基因预测

CAAS08F1A：张荣志 郑永胜 郝峰
李玉荣 张程程 李珊珊
杨菲 郑作良 刘峙
演　讲　人：杨菲

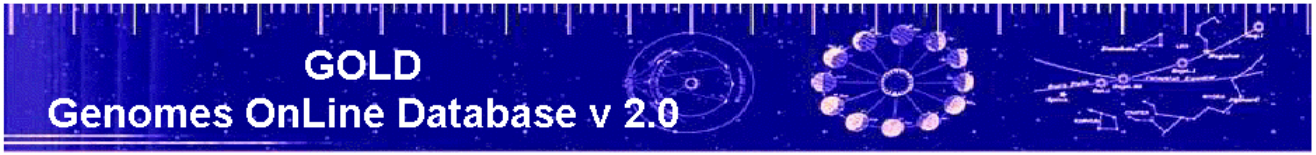
1.基因预测意义

2.基因预测原理

3.基因预测常用软件

4.存在的主要问题

1. 基因预测意义



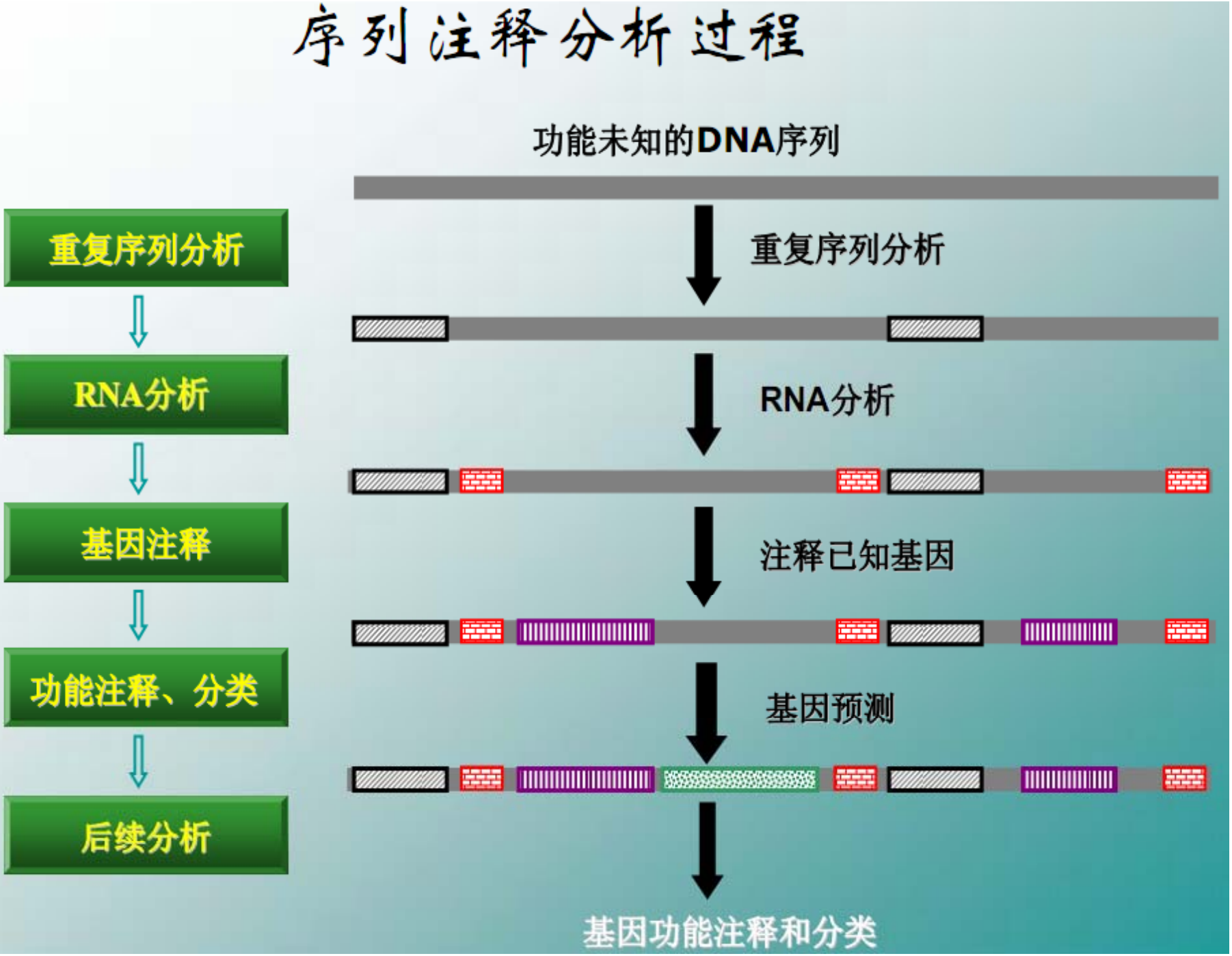
Contact: Genomesonline	Last Update: March 1, 2009	Location www.genomesonline.org
958 Published Complete Genomes	Search GOLD: 4613 genome projects	137 Metagenomes
100 Archaeal Ongoing Genomes	2405 Bacterial Ongoing Genomes	1013 Eukaryotic Ongoing Genomes

	Right-click to save all data: DOWNLOAD	

据GOLD（Genomes OnLine Database）网站统计，截止到2009年3月1日，已经完成测序的基因组有958种，正在进行测序的多达3655种。

大量生物基因组计划的完成提供了极其丰富的生物序列资源，如何进行序列注释是测序后所面临的首要问题。从目前的研究来看，基因组序列由3种成分构成：基因序列、重复序列、基因间区序列。基因序列在高等生物基因组中所占的比例可能并不大，但却是控制生物性状遗传的主要因素，正确鉴定它们对分子遗传学研究至关重要。

