Course 6
P2. Plant breeding

Iulian Gabur,

Department of Plant Breeding, Justus Liebig University Giessen, Germany

06/07 Noiembrie 2018

Plant Breeding Giessen

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

P2 – Plant breeding

Goal: predict performance of plant breeding lines and make intelligent selection decisions to estimate their “behaviour” under potential future crop production scenarios

Methods: multiple kernel learning framework, deep AI networks similar to word embedding in natural language processing (Mikolov et al.2013; Mejia-Guerra and Buckler, 2017)

Terminologie

Genotip = totalitatea materialului genetic al unui organism.

Fenotip = totalitatea trasaturilor observabile ale unui organism, determinate de materialul genetic si de mediul inconjurator.

Acidului dezoxiribonucleic (ADN)
• ADN este format din nucleotide (A, G, C, T).
• Molecula ADN este formata din 2 lanturi polinucleotidice ale caror nucleotide sunt unite intre ele, 2 cate 2, prin legaturi de hidrogen.

Acidul ribonucleic (ARN) este implicat în decodificarea informației ereditare stocate in ADN.

Gena = secventa de nucleotide care contine informatia necesara sintezei unei proteine sau a unei molecule de ARN.

Codon = o succesiune de 3 nucleotide sau baze azotate (ex. ATG, CGT etc) care codifica un aminoacid (**64 codoni**).

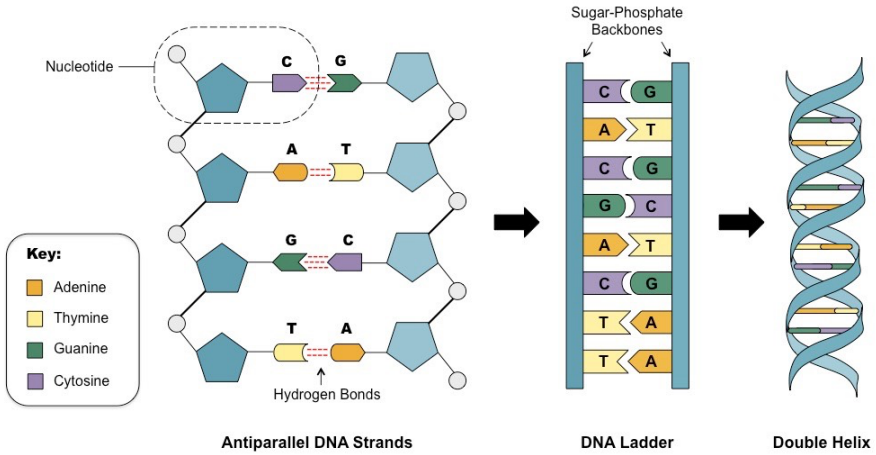
Aminoacid = sunt constituentii fundamentali ai proteinelor (**20 de aminoacizi**).

