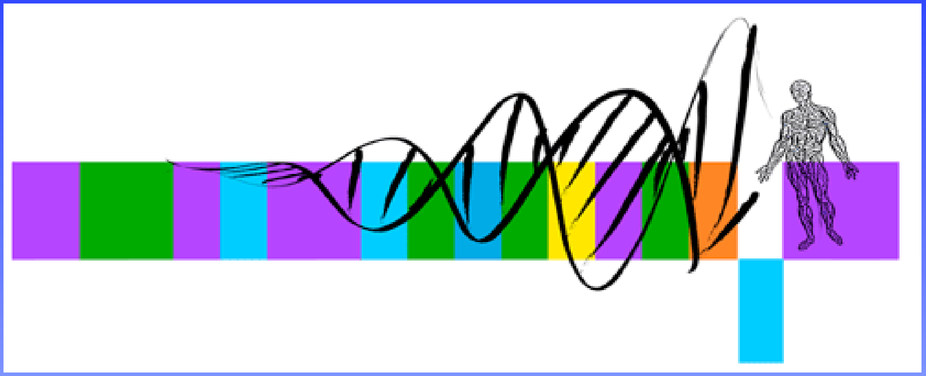


Bioinformatics for Biologists

Sequence Analysis: Part II. Pattern Searching and Gene Finding

Fran Lewitter, Ph.D.
Director
Bioinformatics & Research Computing
Whitehead Institute

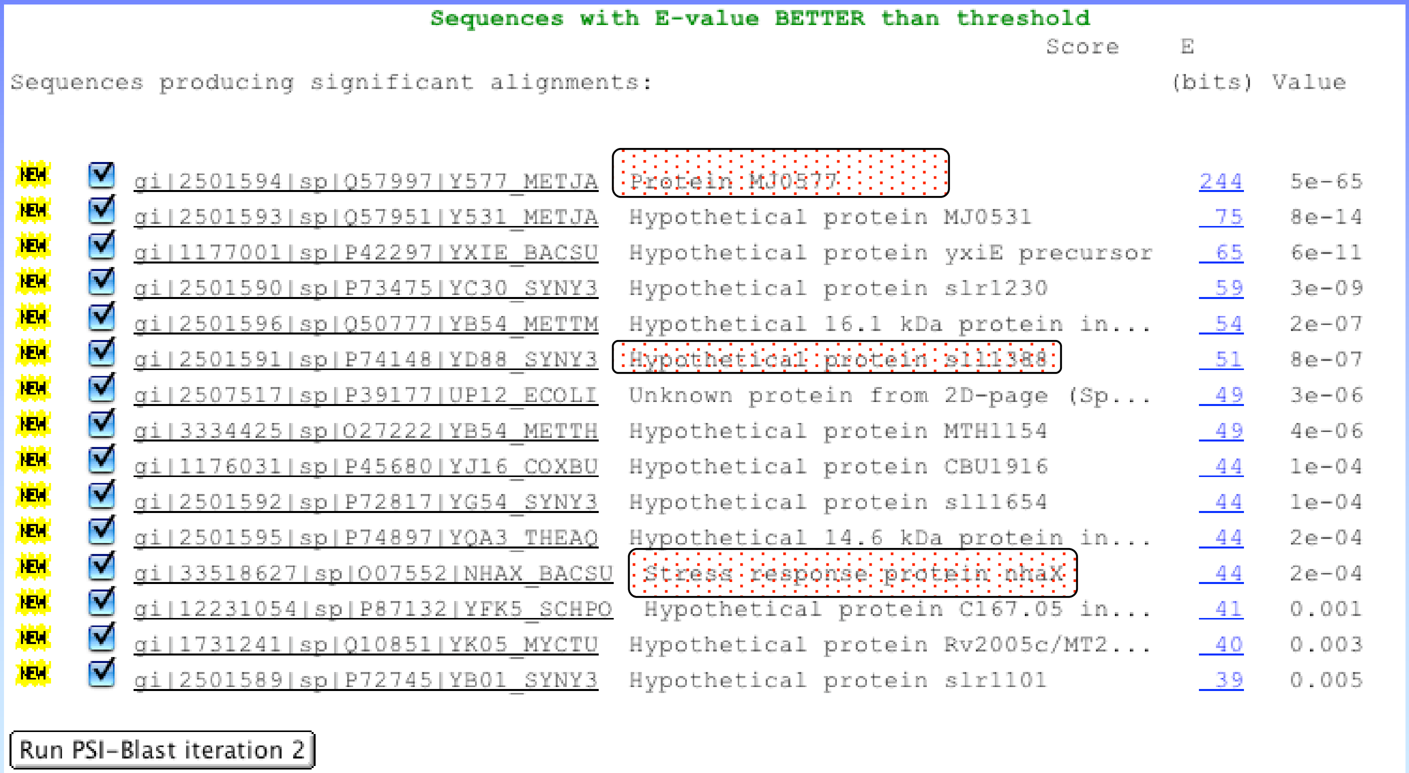
Topics to Cover

- Pattern searching
  - PSI-BLAST
  - PHI-BLAST
  - Finding patterns
- Gene finding

PSI-BLAST

- **<u>P</u>**osition **<u>S</u>**pecific **<u>I</u>**terative BLAST uses a profile (or position specific scoring matrix, PSSM) that is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search.
- The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero.
- The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" is used to refine the profile. This iterative searching strategy results in increased sensitivity.

