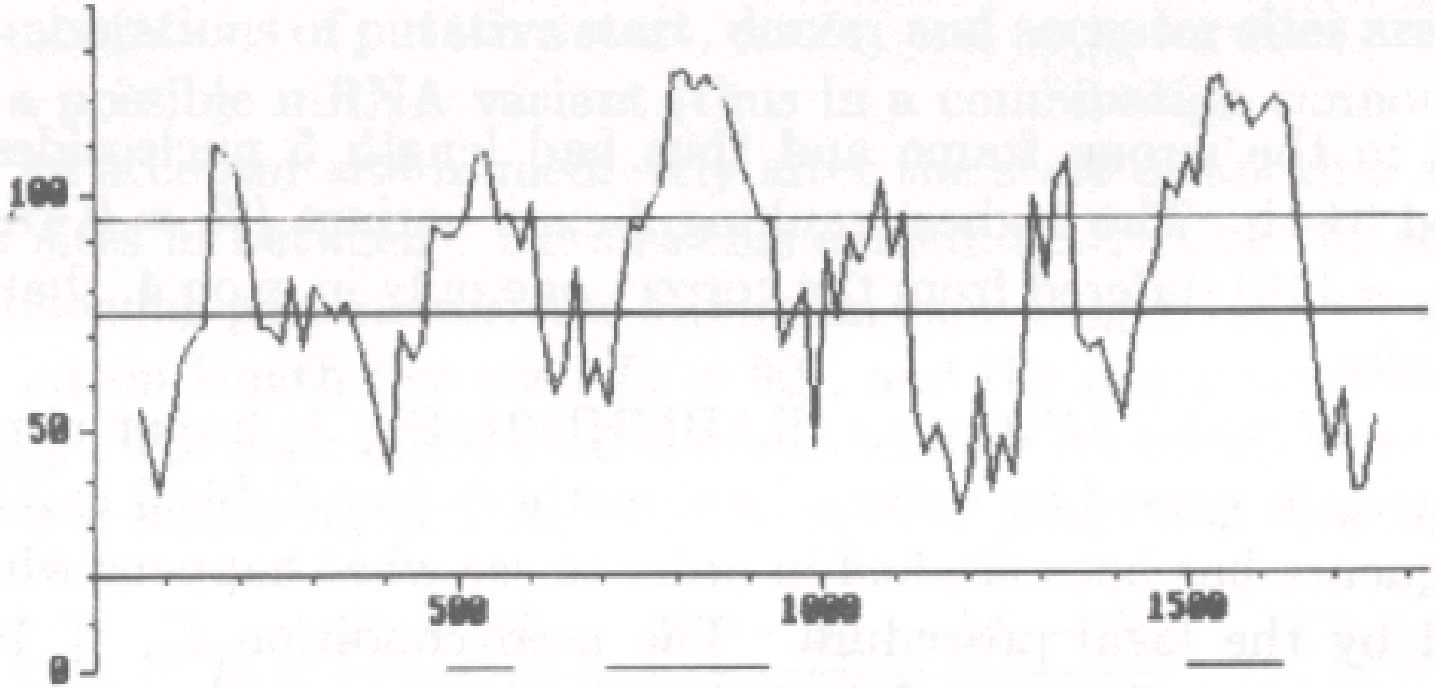


# Длины интронов

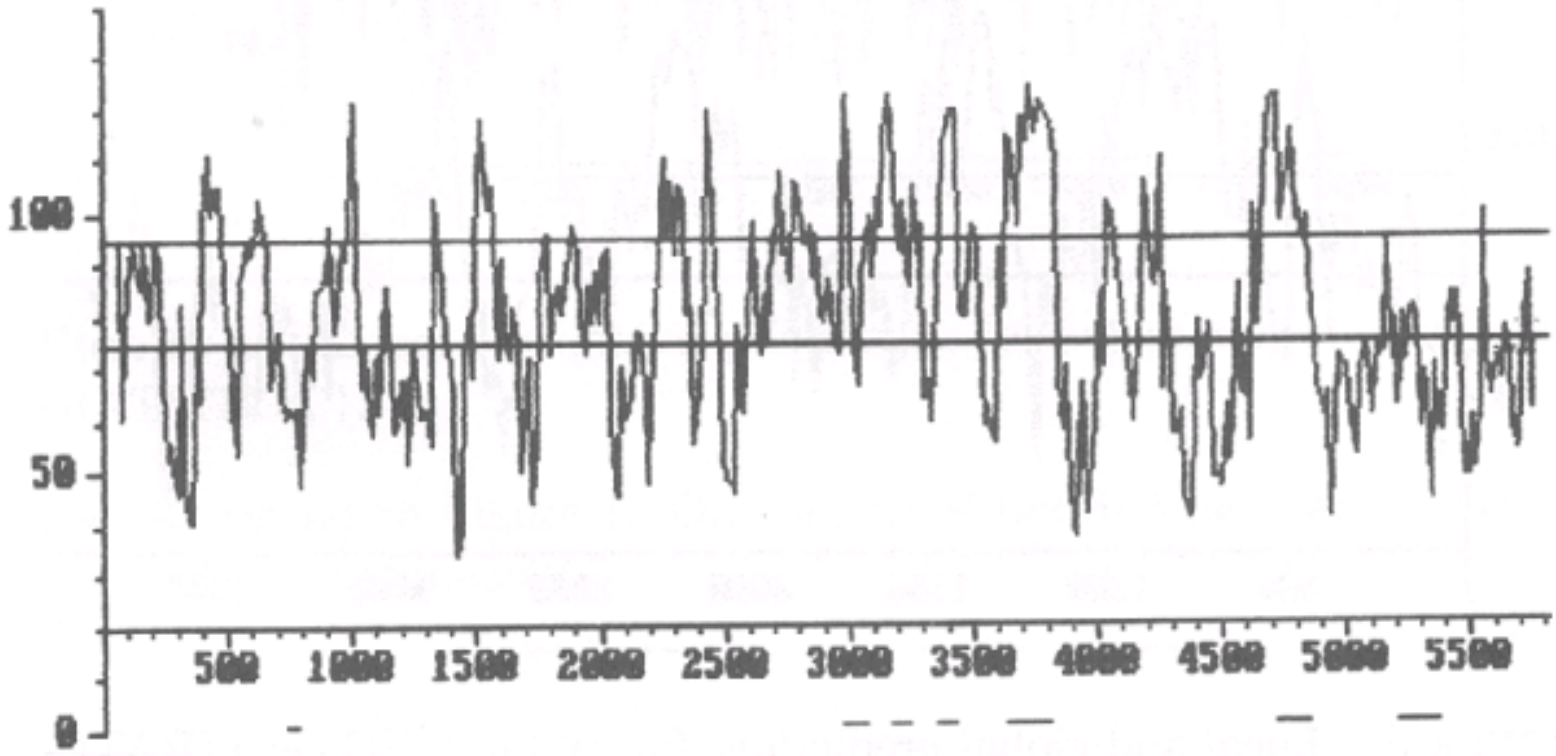


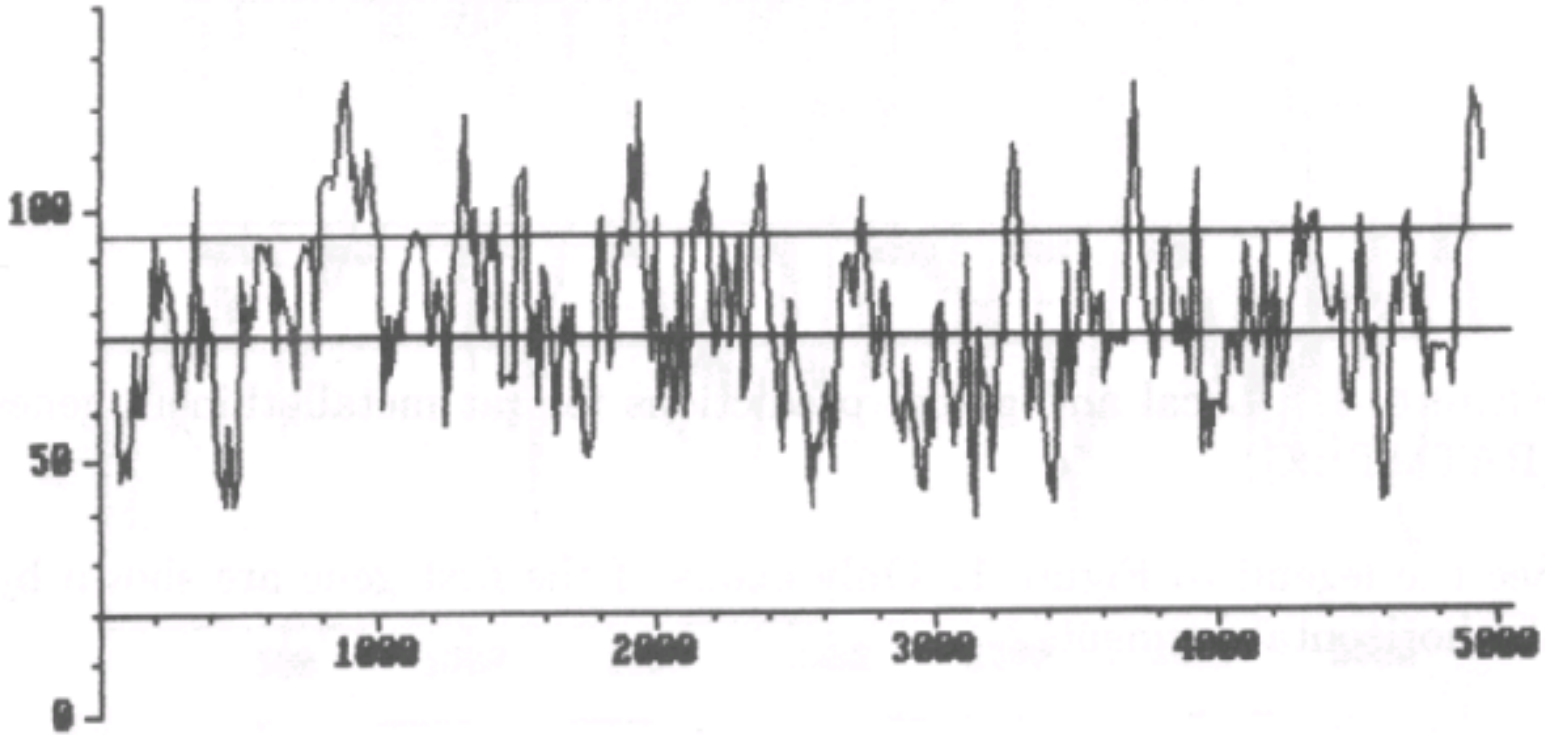
# Бета-глобин человека



# Хемотрипсин крысы



# … ничего … (28S рРНК человека)



# Статистические методы

- Скользящее окно не работает! (~ 1990)
- Статистика кодирующих и некодирующих областей + сайты сплайсинга – ещё одна вариация на тему динамического программирования

# Сайты сплайсинга

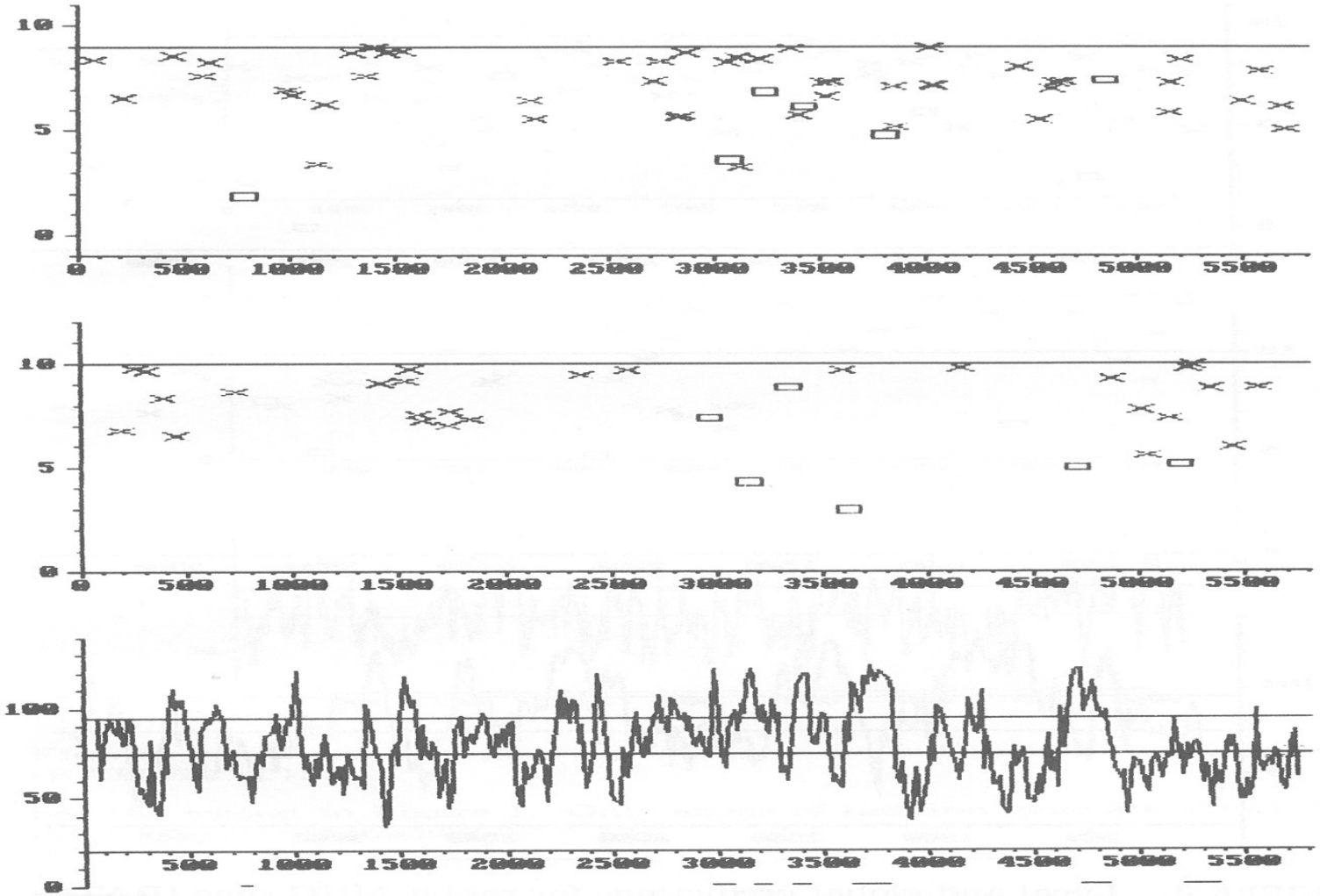
Donor sites

gtgggatgatgtaagtattggggcggcccg
tcaaaacaaggtaagaaatgaggtatgcct
agctcccaaggtaggaggttgagtgttgtg
agtggccaaggtatggtggatggaaattgc
tggaaaaagcgtaagtcactctaattttat
ctctcaaaaagtaagctttgtgagcatttc
atcttcaagggtgagcatgtgtgttatgct
tttcagaattgtaagagtacacattttaag
gccagaaaaggtcagtactttctttcacac
tacctcacaggtatgaattttctagttctt
atctttcaaggtagagtatatgaatgttac
atgtggattcgtaagtattcaacacattca
aaaatatccagtaagcagttctgatgtttg
ccaggagccggtgaggggctggtgggctct
aatggatgaggtgggtacttagggcttctg
atttcaaaaagtaagttttccctggagaaa
aatttgtagagtatccttgatttgacgaat