Structura ADN

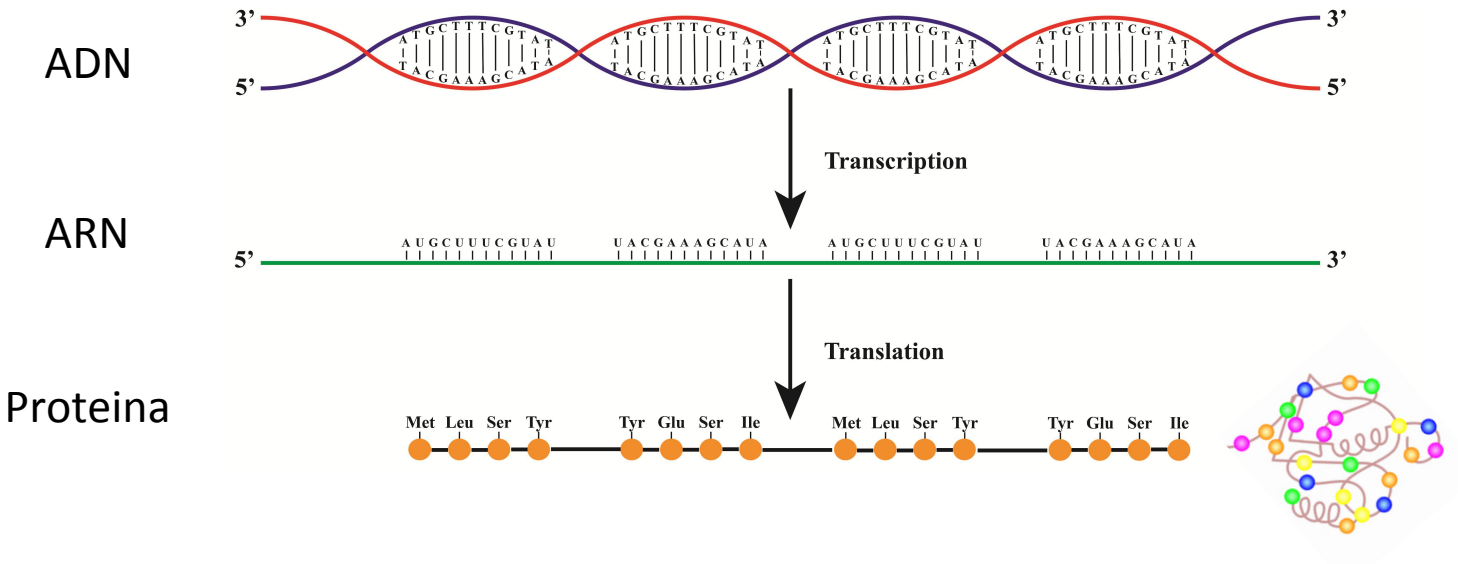
Science 2018, Vol 362, Issue 6413

Dogma centrala a geneticii

a)