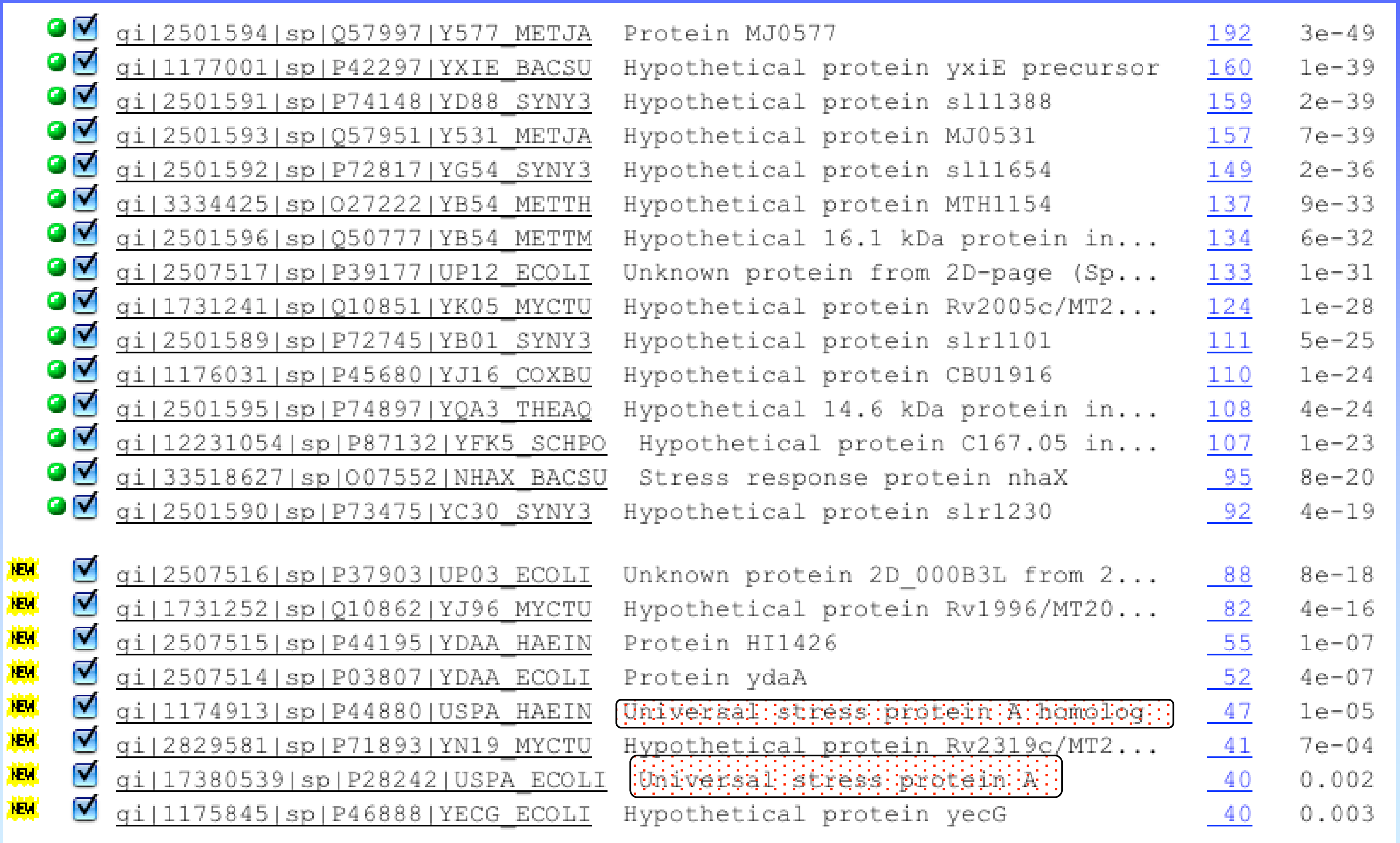
Start with a BLASTP search



```
                      Sequences with E-value BETTER than threshold
                                                                  Score    E
Sequences producing significant alignments:                      (bits) Value


NEW  ☑  gi|2501594|sp|Q57997|Y577_METJA  Protein MJ0577                   244   5e-65
NEW  ☑  gi|2501593|sp|Q57951|Y531_METJA  Hypothetical protein MJ0531       75   8e-14
NEW  ☑  gi|1177001|sp|P42297|YXIE_BACSU  Hypothetical protein yxiE precursor 65   6e-11
NEW  ☑  gi|2501590|sp|P73475|YC30_SYNY3  Hypothetical protein slr1230      59   3e-09
NEW  ☑  gi|2501596|sp|Q50777|YB54_METTM  Hypothetical 16.1 kDa protein in...  54   2e-07
NEW  ☑  gi|2501591|sp|P74148|YD88_SYNY3  Hypothetical protein sll1388      51   8e-07
NEW  ☑  gi|2507517|sp|P39177|UP12_ECOLI  Unknown protein from 2D-page (Sp...  49   3e-06
NEW  ☑  gi|3334425|sp|O27222|YB54_METTH  Hypothetical protein MTH1154      49   4e-06
NEW  ☑  gi|1176031|sp|P45680|YJ16_COXBU  Hypothetical protein CBU1916      44   1e-04
NEW  ☑  gi|2501592|sp|P72817|YG54_SYNY3  Hypothetical protein sll1654      44   1e-04
NEW  ☑  gi|2501595|sp|P74897|YQA3_THEAQ  Hypothetical 14.6 kDa protein in...  44   2e-04
NEW  ☑  gi|33518627|sp|O07552|NHAX_BACSU  Stress response protein nhaX    44   2e-04
NEW  ☑  gi|12231054|sp|P87132|YFK5_SCHPO  Hypothetical protein C167.05 in...  41   0.001
NEW  ☑  gi|1731241|sp|Q10851|YK05_MYCTU  Hypothetical protein Rv2005c/MT2...  40   0.003
NEW  ☑  gi|2501589|sp|P72745|YB01_SYNY3  Hypothetical protein slr1101      39   0.005
```