cagacaatgggtaagtacatgcttgttccc
gtctgttaaggtaggtataccccatcacaa
gttcaaaaaggttggtcacatgttcttgat
attcggccaggtatgggtagtgtgctgaga
acatatgcaggtaaacaacttaactcaaat
aaagaaagaggtgagagggtgttttaattt
ccagctccaggtaagccatctggaaagagc
gtcttaacaggtaaatgccacccttttcccc

Acceptor sites

gtttcttcttacatttctaggactcaacta
ttcacgtttttgccttccaggagacagagc
tttcaatatttattacccaggaccccaaat
gtgttatttacatttttcaggaatggacaa
tttttctgcttctccaacagctatactaaa
ttgttgtgttcacttcacagcatatatcgc
tccgttgttttatttcccagaatgattcaa
tggtttttcattgtttttagtggtgcaaaa
tctaacttcatttcctccaggacaaatatc
gttttgttggtgttttatagctggccaact
acatgtgttctcatttttaggaagtgatag
ctgttcttgttctcccttagcccaaagcag
atgcctttcatttctattagctggaatctg
ctgttattaaaatttgacaggagaagctga
ttttttattcctacttccaggggactgctg
tttgttgttgcttaactcagaaagaaataa
tacttaacatgatggtccagatataacaaa
cttgtgtttttgatactcagacctggctat
ttgatttattgattttctagattatttcag
gtccttaatgtcctttgtaggtggttcttc
gcattattctcaccttccaggctatcacta
aatatctcttccctatttagatgtcatcga
aaggatatttataattttaggctgatcctg
ttttatcttttatattacaggttctgtaaa
ttcatattcatttgttgcagaagtggaagc