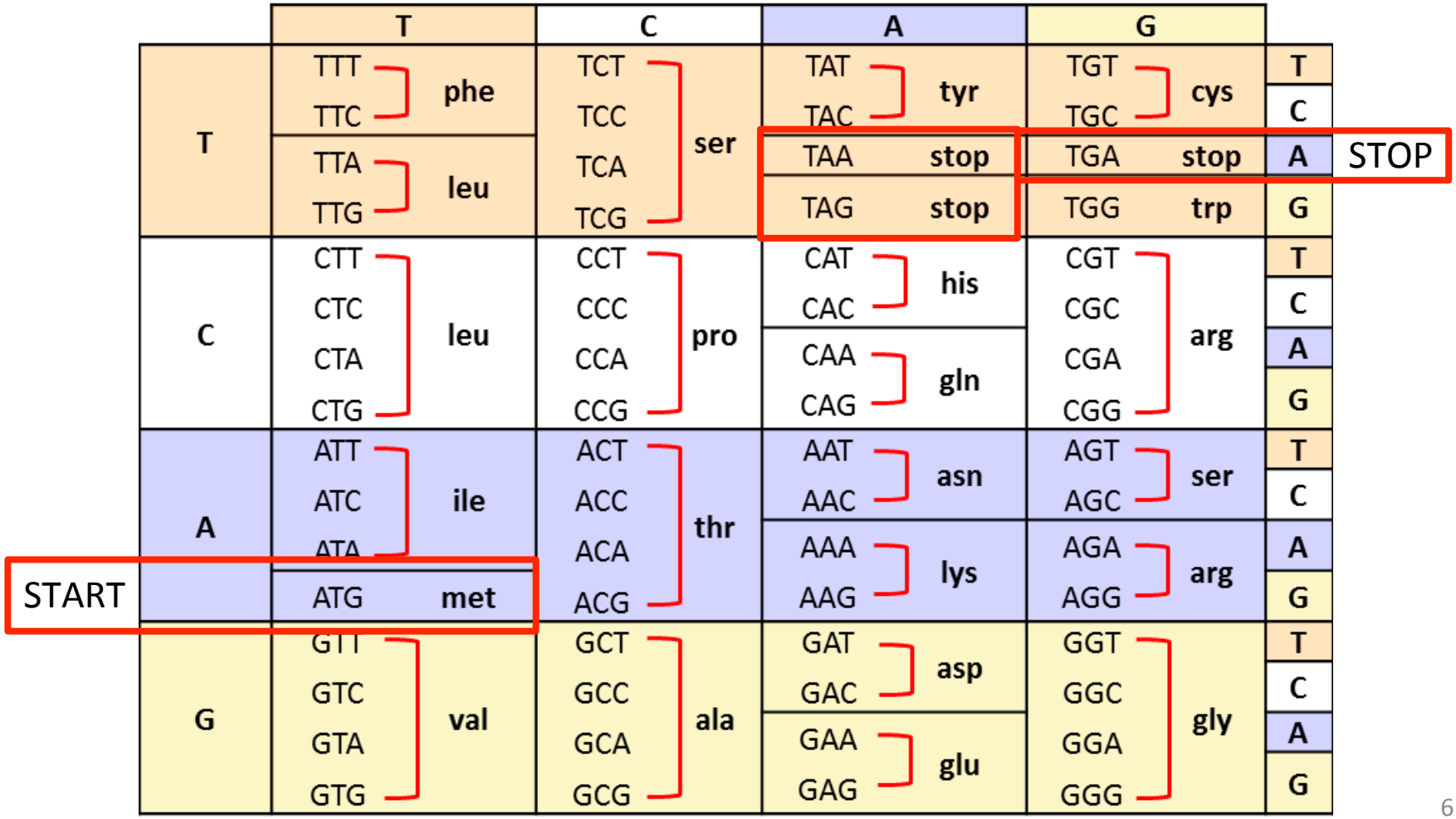


b) Transcriptia si translatia ADN-ului



Codon - Aminoacid

!!!!! 4 nucleotide (ATCG) => 64 codoni (serie de 3 nucleotide) => 20 aminoacizi



Gena

Gena = secventa de nucleotide care contine informatia necesara sintezei unei proteine

Situatie: *este cald afară!*

Gena

Gena = secventa de nucleotide care contine informatia necesara sintezei unei proteine

Situatie: *este cald afară!*

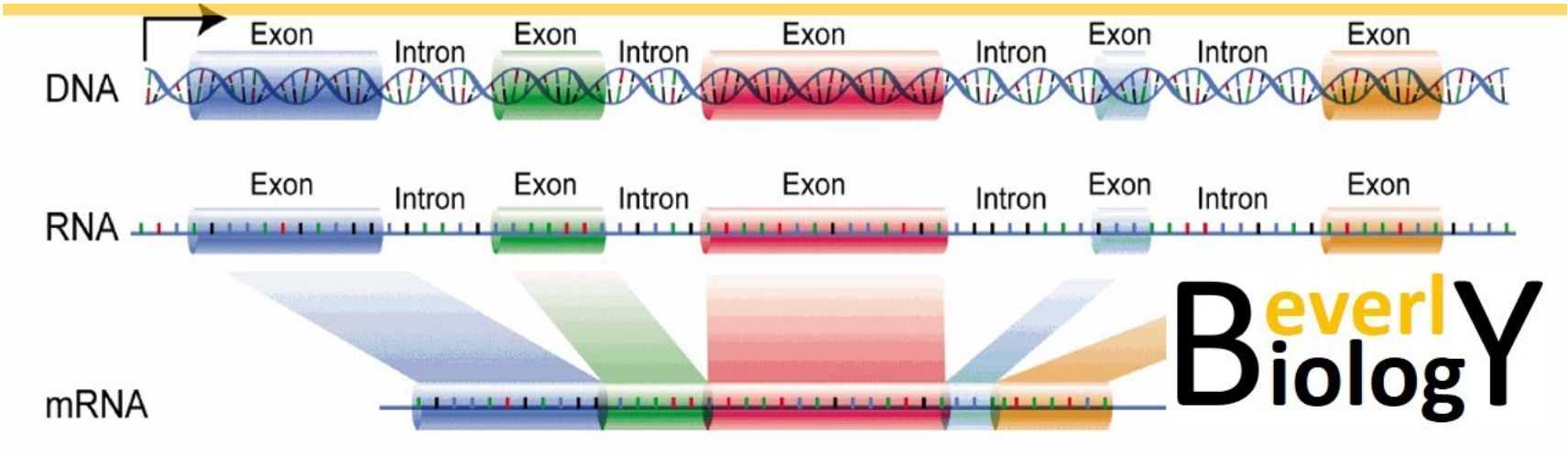
Senzatie: *imi este sete*. Reactie*: ar fi bună o bere rece!* Rezolvare: *găsesc o terasă.*

Receptor 1 → Gena 1 → Proteina 1

Gena

Gena = secventa de nucleotide care contine informatia necesara sintezei unei proteine