序列注释分析过程


重复序列分析
RNA分析
基因注释
功能注释、分类
后续分析
功能未知的DNA序列
重复序列分析
RNA分析
注释已知基因
基因预测
基因功能注释和分类

2. 基因预测原理

- 原核基因结构
- 真核基因结构
- 马尔可夫模型与隐马尔可夫模型
- 基因预测算法的分类
- 原核生物中的基因预测
- 真核生物中的基因预测

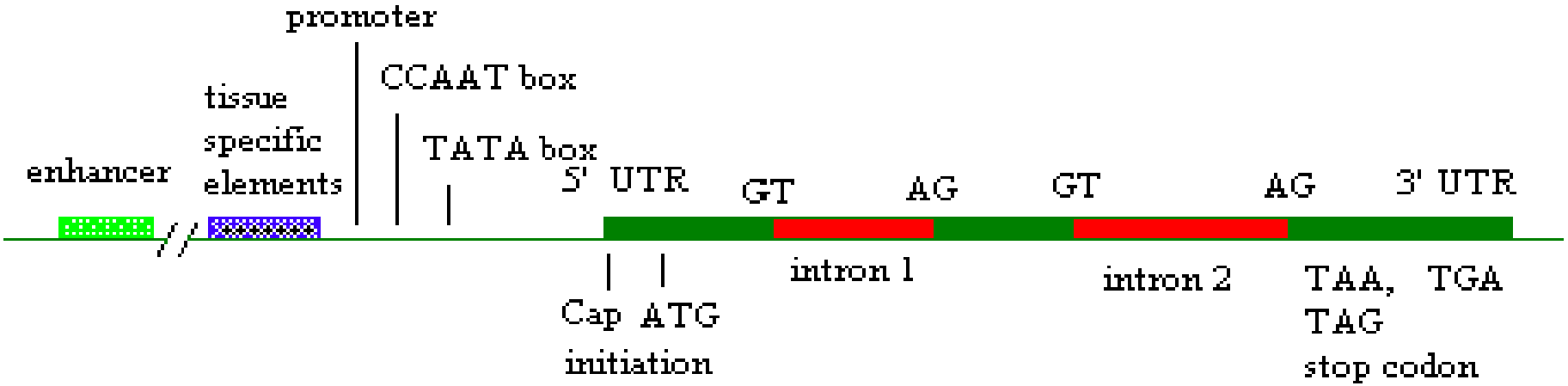
2.1 原核基因结构

- 原核生物基因组小，基因密度高，很少存在重复序列，一个基因是由编码一个蛋白质或RNA的开封阅读框构成，中间没有间断。
- 细菌的起始密码子为: ATG, GTG, TTG
- 核糖体结合位点(*Shine-Delgaron sequence*)
- 终止密码子较容易确定
- 转录终止子
- 密码子偏好性

2.2 真核基因结构

- 基因组较大，基因密度低，富含重复序列和转座元件；最重要的是基因被插入的非编码序列（内含子）切分成小段（外显子）。
- 初生的转录产物需要经过三个步骤转变成成熟的可翻译为蛋白的mRNA。
- 真核基因预测的主要问题是识别外显子、内含子和间接位点。
- 真核基因中存在一些保守序列特征有助于进行计算预测，如：

GT-AG规则，密码子偏好性，六聚体频率，*kozak*序列，CpG岛，poly-A



2.3 马尔可夫模型与隐马尔可夫模型

- 马尔可夫模型是描述一条DNA序列中核苷酸分布的模型。
- 用马尔可夫模型进行基因预测利用以下事实：编码区寡核苷酸分布概率与非编码区不同。
- 统计分析表明密码子对具有相关性。一组六聚体核苷酸在编码区出现的概率要比随机分布概率高，因此，用计算六聚体碱基概率的五阶马尔可夫模型来检测编码区中核苷酸的相关性准确度更高，也较为常用。
- 在基因内容和长度分布上，非典型的基因和典型基因是不同的，预测典型基因的模型可能会漏掉对非典型基因的预测。为了使一个算法适用于整个基因组中的全部基因，就需要更多的马尔可夫模型。结合代表典型与非典型核苷酸分布的不同的马尔可夫模型，建立了隐马尔可夫模型预测算法。

2.4 基因预测程序分类

- 基于从头算的方法（Ab initio-based）：以给定的序列本身来进行预测，主要依赖于以下两个特征：

  基因信号（Gene signals）：包括起始终止密码子，内含子剪接信号，转录因子结合位点，核糖体结合位点以及Poly-A等。

  基因内容（Gene content）：对编码区的统计学上的描述。可以由概率模型：马尔可夫模型或隐马尔可夫模型检测到这一特性，用以区别编码与非编码区。

- 基于同源性的方法（Homology-based）以检索序列与已知基因的序列最大的匹配为基础。
- 基于一致性的算法（Consensus based）以上两种策略相结合。

2.5 原核生物中的基因预测

- 由于原核生物基因组密度较高且没有插入基因，其预测较真核生物简单。目前，基于HMMs的原核生物基因预测算法已经达到相当高的准确度。
- 主要对真核生物中的基因预测做详细介绍。

2.6 真核生物中的基因预测

- 统分为三大类：

以从头算的方法为基础

以同源性的方法为基础

以一致性为基础

- 大部分程序是物种专一的，这是由于用于获得统计参数的训练数据必须由单一生物体取得。

2.6.1 基于从头算（Ab initio-based）的程序

- 此程序的目标是从非编码序列中辨别外显子，随后使外显子以正确的次序排列。
- 主要困难之处在于正确识别外显子。要预测外显子，算法依赖于两个方面特征，基因信号和基因内容。

其中，六聚体频率对识别可能的编码区来说是最具有鉴别力的。

- 以下分别介绍用判别分析和隐马尔可夫模型进行预测的原理。

2.6.1.1 用判别分析进行预测

- 一些基因预测程序依赖于判别分析，线性判别分析（LDA）或二次判别分析（QDA），来提高准确性。
- LDA或QDA用编码信号对应的所有可能的3'端剪接位点作二维图，并用斜线或曲线来划分编码与非编码信号，这是以已知基因结构的培训数据集的知识为基础的。
- E.g: FGENES, MZEF.