# Распознавание сайтов сплайсинга



# Список потенциальных экзонов

```
(a)
   tccatgcagaccatggcggtacaggacatgccggtgcagctctgactt

(b)
      ATGCAGACCATGGCGgt
      ATGCAGACCATGGCGGTACAGGACATGCCGgt
      ATGCAGACCATGGCGGTACAGGACATGCCGGTGCAGCTCTGA
          agACCATGGCGgt
          agACCATGGCGGTACAGGACATGCCGgt
          agACCATGGCGGTACAGGACATGCCGGTGCAGCTCTGA
              ATGGCGgt
              ATGGCGGTACAGGACATGCCGgt
              ATGGCGGTACAGGACATGCCGGTGCAGCTCTGA
                        agGACATGCCGgt
                        agGACATGCCGGTGCAGCTCTGA
                            ATGCCGgt
                            ATGCCGGTGCAGCTCTGA
                                      agCTCTGA
```

# Граф динамического программирования

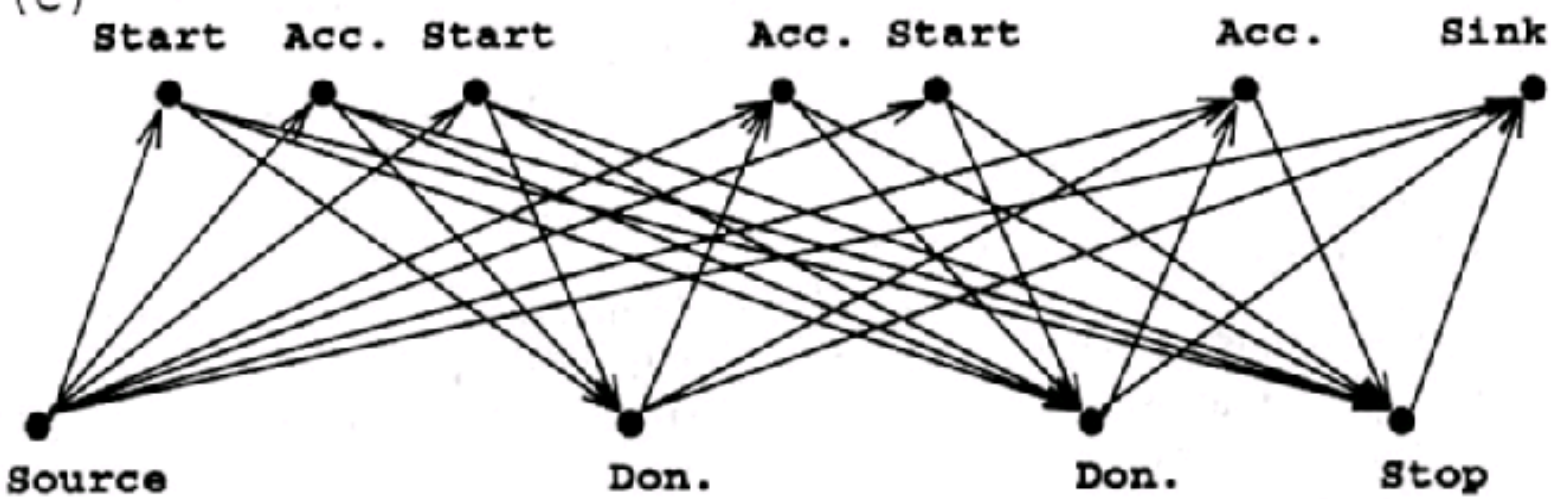
(a)

tccatgcagaccatggcggtacaggacatgccggtgcagctctgactt

(b)

ATGCAGACCATGGCGgt
ATGCAGACCATGGCGGTACAGGACATGCCGgt
ATGCAGACCATGGCGGTACAGGACATGCCGGTGCAGCTCTGA
agACCATGGCGgt
agACCATGGCGGTACAGGACATGCCGgt
agACCATGGCGGTACAGGACATGCCGGTGCAGCTCTGA
ATGGCGgt
ATGGCGGTACAGGACATGCCGgt
ATGGCGGTACAGGACATGCCGGTGCAGCTCTGA
agGACATGCCGgt
agGACATGCCGGTGCAGCTCTGA
ATGCCGgt
ATGCCGGTGCAGCTCTGA
agCTCTGA

(c)

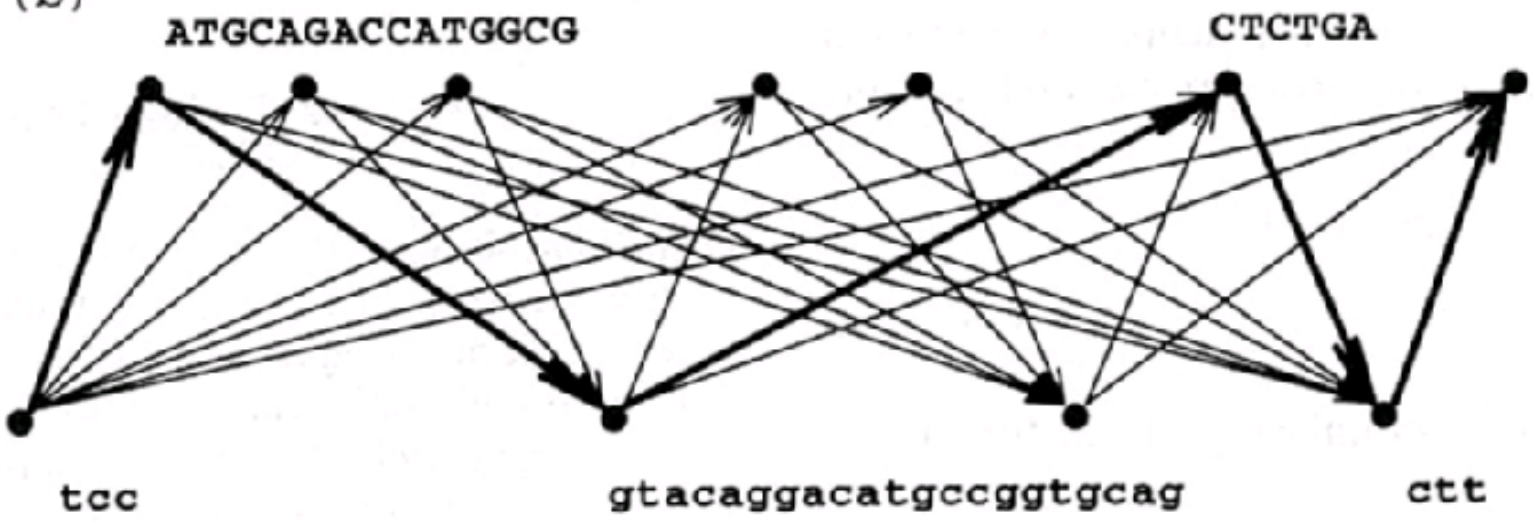


# Путь = экзон-интронная структура

(a)

tcc**ATG**CAGACCATGGCG**gt**acaggacatgccggtgc**ag**CTC**TGA**ctt

(b)



# Gen-Scan

Файл Правка Вид Избранное Сервис Справка

Назад Адрес: http://genes.mit.edu/GENSCAN.html

## The New GENSCAN Web Server at MIT

### Identification of complete gene structures in genomic DNA

```
              \\|//
              (o o)
-. .-.   .-oOOo~(_)~oOOo-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.   .-. .-.
|||\|||\ /|||\|||\ /|||\|||\ /|||\|||\ /|||\|||\ /|||\|||\ /|||\|||\ /|||\|||\ /|||\|||\ /|||\|||\ /|||\|||\
|/ \|||\|||/ \|||\|||/ \|||\|||/ \|||\|||/ \|||\|||/ \|||\|||/ \|||\|||/ \|||\|||/ \|||\|||/ \|||\|||/ \|||
'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-' `-'   `-'
```

For information about Genscan, click here

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

**News:**

- **The server is back online. Apologies for the delay.**

- The interface now supports file upload, allowing you to input long sequences without cutting and pasting. The old cut-and-paste method also works.
- Also, the server now outputs PDF (portable document format) images of predicted gene locations, as well as Postscript.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*e.g.*, Lynx) does not support file upload or multipart forms, use the older version.

Organism: Vertebrate Suboptimal exon cutoff (optional): 1.00

Vertebrate
Arabidopsis
Maize

Sequence n

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored): Обзор...