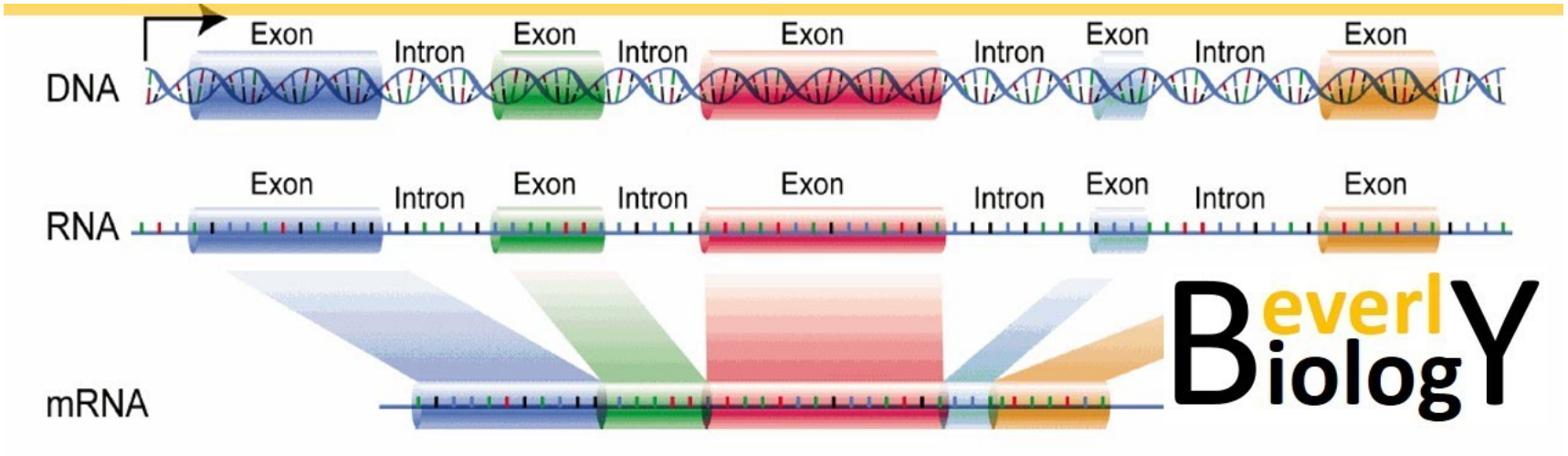
- Codon **START** (**ATG**)
- **Exoni** (Secventa codificatoare) - secvente transcrise, pastrate in ARNm si traduse in proteina
- **Introni** (secvente eliminate, absente dinARNm) – incepe cu **GT** si se termina cu **AG**
- Codon **STOP** (**TAA, TAG, TGA**)



Gena

START STOP

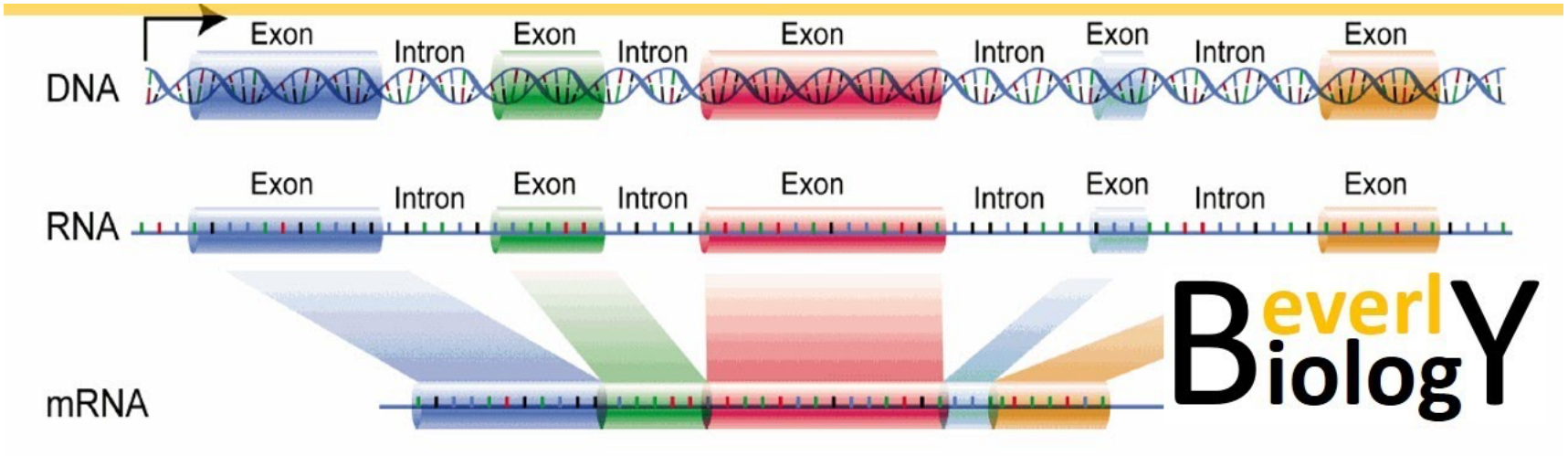
ATG….exon….AAAGT-intron-AG**GTT…exon…AAT**GT-intron-AG**CGT…exon….TAA**