2.6.1.2 用隐马尔可夫模型进行基因预测

- GENESCAN是以网络为基础的基于五阶马尔可夫模型进行基因预测的程序。
- 它结合六聚体频率以及编码信号（起始密码子，TATA box，帽子位点，poly-A等）进行预测。
- 假定的外显子能够成为真外显子的概率得分为P，只有当P大于0.5时，才被认为是可靠的。
- 此程序训练用于脊椎动物、拟南芥（双子叶植物）和玉米（单子叶植物）。也可以用来预测人类基因。

2.6.2 基于同源性（Homology-based）的程序

- 以同源性为基础的程序是以相关物种外显子的结构及序列的高度保守性为基础的。
- 当一条检索序列中编码蛋白质的序列翻译后并与数据库中最为相近的蛋白质序列比对后，如有几乎完全配对的区域，即可显示出检测序列的外显子界限。
- 这种方法假定数据库中的序列都是正确的。它按照以下事实进行合理假设：用于比较的同源序列均来自于同一物种的cDNA或表达序列标签。由于有实验证据的支持，这对于在未知基因组的DNA中寻找基因来说是一种十分有效的方法。
- 这种方法的缺点是，数据库中必须存在同源序列。在数据库中没有匹配的情况下新物种中的新基因则不能被预测。
- E.g: GenomeScan, EST2Genome, SGP-1, TwinScan.

2.6.3 基于一致性（Consensus-based）的程序

- 由于不同的预测程序的灵敏度和特异性的差异，以综合手段为基础的方法将多个程序的结果综合起来进行分析是十分有必要的。
- 该方法将与大多数程序相一致的预测结果保留下来，其余的结果被删除掉。
- 这种方法可以提高特异性，但会遗漏一些有用的新预测（因为新预测可能不被大多数程序认可而被忽略掉了）。
- E.g: GeneComber, DIGIT.

3. 基因预测常用软件

- 适用于原核生物

GeneMark, Glimmer, FGENESB

- 适用于真核生物

GENSCAN, FGENESH, TwinScan

New GENSCAN Web Server ...

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

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.

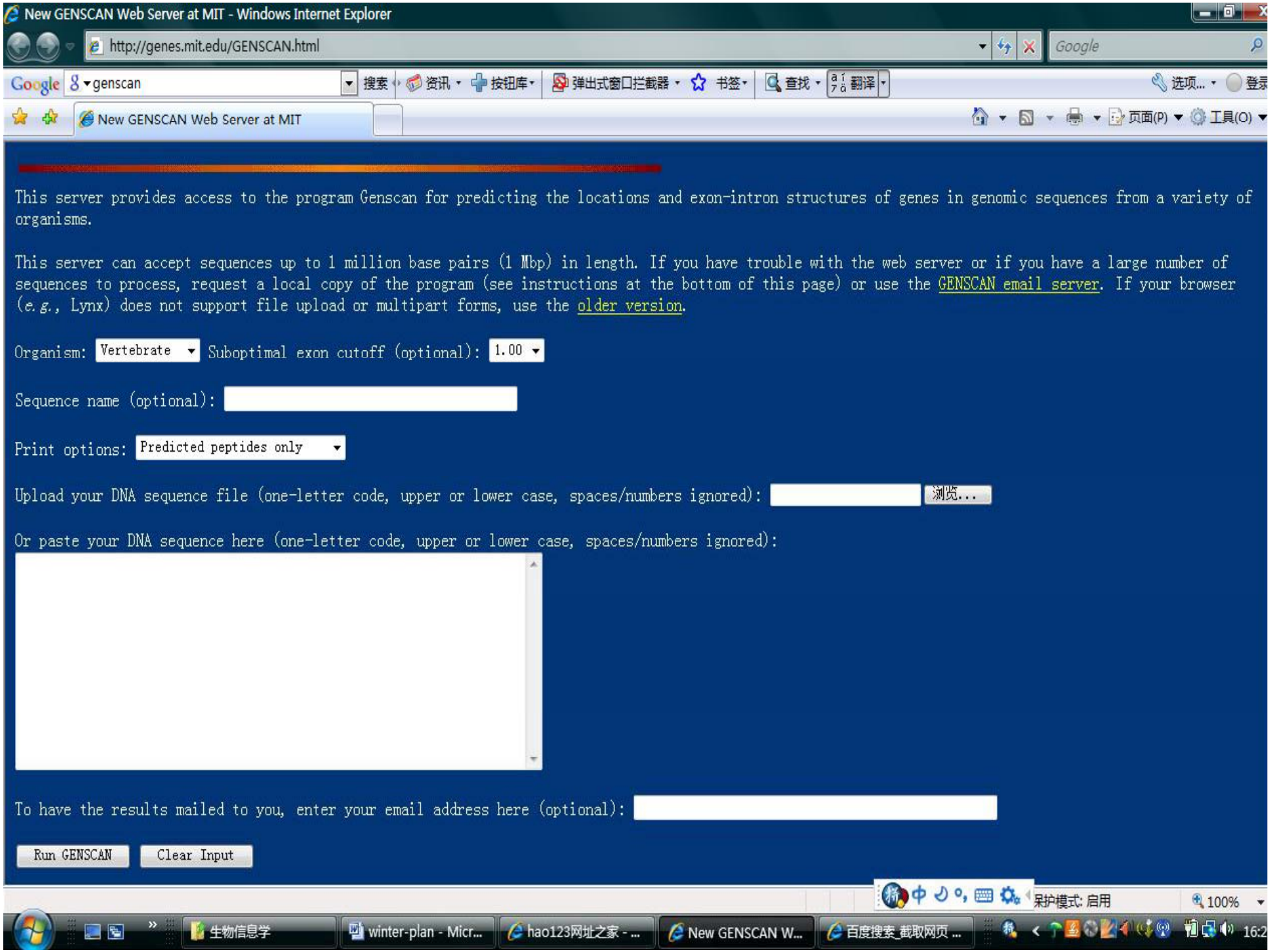
This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the GENSCAN email server. If your browser (*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