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

# Сравнительные методы

- BLASTN: ESTs и альтернативный сплайсинг
- BLASTX
- BLASTX+статистика
- Сравнение с известными белками
- Геномные сравнения
  - выравнивание ДНК
  - выравнивание белков
- All of the above and more…

# ESTs: короткие фрагменты (клонированной) мРНК

- Характерная длина ~300 нт
- Ошибки секвенирования
- Ошибки клонирования
  - несплайсированный транскрипты
  - геномная ДНК
- Обогащение к 3’-концу (PolyA-праймеры)
- Альтернативный сплайсинг: 30-50% генов

# Human Genome Browser – поиск по имени гена

Human Genome Browser Gateway - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Адрес: http://bioinfo.hku.hk/cgi-bin/hgGateway

Home - Genome Browser - Blat - Table Browser - FAQ - Help

## Human Genome Browser Gateway

UCSC Genome Browser created by Jim Kent, Charles Sugnet, Terry Furey, Robert Baertsch, Heather Trumbower, Angie Hinrichs, Fan Hsu, Hiram Clawson, Kate Rosenbloom, Brian Raney, Robert Kuhn, Donna Karolchik, David Haussler, and the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

genome: Human
- Human
- Mouse
- Rat
- C. elegans
- C. briggsae
- SARS

assembly: July 2003

position: magec

clear

image width: 620

Submit

lick here to reset the browser user interface settings to their defaults.

Add Your Own Custom Tracks

### A mo sapiens assembly

The latest human reference sequence (UCSC version hg16) is based on NCBI Build 34 and was produced by the International Human Genome Sequencing Consortium. The sequence covers about 99 percent of the gene-containing regions in the genome, and has been sequenced to an accuracy of 99.99 percent. Of note in this release is the addition of the pseudoautosomal regions of the Y chromosome. This sequence was taken from the corresponding regions in the X chromosome and is an exact duplication of that sequence.

There are 2,843,433,602 finished sequenced bases in the ordered and oriented portion of the

# Результат

Human magec - UCSC Genome Browser v46 - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Адрес: http://bioinfo.hku.hk/cgi-bin/hgTracks?org=Human&db=hg16&position=magec&p

## Known Genes

MAGEC1 at chrX:139685281-139690771 - (AF056334) melanoma antigen, family C, 1
MAGEC2 at chrX:139983739-139986659 - (AF151378) melanoma antigen, family E, 1, cancer/testis specific

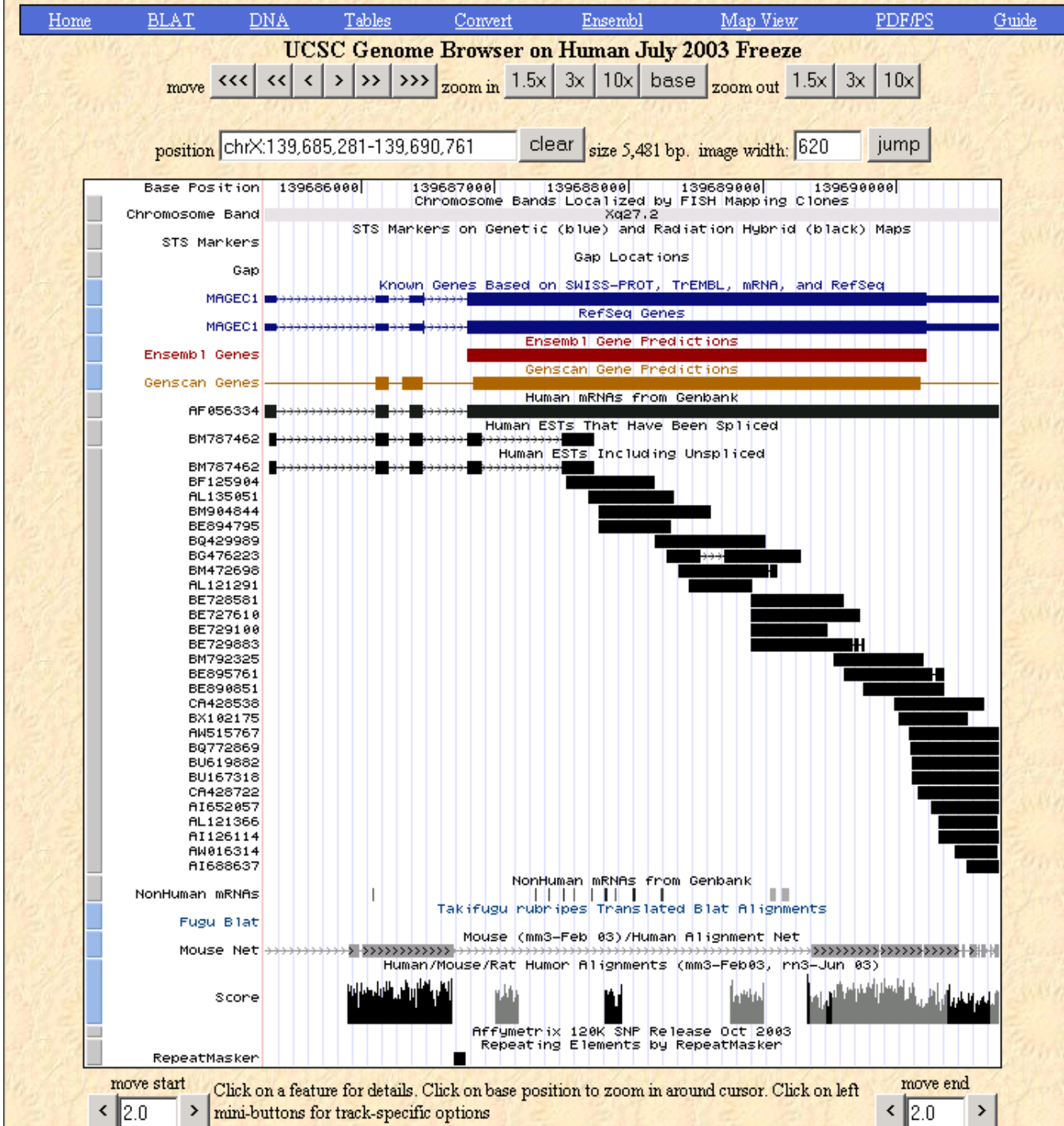
## RefSeq Genes

MAGEC1 at chrX:139685281-139690761 - (NM_005462) melanoma antigen, family C, 1
MAGEC3 at chrX:139619703-139679220 - (NM_138702) MAGE family testis and tumor-specific
MAGEC3 at chrX:139676705-139679220 - (NM_177456) MAGE family testis and tumor-specific

## Human Aligned mRNA Search Results

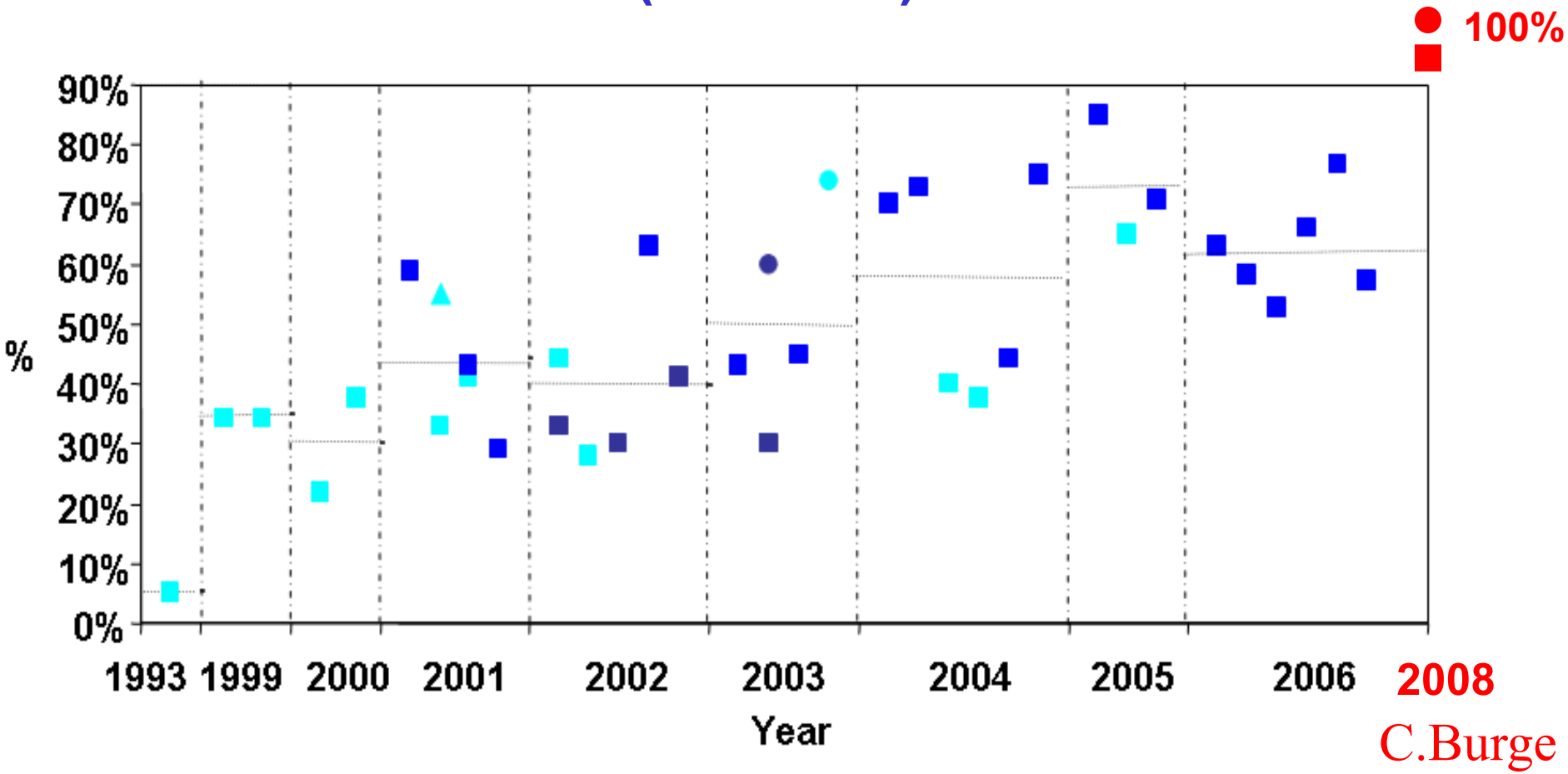
AF196482 - Homo sapiens MAGE-C2 (MAGEC2) mRNA, complete cds.