PSI-BLAST - Iteration 1



●	☑	gi|2501594|sp|Q57997|Y577_METJA	Protein MJ0577	192	3e-49
●	☑	gi|1177001|sp|P42297|YXIE_BACSU	Hypothetical protein yxiE precursor	160	1e-39
●	☑	gi|2501591|sp|P74148|YD88_SYNY3	Hypothetical protein sll1388	159	2e-39
●	☑	gi|2501593|sp|Q57951|Y531_METJA	Hypothetical protein MJ0531	157	7e-39
●	☑	gi|2501592|sp|P72817|YG54_SYNY3	Hypothetical protein sll1654	149	2e-36
●	☑	gi|3334425|sp|O27222|YB54_METTH	Hypothetical protein MTH1154	137	9e-33
●	☑	gi|2501596|sp|Q50777|YB54_METTM	Hypothetical 16.1 kDa protein in...	134	6e-32
●	☑	gi|2507517|sp|P39177|UP12_ECOLI	Unknown protein from 2D-page (Sp...	133	1e-31
●	☑	gi|1731241|sp|Q10851|YK05_MYCTU	Hypothetical protein Rv2005c/MT2...	124	1e-28
●	☑	gi|2501589|sp|P72745|YB01_SYNY3	Hypothetical protein slr1101	111	5e-25
●	☑	gi|1176031|sp|P45680|YJ16_COXBU	Hypothetical protein CBU1916	110	1e-24
●	☑	gi|2501595|sp|P74897|YQA3_THEAQ	Hypothetical 14.6 kDa protein in...	108	4e-24
●	☑	gi|12231054|sp|P87132|YFK5_SCHPO	Hypothetical protein C167.05 in...	107	1e-23
●	☑	gi|33518627|sp|O07552|NHAX_BACSU	Stress response protein nhaX	95	8e-20
●	☑	gi|2501590|sp|P73475|YC30_SYNY3	Hypothetical protein slr1230	92	4e-19
NEW	☑	gi|2507516|sp|P37903|UP03_ECOLI	Unknown protein 2D_000B3L from 2...	88	8e-18
NEW	☑	gi|1731252|sp|Q10862|YJ96_MYCTU	Hypothetical protein Rv1996/MT20...	82	4e-16
NEW	☑	gi|2507515|sp|P44195|YDAA_HAEIN	Protein HI1426	55	1e-07
NEW	☑	gi|2507514|sp|P03807|YDAA_ECOLI	Protein ydaA	52	4e-07
NEW	☑	gi|1174913|sp|P44880|USPA_HAEIN	Universal stress protein A homolog	47	1e-05
NEW	☑	gi|2829581|sp|P71893|YN19_MYCTU	Hypothetical protein Rv2319c/MT2...	41	7e-04
NEW	☑	gi|17380539|sp|P28242|USPA_ECOLI	Universal stress protein A	40	0.002
NEW	☑	gi|1175845|sp|P46888|YECG_ECOLI	Hypothetical protein yecG	40	0.003

PSSM from PSI-BLAST

Amino acids

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	0	2	3	2	4	1	1	4	3	0	3	3	7	3	3	2	1	0	1	2
2	6	0	3	3	5	4	0	3	2	5	0	1	2	2	4	1	3	2	4	2
3	4	3	0	3	3	1	3	2	4	2	3	2	5	0	1	2	1	0	5	7
4	3	2	3	2	4	9	3	3	5	4	0	3	2	5	1	2	2	4	1	2
5	0	1	2	2	4	1	6	3	3	1	3	2	0	4	8	3	1	0	3	0
6	4	3	2	…																
•	…																			
•	…																			
N																				

POSITIONS

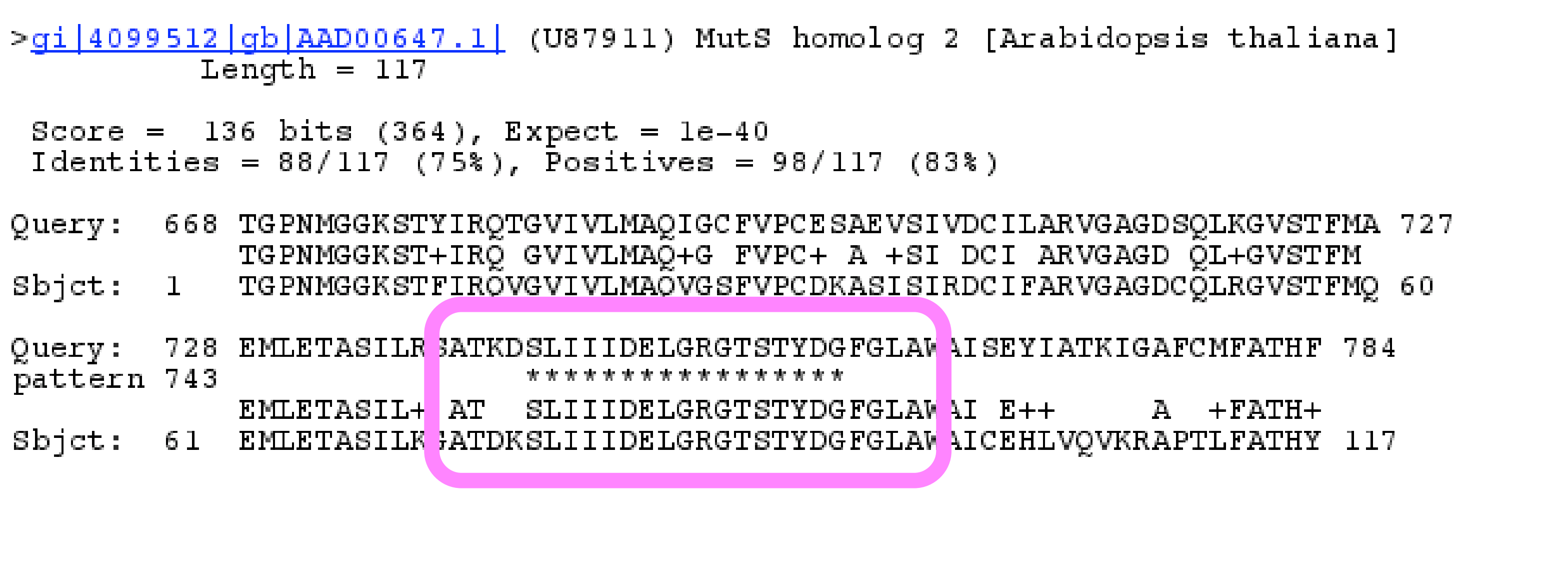
Pattern Hit Initiated (PHI)-BLAST

```
>HUMAN MSH2

MAVQPKETLQLESAAEVGFVRFFQGMPEKPTTTVRLFDRGDFYTAHGEDALLAAREVFKTQGVIKYMGPA
GAKNLQSVVLSKMNFESFVKDLLLVRQYRVEVYKNRAGNKASKENDWYLAYKASPGNLSQFEDILFGNND
MSASIGVVGVKMSAVDGQRQVGVGYVDSIQRKLGLCEFPDNDQFSNLEALLIQIGPKECVLPGGETAGDM
GKLRQIIQRGGILITERKKADFSTKDIYQDLNRLLKGKKGEQMNSAVLPEMENQVAVSSLSAVIKFLELL
SDDSNFGQFELTTFDFSQYMKLDIAAVRALNLFQGSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPL
MDKNRIEERLNLVEAFVEDAELRQTLQEDLLRRFPDLNRLAKKFQRQAANLQDCYRLYQGINQLPNVIQA
LEKHEGKHQKLLLAVFVTPLTDLRSDFSKFQEMIETTLDMDQVENHEFLVKPSFDPNLSELREIMNDLEK
KMQSTLISAARDLGLDPGKQIKLDSSAQFGYYFRVTCKEEKVLRNNKNFSTVDIQKNGVKFTNSKLTSLN
EEYTKNKTEYEEAQDAIVKEIVNISSGYVEPMQTLNDVLAQLDAVVSFAHVSNGAPVPYVRPAILEKGQG
RIILKASRHACVEVQDEIAFIPNDVYFEKDKQMFHIITGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESA
EVSIVDCILARVGAGDSQLKGVSTFMAEMLETASILRSATKDSLIIIDELGRGTSTYDGFGLAWAISEYI
ATKIGAFCMFATHFHELTALANQIPTVNNLHVTALTTEETLTMLYQVKKGVCDQSFGIHVAELANFPKHV
             FQYIGESQGYDIMEPAAKKCYLEREQGEKIIQEFLSKVKQMPFTEMSEENITIKLKQ
             NEIISRIKVTT
```