New GENSCAN Web Server at MIT - Windows Internet Explorer
http://genes.mit.edu/GENSCAN.html
New GENSCAN Web Server at MIT
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.
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
Sequence name (optional):
Print options: Predicted peptides only
Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):
浏览...
Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):
To have the results mailed to you, enter your email address here (optional):
Run GENSCAN
Clear Input

GENSCANW output for sequence 07:55:59

GENSCAN 1.0 Date run: 7-Mar-109 Time: 07:56:11

Sequence gi : 73308 bp : 39.46% C+G : Isochore 1 (0 - 43 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	3434	3638	205	1	1	67	11	121	0.416	1.36
1.02	Intr	+	5012	5152	141	0	0	14	33	172	0.591	3.70
1.03	Intr	+	6238	6416	179	2	2	68	98	154	0.855	13.12
1.04	Term	+	13456	13566	111	0	0	79	38	42	0.124	-4.12
1.05	PlyA	+	13583	13588	6							1.05
2.00	Prom	+	14142	14181	40							-6.35
2.01	Init	+	19541	19632	92	1	2	100	115	100	0.990	13.91
2.02	Intr	+	19755	19977	223	0	1	56	113	170	0.818	13.51
2.03	Term	+	20833	20961	129	0	0	96	44	116	0.867	5.20
2.04	PlyA	+	21055	21060	6							1.05
3.00	Prom	+	34441	34480	40							-6.05
3.01	Init	+	34531	34622	92	0	2	96	102	131	0.999	15.31
3.02	Intr	+	34745	34967	223	2	1	105	113	224	0.996	23.81
3.03	Term	+	35854	35982	129	0	0	83	49	153	0.985	8.10
3.04	PlyA	+	36043	36048	6							1.05
4.00	Prom	+	39377	39416	40							-6.05

Explanation

Gn.Ex : gene number, exon number (for reference)
Type : Init = Initial exon (ATG to 5' splice site)
Intr = Internal exon (3' splice site to 5' splice site)
Term = Terminal exon (3' splice site to stop codon)
Sngl = Single-exon gene (ATG to stop)
Prom = Promoter (TATA box / initation site)
PlyA = poly-A signal (consensus: AATAAA)
S : DNA strand (+ = input strand; - = opposite strand)
Begin : beginning of exon or signal (numbered on input strand)
End : end point of exon or signal (numbered on input strand)
Len : length of exon or signal (bp)
Fr : reading frame (a forward strand codon ending at x has frame x mod 3)
Ph : net phase of exon (exon length modulo 3)
I/Ac : initiation signal or 3' splice site score (tenth bit units)
Do/T : 5' splice site or termination signal score (tenth bit units)
CodRg : coding region score (tenth bit units)
P : probability of exon (sum over all parses containing exon)
Tscr : exon score (depends on length, I/Ac, Do/T and CodRg scores)

Comments

The SCORE of a predicted feature (e.g., exon or splice site) is a log-odds measure of the quality of the feature based on local sequence properties. For example, a predicted 5' splice site with score > 100 is strong; 50-100 is moderate; 0-50 is weak; and below 0 is poor (more than likely not a real donor site).

The PROBABILITY of a predicted exon is the estimated probability under GENSCAN's model of genomic sequence structure that the exon is correct. This probability depends in general on global as well as local sequence properties, e.g., it depends on how well the exon fits with neighboring exons. It has been shown that predicted exons with higher probabilities are more likely to be correct than those with lower probabilities.

Click here to view a PDF image of the predicted gene(s)

Click here for a PostScript image of the predicted gene(s)

Predicted peptide sequence(s):

>gi|GENSCAN_predicted_peptide_1|211_aa
MTEGSLIILGCMNQIINRNVETNYGEAGYERALNEMETAMLPAHSGLQHVEIWGFVKTGL
KSEAPLDKSKFQKIASTTVRDILPVSEPDVNLAMGWDPPPVEPSNQLLQSKSNAAAKHKL
RYCGCEKLEVDIPALWPLLLTFTSWRLEVVVQATVADHTSSTIIAFLQESLREKKLSIVR
ALIYIHTQFLTLPSYHHPKSKEKNNGFAHLC

>gi|GENSCAN_predicted_peptide_2|147_aa
MVHFTAEEKAAVTSLWSKMNVEEAGGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPK
VKAHGKKVLTSFGDAIKNMDNLKPAFAKLSELHCDKLHVDPENFKLLGNVMVIILATHFG
KEFTPEVQAAWQKLVSAVAIALAHKYH