# MAGE-C1


Home
BLAT
DNA
Tables
Convert
Ensembl
Map View
PDF/PS
Guide
UCSC Genome Browser on Human July 2003 Freeze
move
zoom in
zoom out
position chrX:139,685,281-139,690,761
clear
size 5,481 bp. image width: 620
jump
Base Position
Chromosome Band
STS Markers
Gap
MAGEC1
Ensembl Genes
Genscan Genes
NonHuman mRNAs
Fugu Blat
Mouse Net
Score
RepeatMasker
Chromosome Bands Localized by FISH Mapping Clones
Xq27.2
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
Gap Locations
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
RefSeq Genes
Ensembl Gene Predictions
Genscan Gene Predictions
Human mRNAs from Genbank
Human ESTs That Have Been Spliced
Human ESTs Including Unspliced
NonHuman mRNAs from Genbank
Takifugu rubripes Translated Blat Alignments
Mouse (mm3-Feb 03)/Human Alignment Net
Human/Mouse/Rat Humor Alignments (mm3-Feb03, rn3-Jun 03)
Affymetrix 120K SNP Release Oct 2003
Repeating Elements by RepeatMasker
move start
move end
Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options

# Ещё о сравнении предсказаний

# Альтернативный сплайсинг генов человека (и мыши)



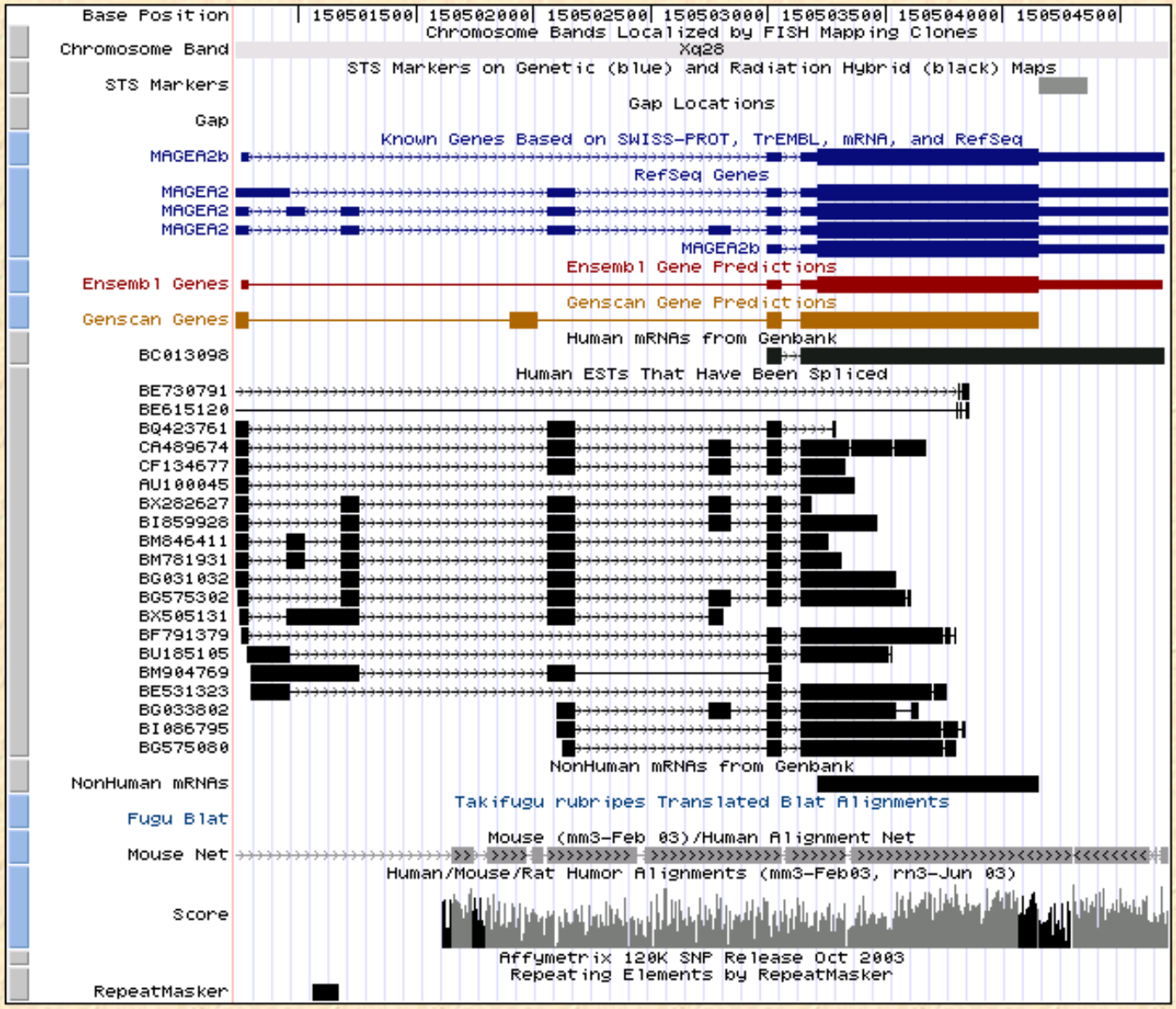

- Human (genome / random sample)
- Human (individual chromosomes)
- Mouse (genome / random sample)

- All genes
- Only multiexon genes
- Genes with high EST coverage

# MAGE-A2

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position chrX:150,500,731-150,504,708 clear size 3,978 bp. image width: 620 jump