DNA mismatch repair proteins mutS family signature

PHI-BLAST



```
>gi|4099512|gb|AAD00647.1| (U87911) MutS homolog 2 [Arabidopsis thaliana]
          Length = 117

 Score =  136 bits (364), Expect = 1e-40
 Identities = 88/117 (75%), Positives = 98/117 (83%)

Query:  668 TGPNMGGKSTYIRQTGVIVLMAQIGCFVPCESAEVSIVDCILARVGAGDSQLKGVSTFMA 727
            TGPNMGGKST+IRQ GVIVLMAQ+G FVPC+ A +SI DCI ARVGAGD QL+GVSTFM
Sbjct:  1   TGPNMGGKSTFIRQVGVIVLMAQVGSFVPCDKASISIRDCIFARVGAGDCQLRGVSTFMQ 60

Query:  728 EMLETASILRGATKDSLIIIDELGRGTSTYDGFGLAWAISEYIATKIGAFCMFATHF 784
pattern 743                *****************
            EMLETASIL+GAT  SLIIIDELGRGTSTYDGFGLAWAI E++      A  +FATH+
Sbjct:  61  EMLETASILKGATDKSLIIIDELGRGTSTYDGFGLAWAICEHLVQVKRAPTLFATHY 117
```

Pattern Searching

RRRRYYYY or R(4)Y(4)	4 purines followed by 4 pyrimidines
TATAA	Exact pattern: TATAA
[DE](2)HS{P}X(2)PX(2,4)C	[ ] Acceptable { } Not acceptable X(2) 2 of anything in a row X(2,4) from 2 to four of of anything
p1=6...8 GAGA ~p1	hairpin with GAGA as the loop
p1=6...6 3...8 p1	exact 6 character repeat separated by up to 8
p1=6...6 3..8 p1[1,1,1]	allow one mismatch, deletion and insertion