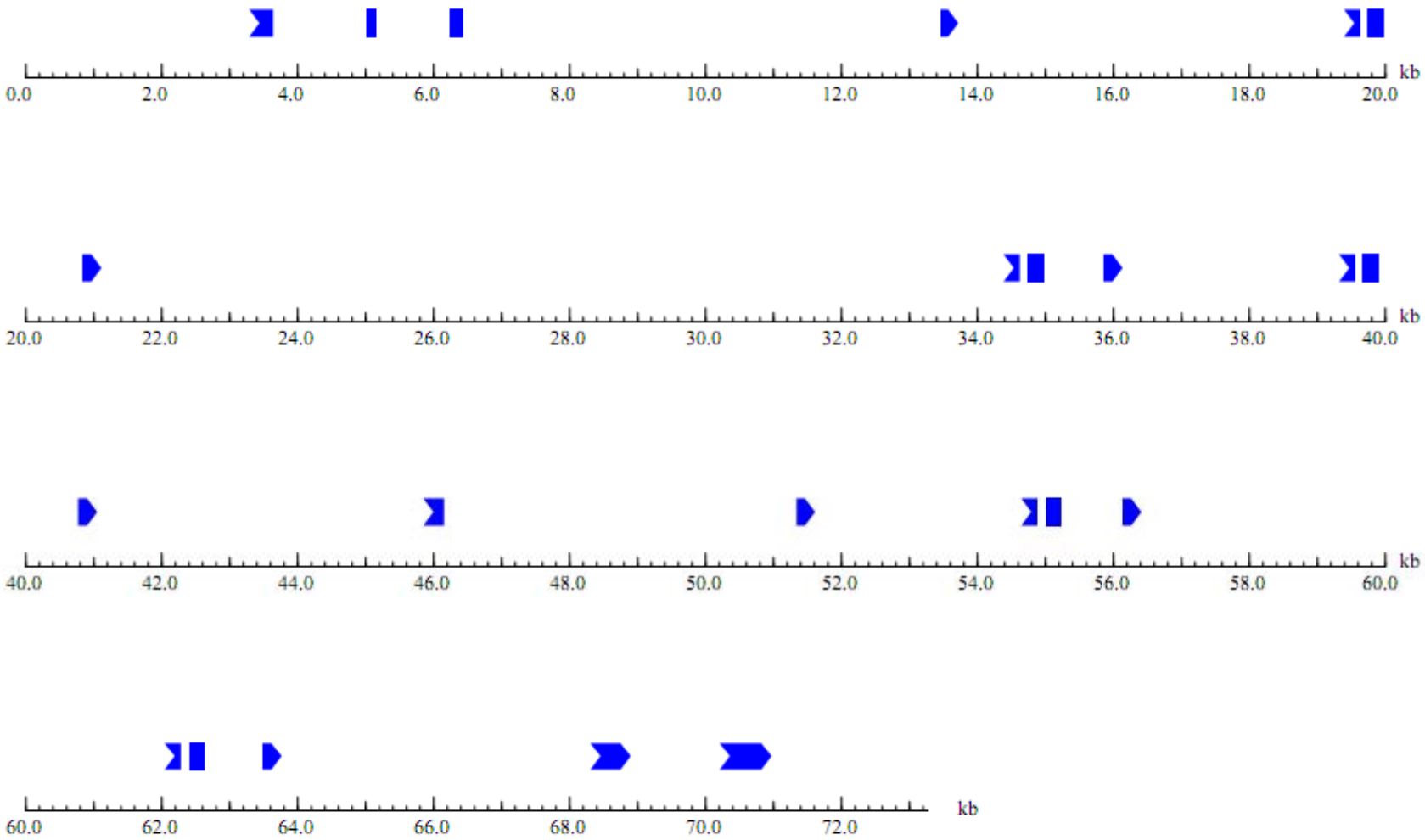
>gi|GENSCAN_predicted_peptide_3|147_aa
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPK
VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFG
KEFTPEVQASWQKMVTGVASALSSRYH