Gena

START STOP

ATG….exon….AAAGT-intron-AG**GTT…exon…AAT**GT-intron-AG**CGT…exon….TAA**

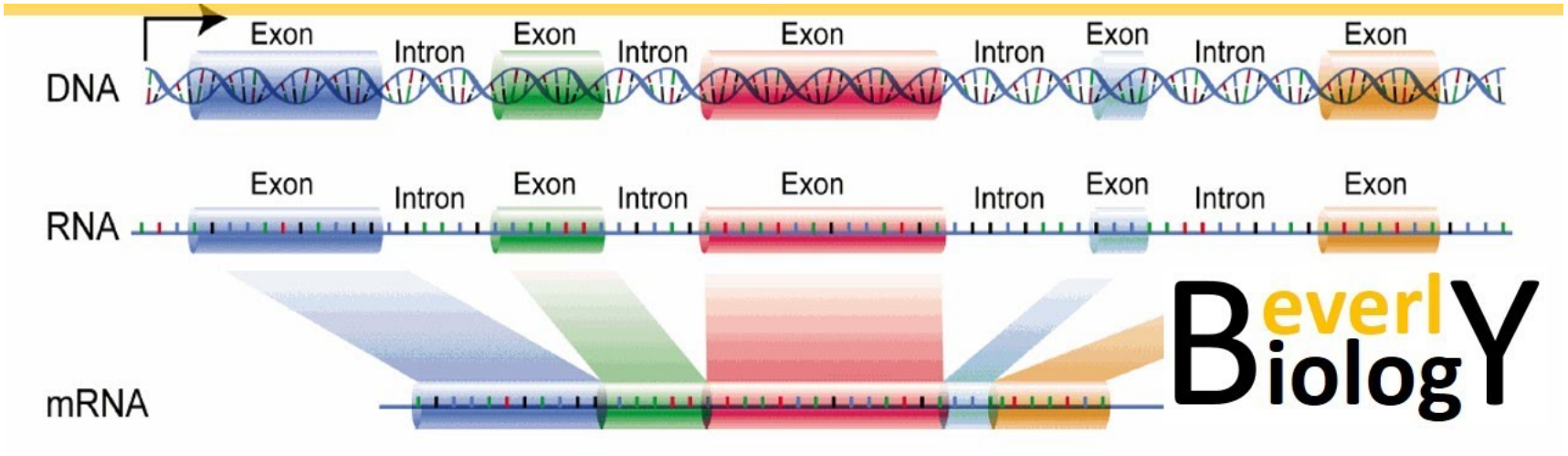


ATG….exon….AAAGTT…exon…AATCGT…exon….TAA

Gena

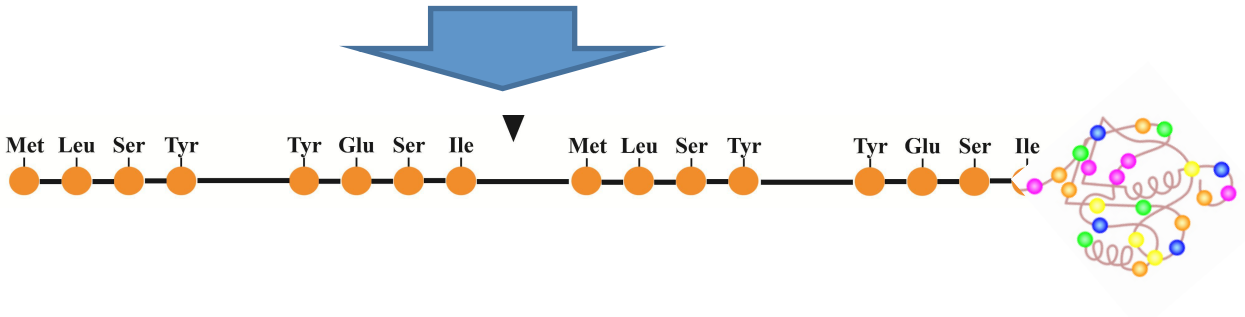
START STOP

ATG….exon….AAAGT-intron-AG**GTT…exon…AAT**GT-intron-AG**CGT…exon….TAA**



ATG….exon….AAAGTT…exon…AATCGT…exon….TAA

Proteina



Gene/Genom

In general, la eucariote 1-2% din genom este alcatuit din gene.
Ex: La *H. sapiens*, 98% din genom contine secvente non-codificatoare.