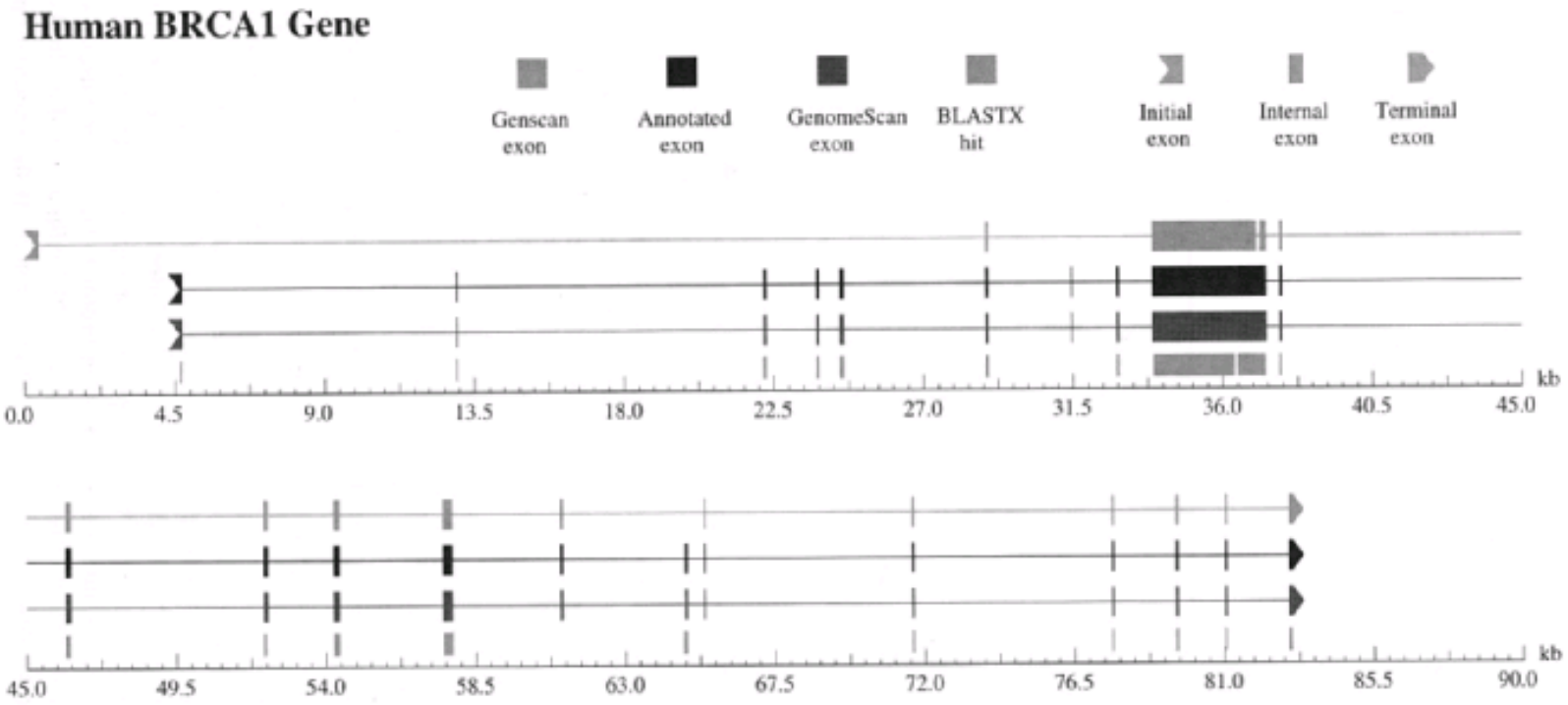



move start < 2.0 >

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for track-specific options

move end < 2.0 >

# GenomeScan=GenScan+BLASTX

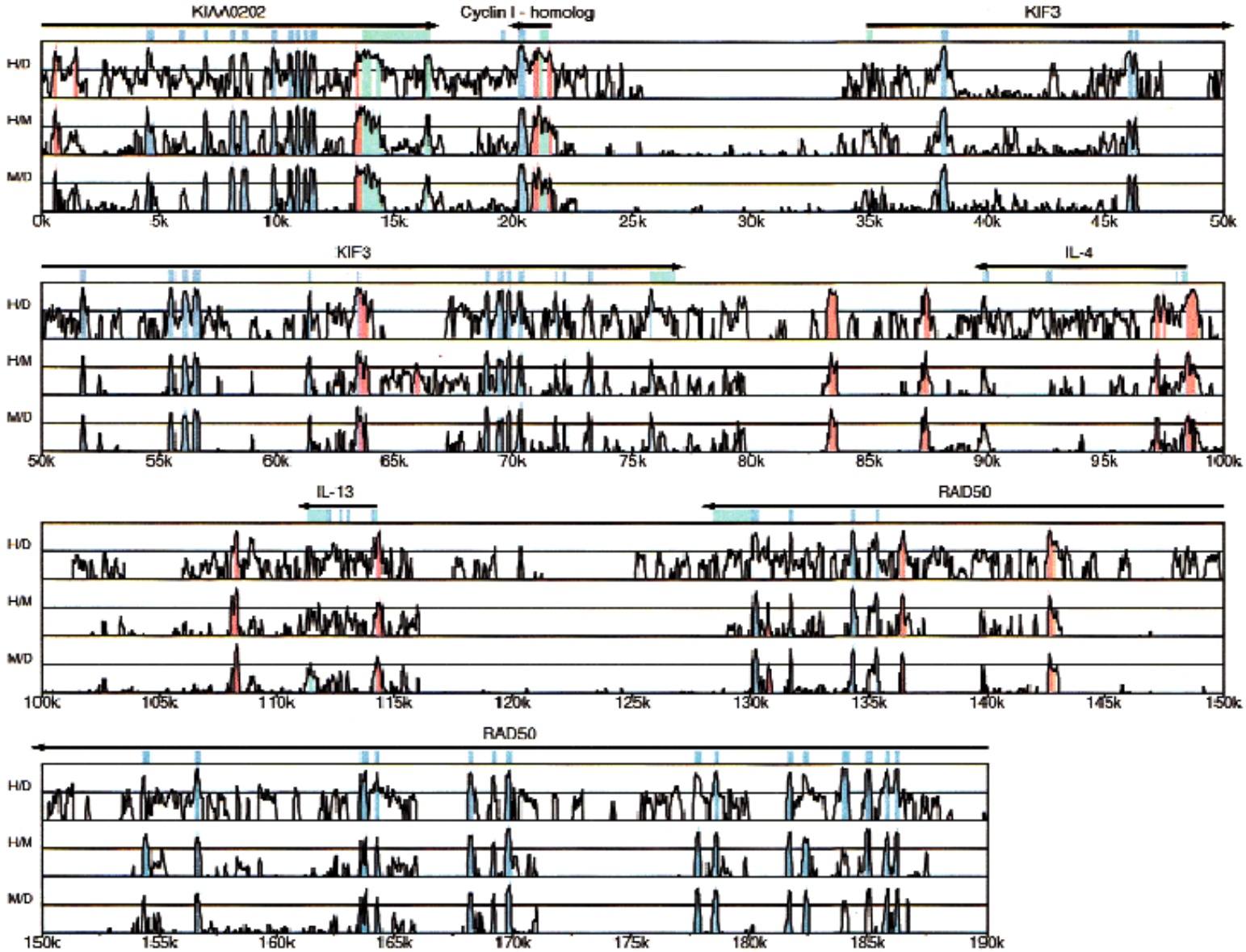


# Сплайсированное выравнивание

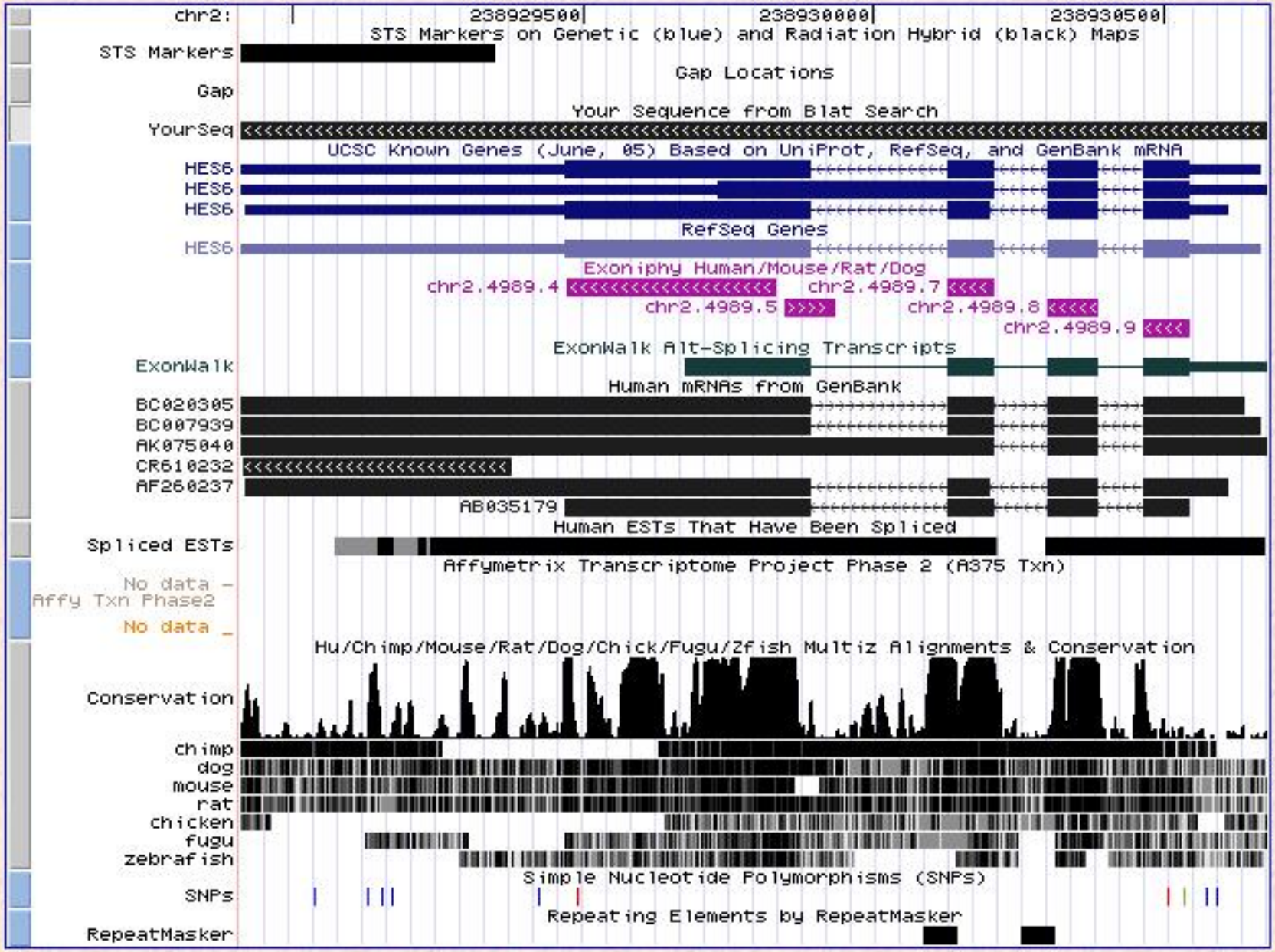
Сравнение (формально транслированной) ДНК с аминокислотной последовательностью родственного белка.