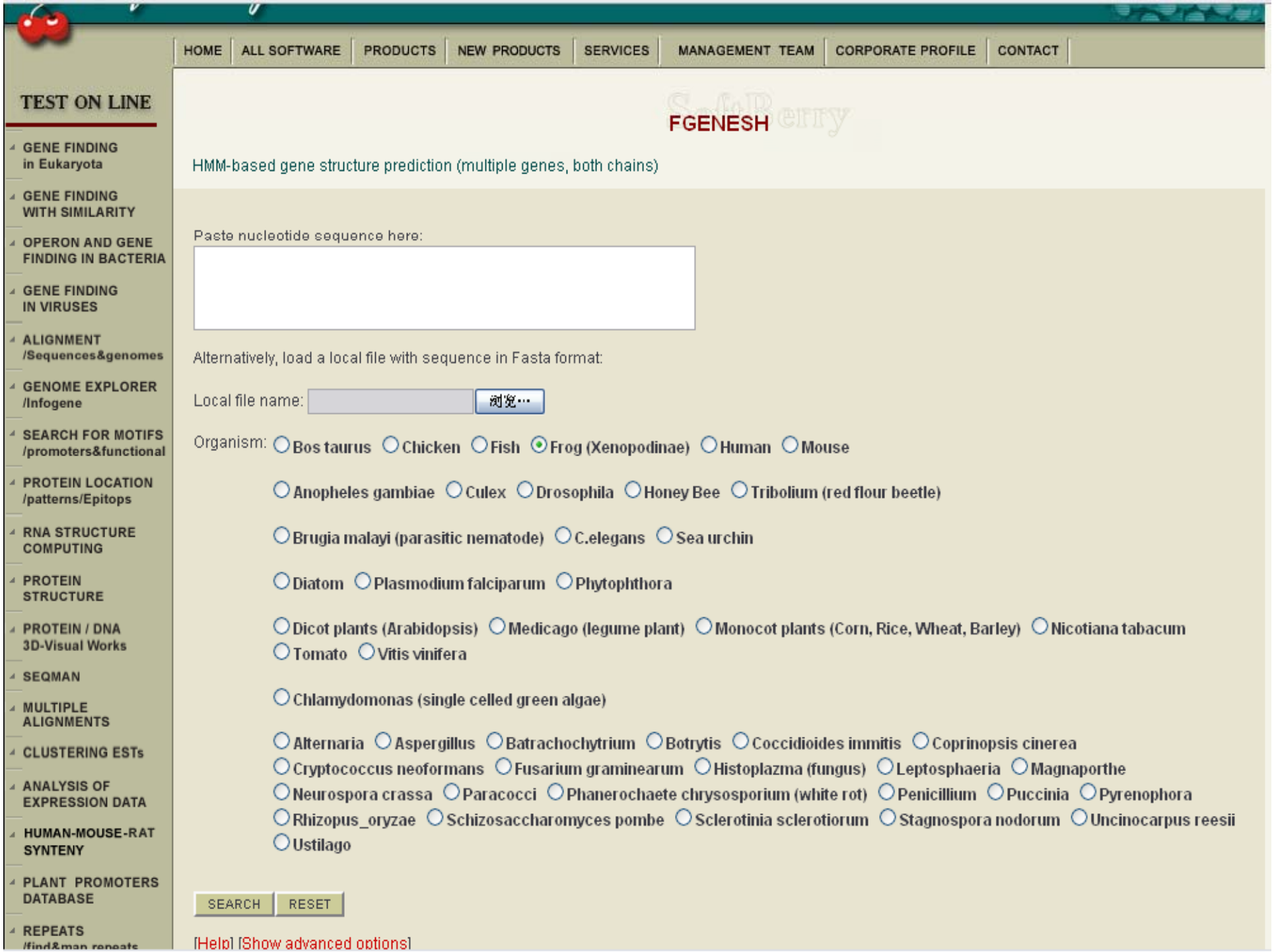
>gi|GENSCAN_predicted_peptide_4|147_aa
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPK
VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFG
KEFTPEVQASWQKMVTAVASALSSRYH

>gi|GENSCAN_predicted_peptide_5|92_aa
MGNPKVKAHGKKVLISFGKAVMLTDDLKGTFATLSDLHCNKLHVDPENFLVSTFHCVSKS
RTLSPTCSETNSKLTSILTHIGITLIKIDTEF

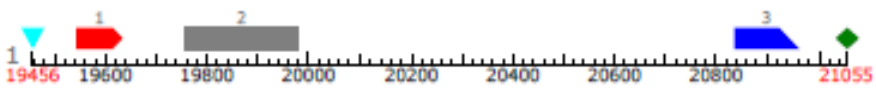
>gi|GENSCAN_predicted_peptide_6|147_aa
MVHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPENFRLLGNVLVCVLARNFG
KEFTPQMQAAYQKVVAGVANALAHKYH

GENSCAN predicted genes in sequence gi

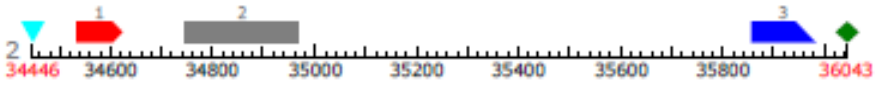



HOME
ALL SOFTWARE
PRODUCTS
NEW PRODUCTS
SERVICES
MANAGEMENT TEAM
CORPORATE PROFILE
CONTACT
TEST ON LINE
GENE FINDING in Eukaryota
GENE FINDING WITH SIMILARITY
OPERON AND GENE FINDING IN BACTERIA
GENE FINDING IN VIRUSES
ALIGNMENT /Sequences&genomes
GENOME EXPLORER /Infogene
SEARCH FOR MOTIFS /promoters&functional
PROTEIN LOCATION /patterns/Epitops
RNA STRUCTURE COMPUTING
PROTEIN STRUCTURE
PROTEIN / DNA 3D-Visual Works
SEQMAN
MULTIPLE ALIGNMENTS
CLUSTERING ESTs
ANALYSIS OF EXPRESSION DATA
HUMAN-MOUSE-RAT SYNTENY
PLANT PROMOTERS DATABASE
REPEATS
FGENESH
HMM-based gene structure prediction (multiple genes, both chains)
Paste nucleotide sequence here:
Alternatively, load a local file with sequence in Fasta format:
Local file name:
浏览…
Organism:
Bos taurus
Chicken
Fish
Frog (Xenopodinae)
Human
Mouse
Anopheles gambiae
Culex
Drosophila
Honey Bee
Tribolium (red flour beetle)
Brugia malayi (parasitic nematode)
C.elegans
Sea urchin
Diatom
Plasmodium falciparum
Phytophthora
Dicot plants (Arabidopsis)
Medicago (legume plant)
Monocot plants (Corn, Rice, Wheat, Barley)
Nicotiana tabacum
Tomato
Vitis vinifera
Chlamydomonas (single celled green algae)
Alternaria
Aspergillus
Batrachochytrium
Botrytis
Coccidioides immitis
Coprinopsis cinerea
Cryptococcus neoformans
Fusarium graminearum
Histoplazma (fungus)
Leptosphaeria
Magnaporthe
Neurospora crassa
Paracocci
Phanerochaete chrysosporium (white rot)
Penicillium
Puccinia
Pyrenophora
Rhizopus_oryzae
Schizosaccharomyces pombe
Sclerotinia sclerotiorum
Stagnospora nodorum
Uncinocarpus reesii
Ustilago
SEARCH
RESET
[Help] [Show advanced options]