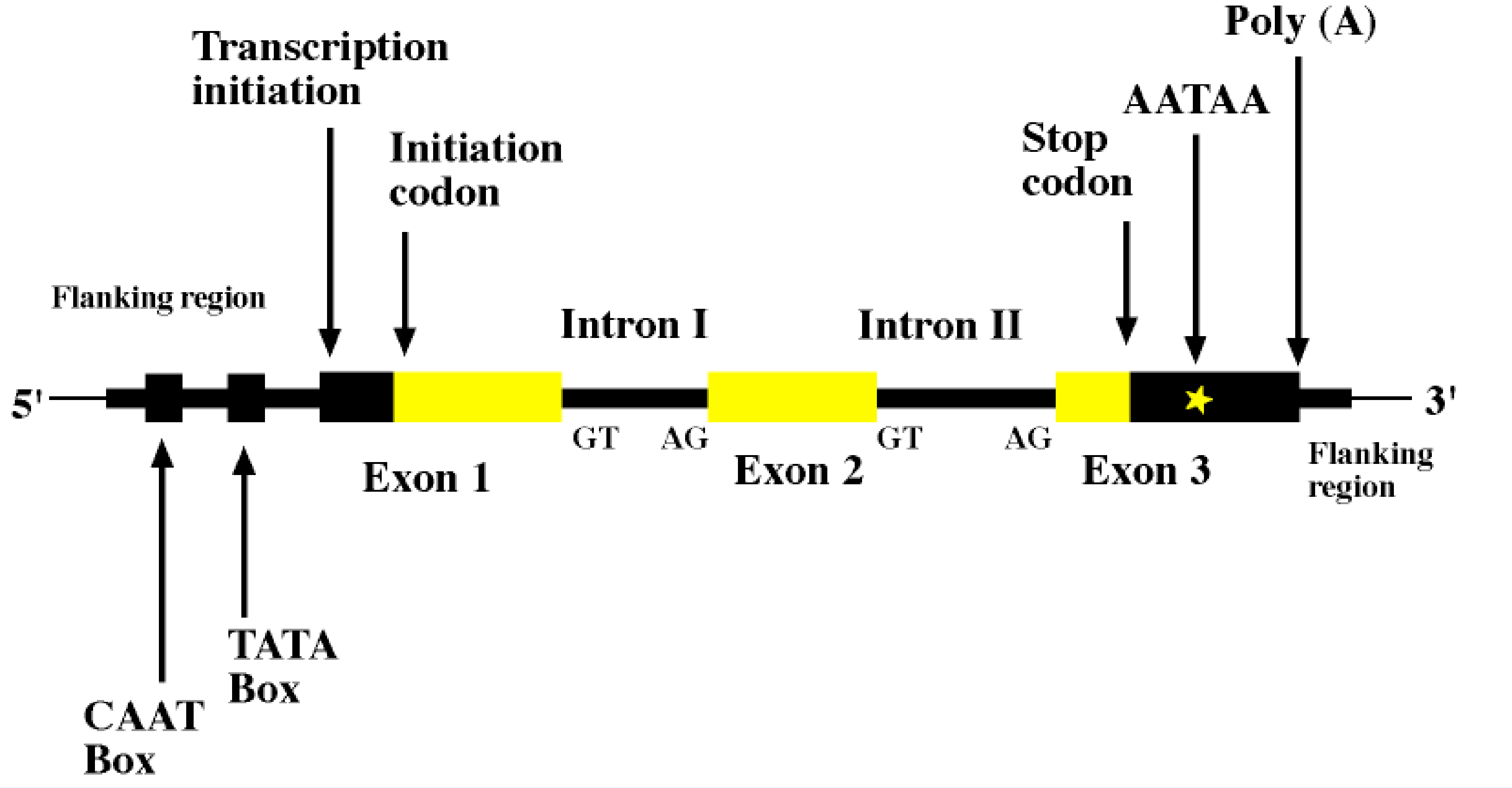
Pattern Searching Programs

Patscan	scan_for_matches patfile < inputfile
fuzznuc, fuzzprot, fuzztrans, dreg	EMBOSS programs; web and Unix

Interesting features in DNA

- ***Structure of genes*** - exons, introns, etc.
- ***Non-coding RNAs*** - including micro-RNAs and other small RNAs
- ***Promoter sites***
- ***Alternative splice forms***

Problem to Solve



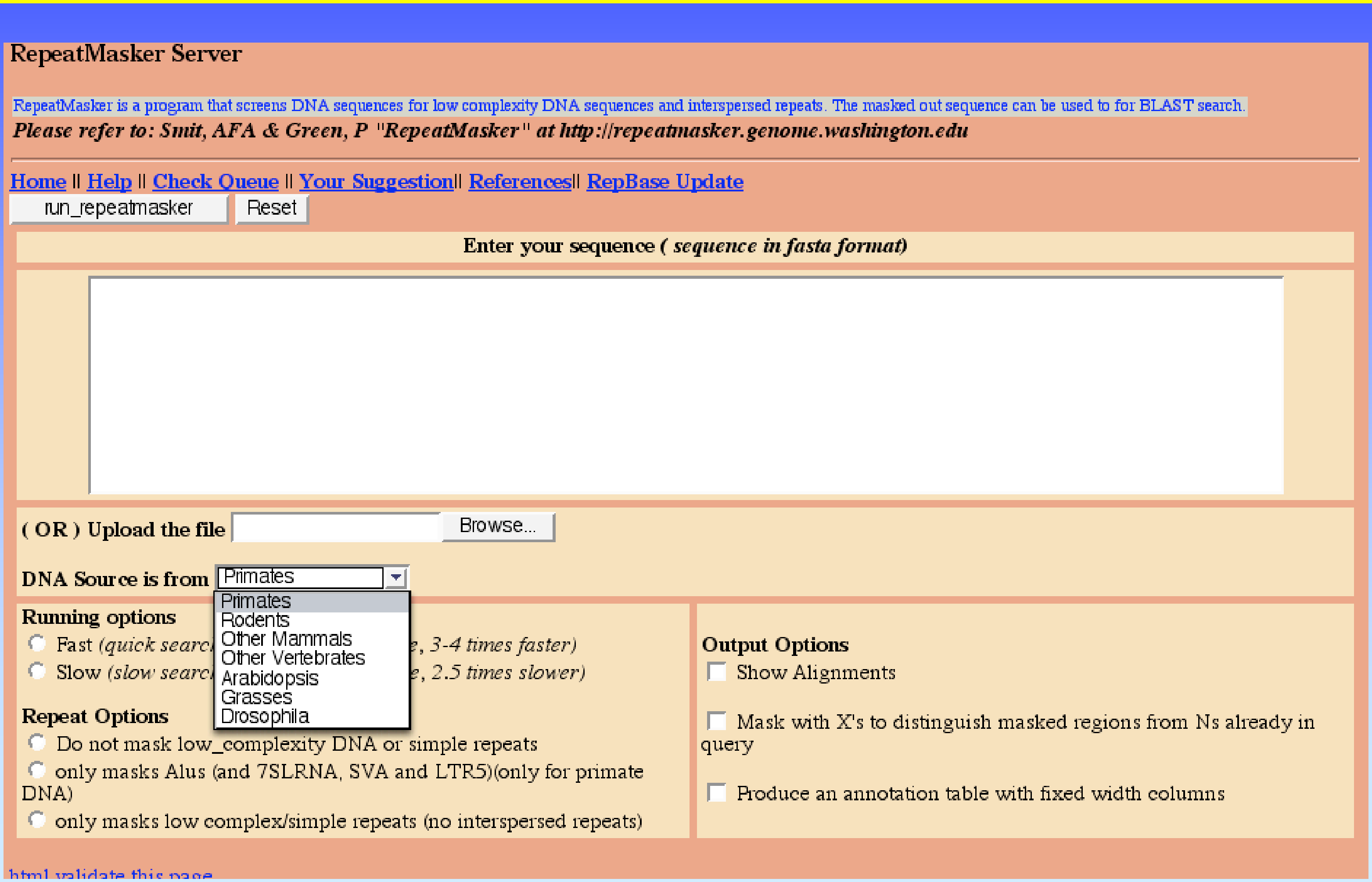
Types of Signals to Detect

- Transcriptional
  - TSS
  - TATA box
  - PolyA
- Translational
  - Kozak (CC A/G CCAUGG)
  - Termination codon (UAA, UAG, UGA)
- Splicing
  - Introns - GT……AG

Gene Finding Strategies

- Content-based methods
  - codon usage, compositional complexity
- Site-based methods
  - presence or absence of specific pattern or sequence
- Comparative methods
  - determination based on homology

RepeatMasker



RepeatMasker Server

RepeatMasker is a program that screens DNA sequences for low complexity DNA sequences and interspersed repeats. The masked out sequence can be used to for BLAST search.