	Organism	# of protein-coding genes	# of genes naïve estimate: (genome size /1000)
viruses	HIV 1	9	10
	Influenza A virus	10-11	14
	Bacteriophage λ	66	49
	Epstein Barr virus	80	170
prokaryotes	*Buchnera sp.*	610	640
	T. maritima	1,900	1,900
	S. aureus	2,700	2,900
	V. cholerae	3,900	4,000
	B. subtilis	4,400	4,200
	E. coli	4,300	4,600
eukaryotes	*S. cerevisiae*	6,600	12,000
	C. elegans	20,000	100,000
	A. thaliana	27,000	140,000
	D. melanogaster	14,000	140,000
	F. rubripes	19,000	400,000
	Z. mays	33,000	2,300,000
	M. musculus	20,000	2,800,000
	H. sapiens	21,000	3,200,000
	T. aestivum (hexaploid)	95,000	16,800,000

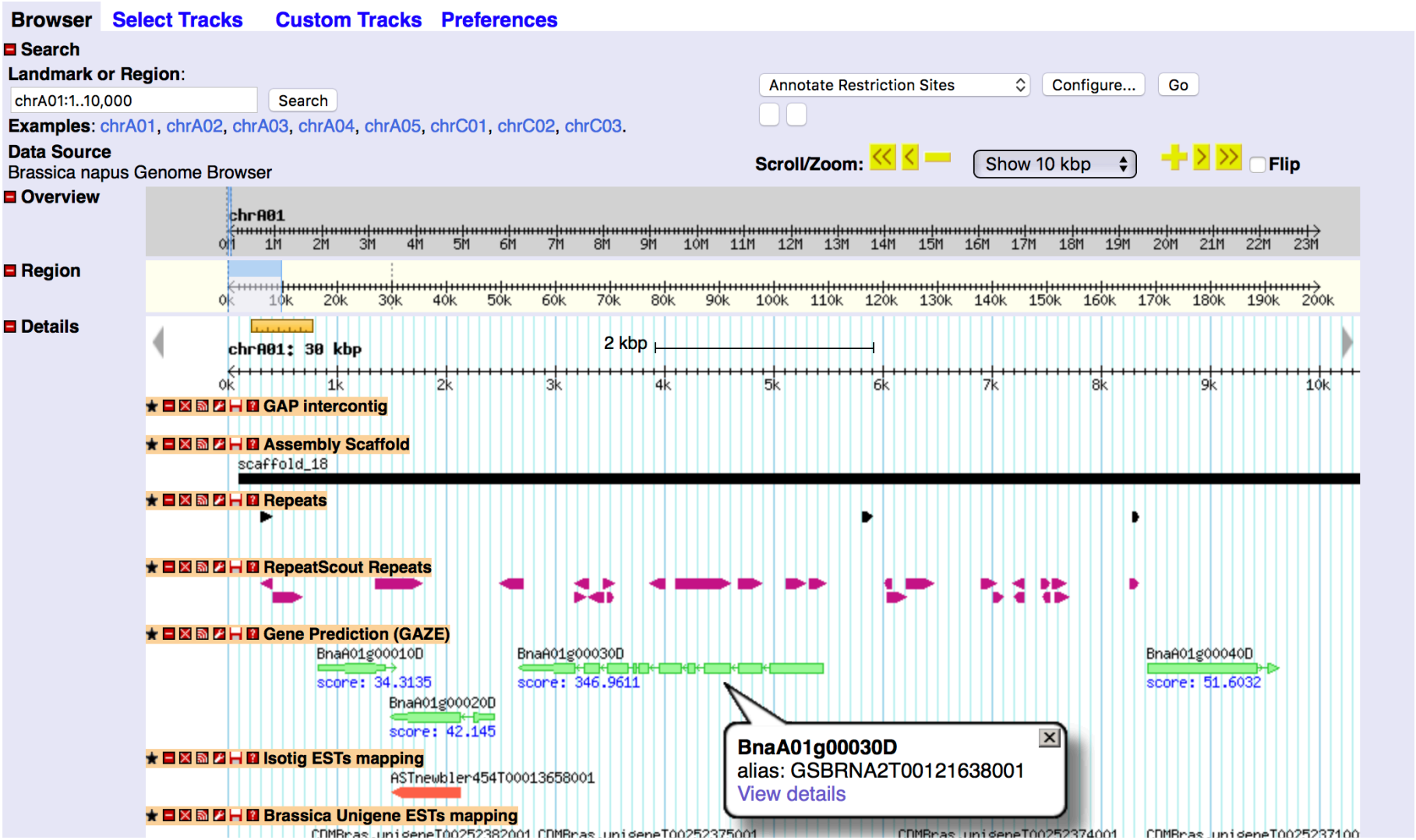
Adapted from M. Lynch, The Origins of Genome Architecture