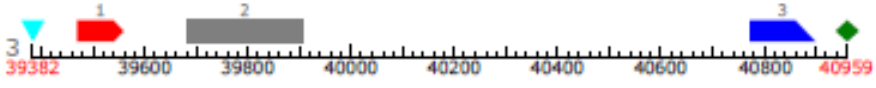
FGENESH 2.6 Prediction of potential genes in Homo_sapiens genomic DNA
Seq name: gi|13907843|ref|NG_000007.1| Homo sapiens genomic beta globin region (HBB@) on
Length of sequence: 73308
Number of predicted genes 10: in +chain 10, in -chain 0.
Number of predicted exons 21: in +chain 21, in -chain 0.
Positions of predicted genes and exons: Variant 1 from 1, Score:180.171887



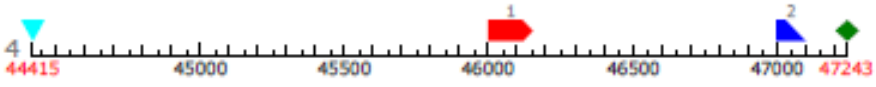
1	+		TSS	19456		-7.09			
1	+	1	CDSf	19541	- 19632	16.13	19541	- 19630	90
1	+	2	CDSl	19755	- 19977	13.37	19756	- 19977	222
1	+	3	CDSl	20833	- 20961	3.34	20833	- 20961	129
1	+		PolA	21055		1.12			



2	+		TSS	34446		-7.09			
2	+	1	CDSf	34531	- 34622	13.42	34531	- 34620	90
2	+	2	CDSl	34745	- 34967	21.52	34746	- 34967	222
2	+	3	CDSl	35854	- 35982	2.92	35854	- 35982	129
2	+		PolA	36043		1.12			

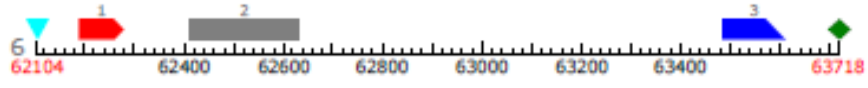


3	+		TSS	39382		-7.09			
3	+	1	CDSf	39467	- 39558	13.42	39467	- 39556	90
3	+	2	CDSl	39681	- 39903	21.52	39682	- 39903	222
3	+	3	CDSl	40770	- 40898	3.66	40770	- 40898	129
3	+		PolA	40959		1.12			

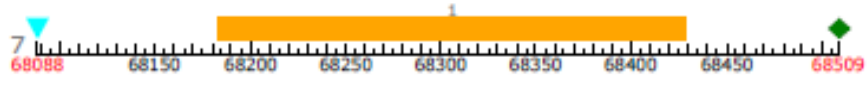


4	+		TSS	44415		-8.69			
4	+	1	CDSf	45995	- 46151	16.58	45995	- 46150	156
4	+	2	CDSl	46997	- 47100	-1.94	46999	- 47100	102
4	+		PolA	47243		1.12			

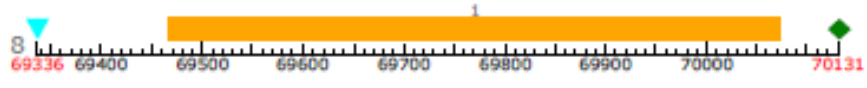
5	+		TSS	54707		-4.39			
5	+	1	CDSf	54790	- 54881	13.44	54790	- 54879	90
5	+	2	CDSl	55010	- 55232	17.01	55011	- 55232	222
5	+	3	CDSl	56425	- 56535	2.53	56425	- 56535	111
5	+		PolA	56931		1.12			



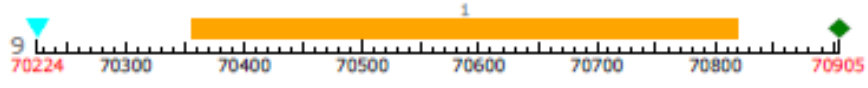
6	+		TSS	62104		-6.59			
6	+	1	CDSf	62187	- 62278	12.99	62187	- 62276	90
6	+	2	CDSl	62409	- 62631	20.06	62410	- 62631	222
6	+	3	CDSl	63482	- 63610	9.54	63482	- 63610	129
6	+		PolA	63718		1.12			



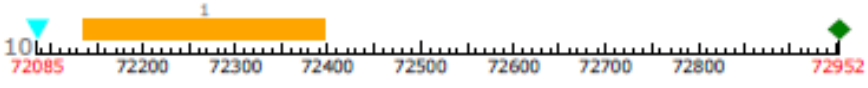
7	+		TSS	68088		-9.39			
7	+	1	CDSo	68183	- 68428	19.52	68183	- 68428	246
7	+		PolA	68509		1.12			



8	+		TSS	69336		-10.29			
8	+	1	CDSo	69467	- 70072	16.45	69467	- 70072	606
8	+		PolA	70131		-1.07			