Please refer to: Smit, AFA & Green, P "RepeatMasker" at http://repeatmasker.genome.washington.edu

Home || Help || Check Queue || Your Suggestion|| References|| RepBase Update
run_repeatmasker
Reset

Enter your sequence (*sequence in fasta format)*

(OR) Upload the file

DNA Source is from Primates

Running options

○ Fast *(quick searc* ... *e, 3-4 times faster)*

○ Slow *(slow searc* ... *e, 2.5 times slower)*

Repeat Options

○ Do not mask low_complexity DNA or simple repeats

○ only masks Alus (and 7SLRNA, SVA and LTR5)(only for primate DNA)

○ only masks low complex/simple repeats (no interspersed repeats)

Output Options

□ Show Alignments

□ Mask with X's to distinguish masked regions from Ns already in query

□ Produce an annotation table with fixed width columns

html validate this page

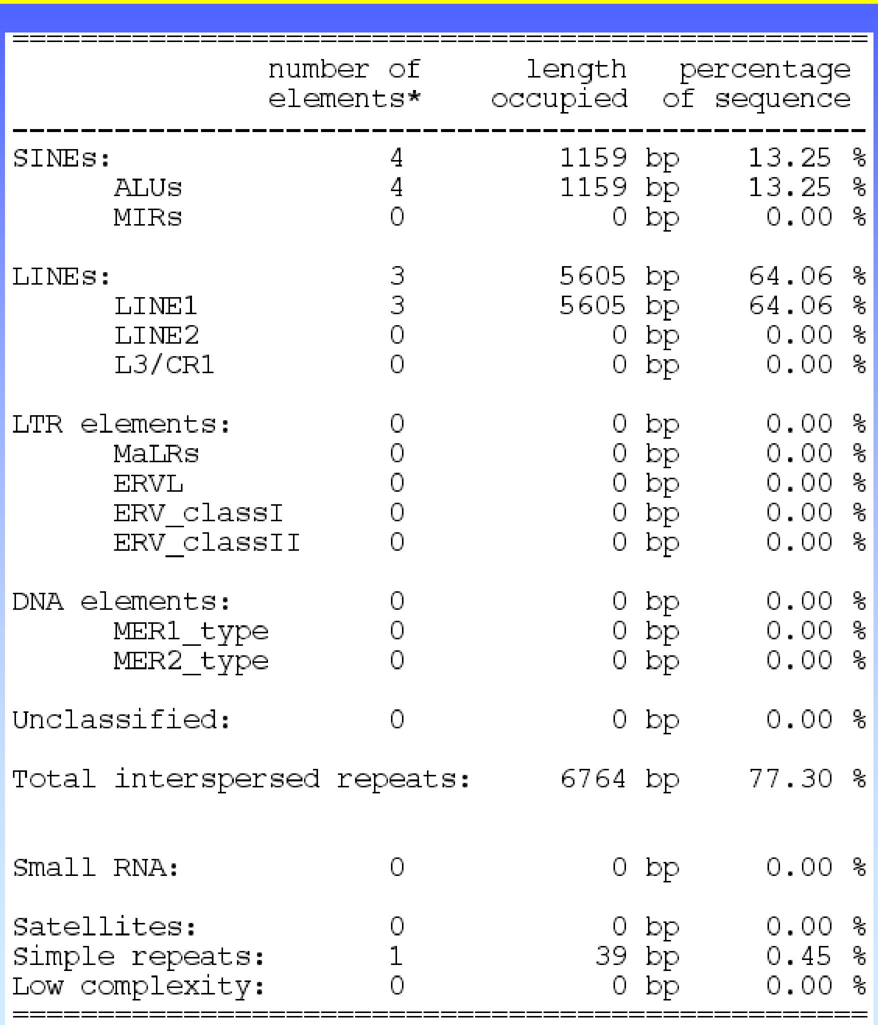
RepeatMasker

Summary:

Total length: 8750 bp

GC level: 35.61%

Bases masked: 6803 bp (77.75%)



```
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:                4         1159 bp    13.25 %
      ALUs            4         1159 bp    13.25 %
      MIRs            0            0 bp     0.00 %

LINEs:                3         5605 bp    64.06 %
      LINE1           3         5605 bp    64.06 %
      LINE2           0            0 bp     0.00 %
      L3/CR1          0            0 bp     0.00 %

LTR elements:         0            0 bp     0.00 %
      MaLRs           0            0 bp     0.00 %
      ERVL            0            0 bp     0.00 %
      ERV_classI      0            0 bp     0.00 %
      ERV_classII     0            0 bp     0.00 %

DNA elements:         0            0 bp     0.00 %
      MER1_type       0            0 bp     0.00 %
      MER2_type       0            0 bp     0.00 %

Unclassified:         0            0 bp     0.00 %

Total interspersed repeats:     6764 bp    77.30 %


Small RNA:            0            0 bp     0.00 %

Satellites:           0            0 bp     0.00 %
Simple repeats:       1           39 bp     0.45 %
Low complexity:       0            0 bp     0.00 %
==================================================
```

Coding Measures

- Look at frequencies of codons (e.g. redundancy of genetic code; Leucine = UUA, UUG, CUU, CUC, CUA, CUG)

- 6-tuple or hexamer approach
  ACCTCG TACTCG GCCCTC
  Thr Ser Tyr Ser Ala Leu

GENSCAN

- MM - prob for a given nuc to occur at position *p* depends on nuc occupying previous *k* positions
- Generalized Hidden Markov Model (GHMM)
- Optimize module performing signal recognition
- Incorporates influence of C+G content
- Considers gene models on both strands
- Can identify multiple genes