- Динамическое программирование, дополнительная операция – интрон
  - Только на потенциальных сайтах сплайсинга
  - Небольшой штраф
  - Учёт особенностей экзон-интронной структуры – минимальная длина интрона (зависит от генома)

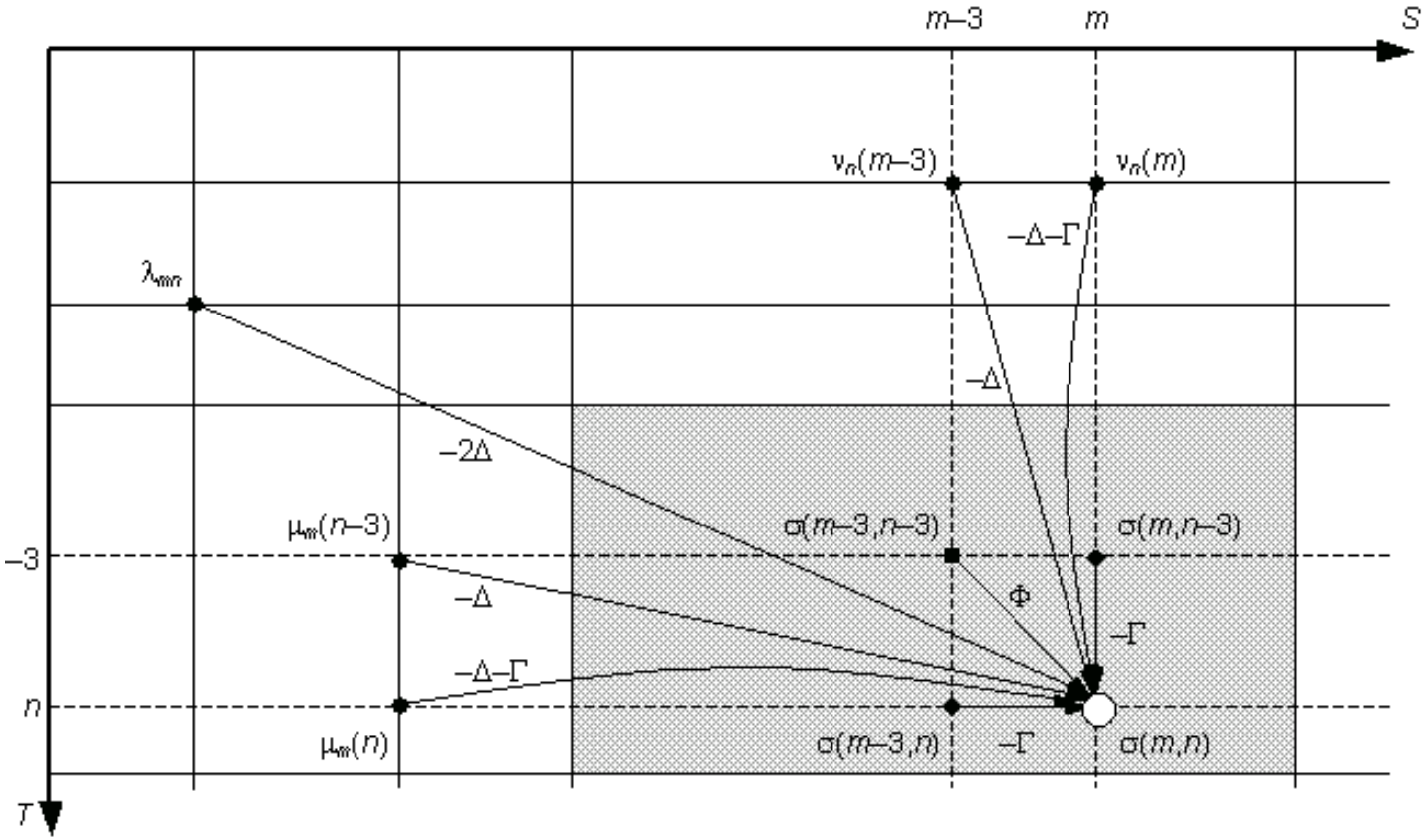
# Геномное сравнение. VISTA (human-dog-mouse)



# HGB: mRNAs, ESTs, repeats, conservation



# Сплайсированое выравнивание геномных последовательностей



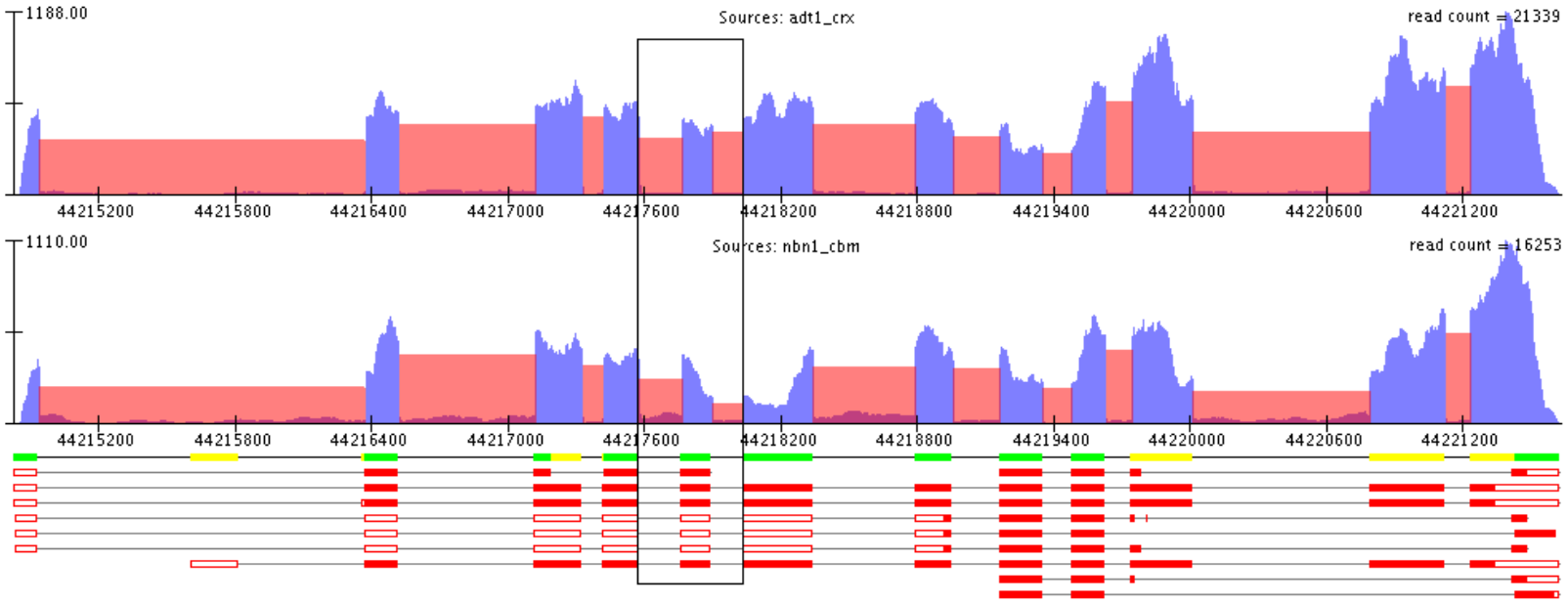
# RNA-Seq

- Тотальное секвенирование транскриптомов с очень большим покрытием.
- Картирование на геном.
- Картирование на всевозможные комбинации экзонов (в правильном порядке)