Rapita (*Brassica napus*)

Rapita (*Brassica napus*)

Rapita (*Brassica napus*)

Cultivar	D41
Nucleotide	848.200.303
Cromozomi	41 (A01-10,C01-09; A01-random,…)
Gene	101.040


Browser
Select Tracks
Custom Tracks
Preferences
Search
Landmark or Region:
chrA01:1..10,000
Search
Examples: chrA01, chrA02, chrA03, chrA04, chrA05, chrC01, chrC02, chrC03.
Data Source
Brassica napus Genome Browser
Annotate Restriction Sites
Configure...
Go
Scroll/Zoom:
Show 10 kbp
Flip
Overview
Region
Details
chrA01
chrA01: 30 kbp
2 kbp
GAP intercontig
Assembly Scaffold
scaffold_18
Repeats
RepeatScout Repeats
Gene Prediction (GAZE)
BnaA01g00010D
score: 34.3135
BnaA01g00020D
score: 42.145
BnaA01g00030D
score: 346.9611
BnaA01g00040D
score: 51.6032
Isotig ESTs mapping
ASTnewbler454T00013658001
Brassica Unigene ESTs mapping
BnaA01g00030D
alias: GSBRNA2T00121638001
View details

Exemplu - exoni in rosu si introni in negru

>BnaA01g00030D

ATGGCGGAGGAAGACGTGTCGCGCAATTTCCGAGCCCTGGTGGAGAACGCGGACCGCAAGTTCGCGAGAGTCCGCG
ATCTCCCGGCCTTCGGGCGAGCGCAGAGCCACTACTTCCACAAGGTCTTCAAGGCCTACATGAAGCTCTGGAATTACCA
GCAGTCCCACCGGCCCAAGCTCGTCGCGTCGGGGCTGAACCGGTGGGAGATCGGCGAGATCGCGAGCCGGATTGGG
CAGCTCTACTTCAGCCAGTACATGAGGACCAGCGAGGCTCGCTTCCTCCTCGAGTCCTACGTCTTCTACGAGGCCATTC
TCAAGCGGAGGTACTTCGAGGAAGGGGAGGAGAGGAAGGATCTCAGCGCGAGGTCCAAGGAGCTGAGGTTCTACGC
GAGGTTCTTGCTCGTTGCGTTGATTGTTGATCGGAGGGAGATGGTTTTGCACTTGGCTGAGAAGCTTAGGGCTTTGGTT
GATGACAGCAAGTCTAATTTTAGGGTAAAAGATTTAATCTTTTTGATTTGAGGCTTTTGAGAATTGAGCCTCTGGTTGAATA
AAGTTTTGATCTTTTGGTGCAGGAGACCAACTTTAAGGAGTGGAGGATAGTTGTGCAAGAGATCACTCGGTTTACTAAAG
CTGACGTGGACTTAACTTATGTCAGGCCGCTTCGTTACTGCGCTATGCTTGATTCCTATCCAGCTTCTCACACGTACCTA
GCGAGGTTCCACGCTAAGAAGCTCTTCAAGTTTCGTGATGCTCTATTAGCTAGCTATCACCGTAACGAGGTAAGCATTAA
ACAAGCAGTGTTTTGTTTTTATTGAACCTAAGATTCATACGCTGTGGCCTTTGTAATCATTTAGGTGAAATACGCTGAAGTG
ACGTTGGATACGTATAGAATGATGCAGTGTTTAGAGTGGGAGCCTAGTGGGTCTTTTTACCAGAAGCGGCCTGTTGAGA
CGAAAGAGAATGGCTTTGTGGTTGATCATACGCTGACTTCTGGGATTATAGATATGAAATTGGCTGCTGATATGGCTGATC
CGTCGTTACCTCCTAATCCTAGGAAGGCAGTTCTTTATCGTCCCACAGTCTCTCACTTGCTTGCGGTGAGTTTTTTTTTTA
TCTGTTTCATATAGGAAAATAATCTTTAGGAGTTTCTGATTTTTTTTTCCTTTATATATGTTCAGGTCTTGGCGATGATCTGT
GATGAGCTCTCTCCAGAGTGTGTGATGCTGATATATCTATCGGCTTCAGGTTTAGTTTTTTTTTTGTTGTCACTTTGGAACT
TCTTGATGTTAACATGGTGACGGTTATCCAGGTGGTCCTGCTGCTCGTGAGAATGTTGCGCAACCGGAGAACGCTGTCG
GTTCAAGCAGAACATCAAAAAGCAAGCTGCTTGGCCGAGCTTCCCAAGAGCAGAAGAGTTACAAGTCAGAACCTCTTAG
CAACGGACACAAATCTTCTGGGGGTTATTATGATGATTATCTTTGGTTAGGTCCTAGAGGAGGCTCTGGTAAGTTGTTTTC
CTCGGATTTACATTGTTATATGTCTCCAGTGGGCAGTGAATGAGCAGACTTTTCTCGCTCACTGTCTTGTTTTTTTTTGTTT
TCAGGTTCAAACAACCTTTACCCTGGTGATCTAATACCTTTTACAAGGAAACCGCTTTTCTTGGTCATTGATAGCGACACT
AGCCGAGCATTCAAGGCAGGTTTGTCCTAACTTAGTCTTATCATTCTCTCAGGAGATTCTTATAGTTGTTCTTTGTCACTG
TGCGATATTGGCTAACGGATTGTGAGCTTCAGGTTTTGAATGGTGCAGAGAGAGGGGAGCCTGTTGCTCTACTTCTTTCT
CCTCTGAAGCCATCTCTGGAGAACCCATCTGCTGATGATACAGCAGCAGCATTAAACGGAAGCCAGTTTACTTTTTTCTT
GACAGCTCCTTTACAAGCATTCTGTCAAATGCTGGGCCTCTCGAACTCGGATCCCGAACCTGTAAGTACCTTAAATGAAG
CAAAGCATATGTTATCATGATGTGGTGGCATTGCTTCTTACCATCTTGATCACATATCTTCAGGAGGTTTATGATGAAGCAG
AGAGCATACTATCTGCTTCTTTCTCTGAGTGGGAAACGATCCTCCTGACTTCCAAAGTGTTGAATCTTGTCTGGGCTCAG
GTCTTGCCTGATCAGTTCTTGAGACGGCTGATTCTCAGTAA