Burge and Karlin, JMB:268:78-94, 1997

Gene Finding Programs

- FGENESH - Softberry
- GeneID - Barcelona
- Genscan - Stanford and MIT
- GenomeScan - MIT
- MZEF/First exon - Cold Spring Harbor
- Twinscan - WU
- Ecgene - Korea (EST clustering)
- SGP - Barcelona (human-mouse homology)

HMR195 Test Set

- 103 human, 82 mouse, 10 rat sequences
- Sequence new since August, 1997
- Genomic sequences containing exactly one gene
- No mRNA sequences, pseudogenes or alternatively spliced genes
- The mean length of sequences is 7,096 bp

Rogic, Mackworth and Ouellette, Genome Research 11:817-832, 2001

Definitions

- ***Sensitivity:*** the proportion of true sites(e.g., exons or donor splice sites) that are correctly predicted = TP/(TP + FN)

- ***Predictive value of positive results ("Specificity" in gene finding literature):*** the proportion of predicted sites that are correct = TP/(TP + FP)

- ***Specificity:*** the proportion of non-sites that are predicted to be non-sites = TN/(TN +FP)

Program Comparisons Results

- Genscan and HMMgene had reliable scores for exons
- Nucleotide Sn = .95 for Genscan and .93 for HMMgene.
- Sp = .90 and .93, respectively
- Accuracy dependent on G+C content

Rogic, Mackworth and Ouellette, Genome Research 11:817-832, 2001

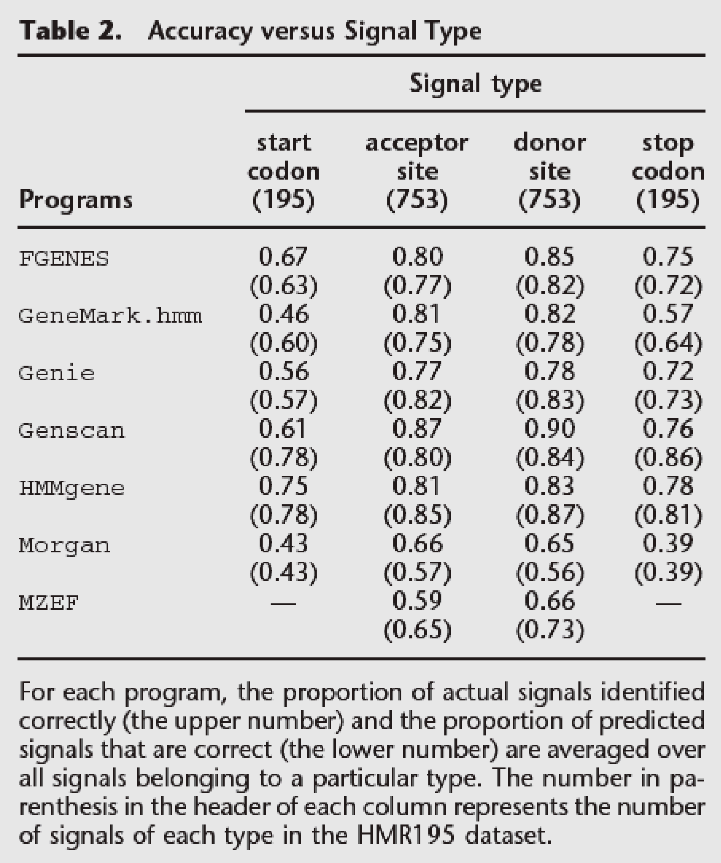
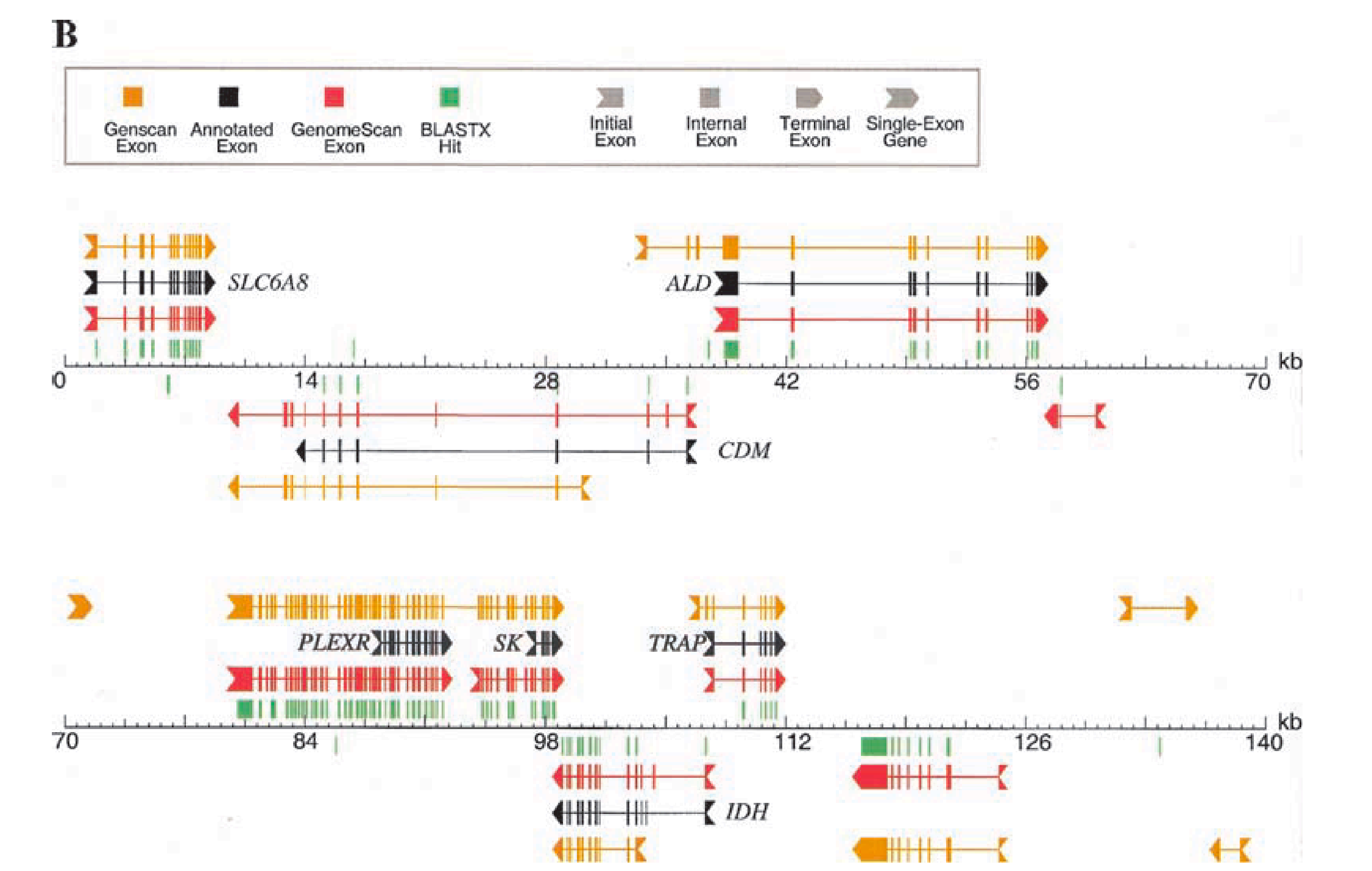