9	+		TSS	70224		-12.49			
9	+	1	CDSo	70355	- 70819	17.10	70355	- 70819	465
9	+		PolA	70905		1.12			



```
10 +      TSS      72085                -6.39
10 +    1 CDSo     72135 -     72395    7.31      72135 -     72395    261
10 +      PolA     72952                1.12

Predicted protein(s):
>FGENESH:[mRNA] 1 3 exon (s) 19541 - 20961 444 bp, chain +
ATGGTGCATTTTACTGCTGAGGAGAAGGCTGCCGTCACTAGCCTGTGGAGCAAGATGAAT
GTGGAAGAGGCTGGAGGTGAAGCCTTGGGCAGACTCCTCGTTGTTTACCCCTGGACCCAG
AGATTTTTTGACAGCTTTGGAAACCTGTCGTCTCCCTCTGCCATCCTGGGCAACCCCAAG
GTCAAGGCCCATGGCAAGAAGGTGCTGACTTCCTTTGGAGATGCTATTAAAAACATGGAC
AACCTCAAGCCCGCCTTTGCTAAGCTGAGTGAGCTGCACTGTGACAAGCTGCATGTGGAT
CCTGAGAACTTCAAGCTCCTGGGTAACGTGATGGTGATTATTCTGGCTACTCACTTTGGC
AAGGAGTTCACCCCTGAAGTGCAGGCTGCCTGGCAGAAGCTGGTGTCTGCTGTCGCCATT
GCCCTGGCCCATAAGTACCACTGA
>FGENESH: 1 3 exon (s) 19541 - 20961 147 aa, chain +
MVHFTAEEKAAVTSLWSKMNVEEAGGEALGRLLVVYPWTQRFFDSPGNLSSPSAILGNPK
VKAHGKKVLTSFGDAIKNMDNLKPAFAKLSELHCDKLHVDPENFKLLGNVMVIILATHFG
KEFTPEVQAAWQKLVSAVAIALAHKYH
>FGENESH:[mRNA] 2 3 exon (s) 34531 - 35982 444 bp, chain +
ATGGGTCATTTCACAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAAT
GTGGAAGATGCTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAG
AGGTTCTTTGACAGCTTTGGCAACCTGTCCTCTGCCTCTGCCATCATGGGCAACCCCAAA
GTCAAGGCACATGGCAAGAAGGTGCTGACTTCCTTGGGAGATGCCATAAAGCACCTGGAT
GATCTCAAGGGCACCTTTGCCCAGCTGAGTGAACTGCACTGTGACAAGCTGCATGTGGAT
CCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTTGGCAATCCATTTCGGC
AAAGAATTCACCCCTGAGGTGCAGGCTTCCTGGCAGAAGATGGTGACTGGAGTGGCCAGT
GCCCTGTCCTCCAGATACCACTGA
>FGENESH: 2 3 exon (s) 34531 - 35982 147 aa, chain +
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSPGNLSSASAIMGNPK
VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFG
KEFTPEVQASWQKMVTGVASALSSRYH
>FGENESH:[mRNA] 3 3 exon (s) 39467 - 40898 444 bp, chain +
ATGGGTCATTTCACAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAAT
GTGGAAGATGCTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAG
AGGTTCTTTGACAGCTTTGGCAACCTGTCCTCTGCCTCTGCCATCATGGGCAACCCCAAA
GTCAAGGCACATGGCAAGAAGGTGCTGACTTCCTTGGGAGATGCCATAAAGCACCTGGAT
GATCTCAAGGGCACCTTTGCCCAGCTGAGTGAACTGCACTGTGACAAGCTGCATGTGGAT
CCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTTGGCAATCCATTTCGGC
AAAGAATTCACCCCTGAGGTGCAGGCTTCCTGGCAGAAGATGGTGACTGCAGTGGCCAGT
GCCCTGTCCTCCAGATACCACTGA
>FGENESH: 3 3 exon (s) 39467 - 40898 147 aa, chain +
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSPGNLSSASAIMGNPK
VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFG
KEFTPEVQASWQKMVTAVASALSSRYH
>FGENESH:[mRNA] 4 2 exon (s) 45995 - 47100 261 bp, chain +
ATGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGATCTCCTTCGGAAAAGCT
GTTATGCTCACGGATGACCTCAAAGGCACCTTTGCTACACTGAGTGACCTGCACTGTAAC
AAGCTGCACGTGGACCCTGAGAACTTCCTGGTGAGTACTCTTAGGCAACCTGATATTGAT
```

常用基因预测软件网址

基于同源性的预测软件

Genemark http://opal.biology.gatech.edu/GeneMark/

Glimmer http://cbcb.umd.edu/software/glimmer/

FGENES http://linux1.softberry.com/berry.phtml?topic=fgenes&group=programs&subgroup=gfind

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

HMMgene http://www.cbs.dtu.dk/services/HMMgene/

基于比较基因组学的预测软件

Twinscan http://mblab.wustl.edu/nscan/submit/

4. 存在的主要问题

- 假阳性（False Positive, FP）：多预测了假的编码区，即在非编码区预测出编码区。
- 假阴性False Negative, FN）：漏掉了真实的编码区，即将编码区预测为非编码区。
- 过界预测（Over Prediction, OP）：由于基因边界很难准确定位，预测经常会超出实际边界。
- 片段化（Fragmentation）：内含子过大的基因，在预测时容易断裂成两个或多个基因。
- 融合化（Fusion）：距离过近的两个或多个基因，在预测时容易被融合成一个很大的基因。

小结

基因的计算预测是基因组序列分析的各个过程中最为重要的一步。由于原核生物基因组密度较高且没有插入基因，其预测较真核生物简单。目前，基于HMMs的原核生物基因预测算法已经达到相当高的准确度，但是对于真核生物预测还存在着许多问题。对于ab-initio算法，在进行真核生物基因组预测时HMM算法能很好的区分外显子-内含子的界限，其主要的限制是对于统计模型训练的依赖性，训练使此方法变得物种专一。同源性为基础的算法结合HMM可以获得进一步的准确性，这种算法受限于数据库中同源序列的可用性。结合统计和同源信息的综合算法通过准确地检测更多的基因和外显子来得到更好的结果。随着计算技术的进步和对于剪接机制的进一步认识，在不久的将来可信度高的真核基因组预测可以成为现实。

Thank you!

2009.03.08