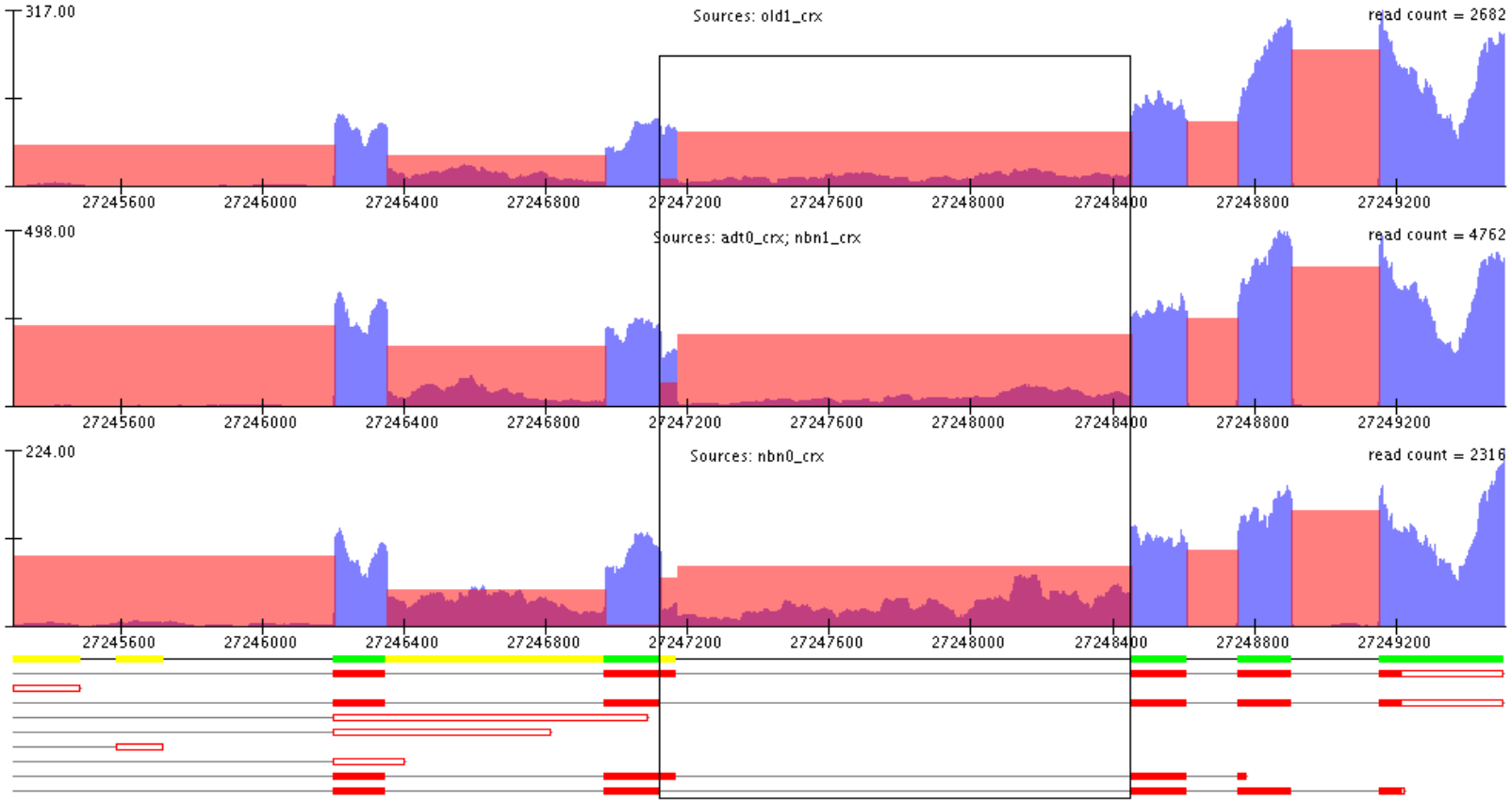
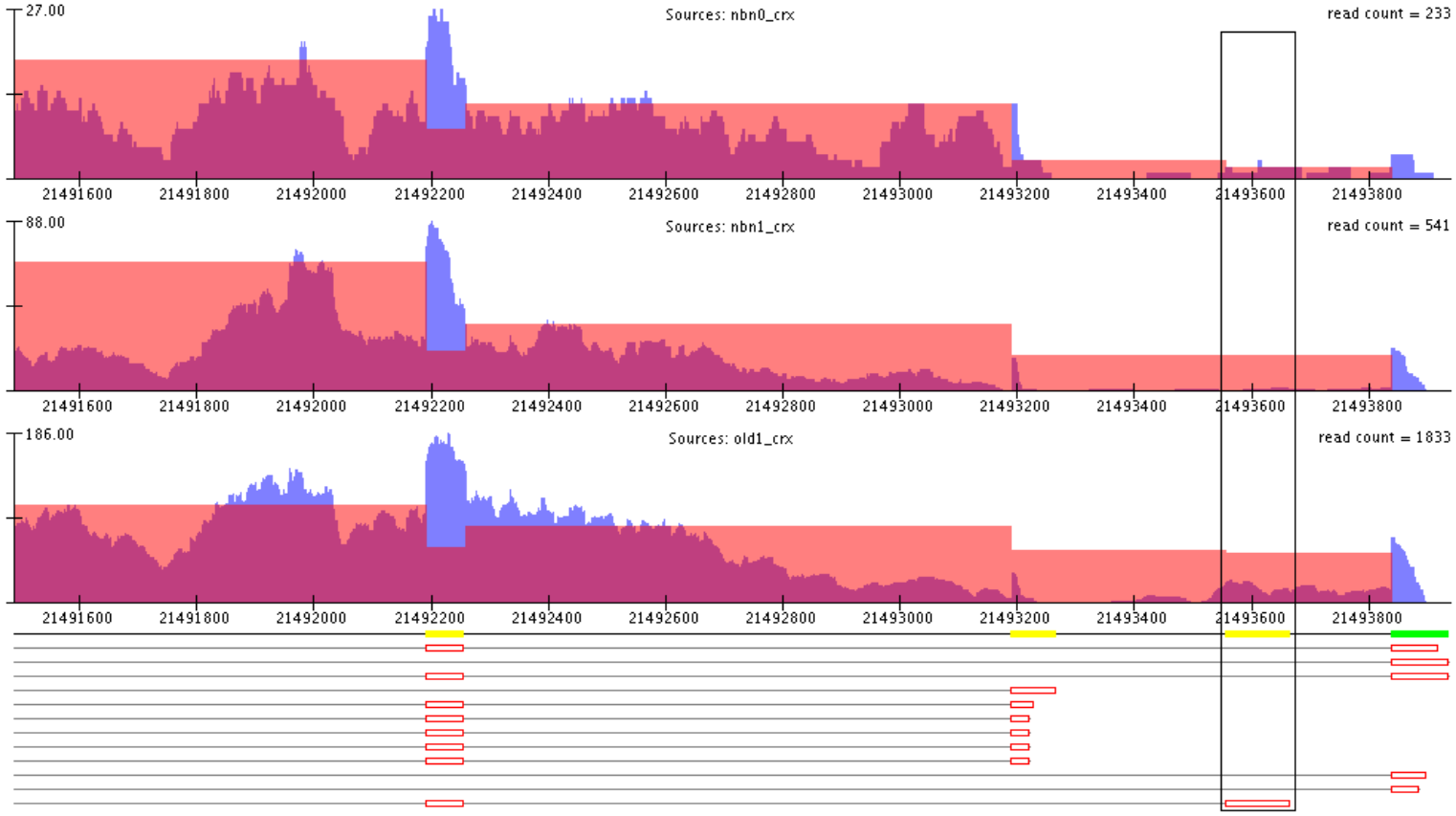
# Методы (и чего от них ждать)

- Roche, 454 (2004)
  - До 400-500 нт (парные)
  - Плохое разрешение polyN
  - 100 мегабаз за проход (обещано 500)
  - Дорого ($1000 за мегабазу)
- Solexa/Illumina, GA II (2005) (уже есть HiSeq)
  - 75 (в обзоре 35, обещано 100) нт (парные)
  - 50 млн. фрагментов за проход (до 3 гигабаз), 4 дня
  - $500 за мегабазу (?? – сейчас явно меньше, ~30 тыс за проход)
- Applied Biosystems, Solid 2.0 (2007)
  - 35 нт, одиночные
  - 3-10 гигабаз за проход, 5-9 дней
  - $10 за мегабазу
- Helicos (2009). Single molecule (no amplification)
  - Доллар за мегабазу
  - 23 нт (кажется, уже 35).
  - Делеции в polyN
  - Миллионы молекул, большие гигабазы

# Хорошее соответствие экзонов и покрытых участков



# Ненулевое покрытие интронов, провалы в экзонах



# Ненулевое покрытие интронов – сравнимо с экзонами. Пропуски экзонов – альтернативный сплайсинг?