Table 2. Accuracy versus Signal Type

	Signal type			
Programs	start codon (195)	acceptor site (753)	donor site (753)	stop codon (195)
FGENES	0.67 (0.63)	0.80 (0.77)	0.85 (0.82)	0.75 (0.72)
GeneMark.hmm	0.46 (0.60)	0.81 (0.75)	0.82 (0.78)	0.57 (0.64)
Genie	0.56 (0.57)	0.77 (0.82)	0.78 (0.83)	0.72 (0.73)
Genscan	0.61 (0.78)	0.87 (0.80)	0.90 (0.84)	0.76 (0.86)
HMMgene	0.75 (0.78)	0.81 (0.85)	0.83 (0.87)	0.78 (0.81)
Morgan	0.43 (0.43)	0.66 (0.57)	0.65 (0.56)	0.39 (0.39)
MZEF	—	0.59 (0.65)	0.66 (0.73)	—

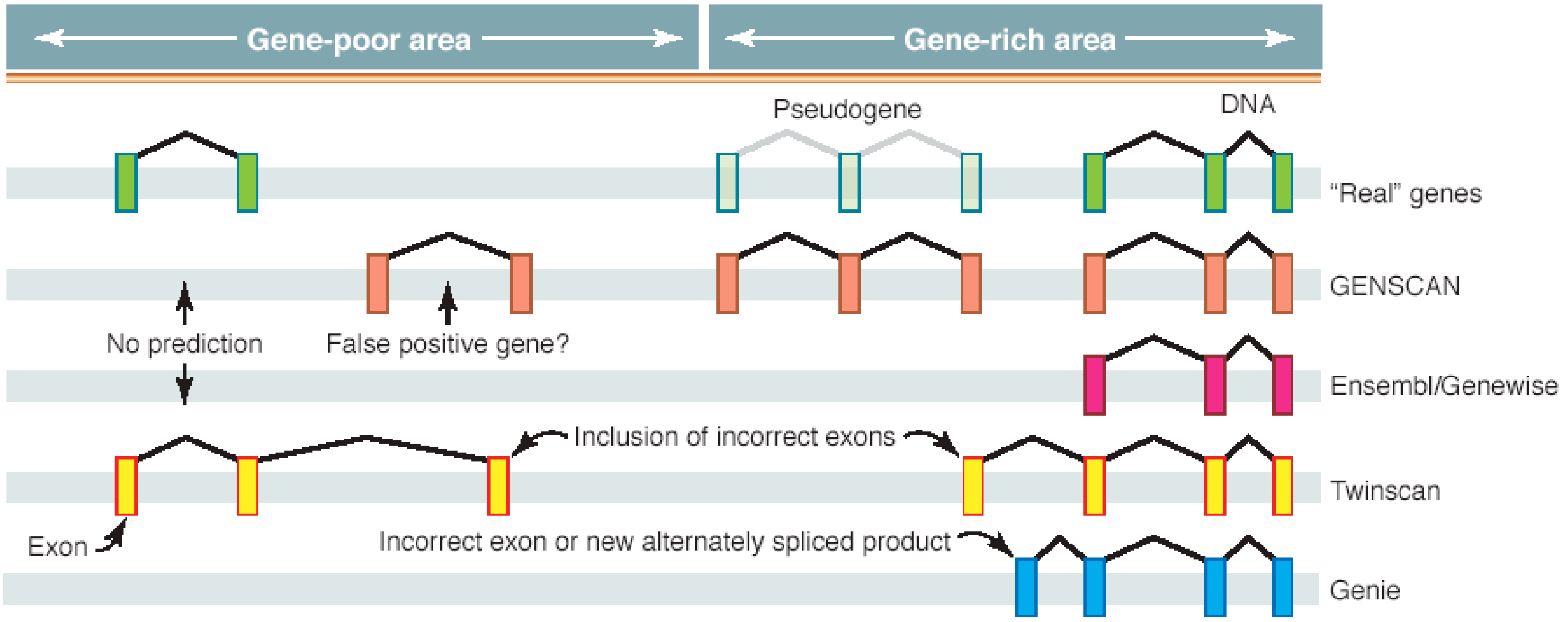
For each program, the proportion of actual signals identified correctly (the upper number) and the proportion of predicted signals that are correct (the lower number) are averaged over all signals belonging to a particular type. The number in parenthesis in the header of each column represents the number of signals of each type in the HMR195 dataset.

GenomeScan



Yeh, Lim, and Burge, Genome Research 11:803-816, 2001.

E. Pennisi, Science 301:1040, 2003



Never perfect. No program calls all genes correctly. Some see genes (shown here as coding regions, or exons, connected by bent lines) where there are none; some miss a gene altogether; and some don't put all the gene's parts in the right places.