Efectul gene *BnaA01g00030D*

Locatie :

Cromozomul: A01
Pozitie start: 2665
Pozitie stop: 5455

Codeaza: o proteina *scai-like*

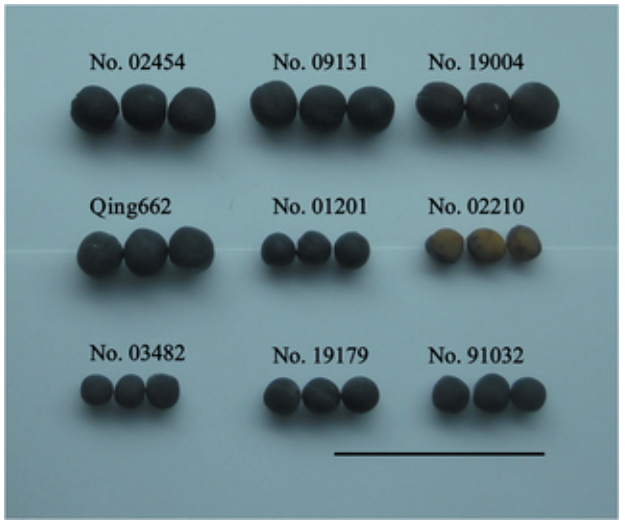
Rol: receptor responsabil cu intreruperea sintezei altei proteine

Variatia genomică

Trasatura monogenica

P1 D41 P3 P4 D81 P6 P7 Tp

Trasatura poligenica



Problema

1. Identificarea genelor dintr-un genom complex pornind de la secventa nucleotidică

2. Prezicerea efectului genelor in alte doua genomuri

Input:

- secventa pentru D41, D81, Tp:

“1.1.D41_genom_secventa_nucleotide.txt”;
“2.D81_genom_secventa_nucleotide.txt”;
“3.Tp_genom_secventa_nucleotide.txt”.

- lista genelor din D41 (cromozom, start, stop, intron, exon)

“1.2.D41-lista_gene_v5.cds.txt”.

- efectul genelor prezente in D41

“1.3.D41_efect_gene_cds.csv”.

- secventa genom pentru D41, D81, Tp

>chrA01
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGTCCATACCTTAGTAGCAAGCTGCTTCCTCGGGGCCTTTCCTCCGGTAGACTTA
CGAGCGGTCTGCTTGGTACGAGCCATCTGAAAACCATTCAAACAATACTAAGATTTACACCATGACCATTTTAATCGATAAATCAACAAATCATCAATAAGGTT
TAAACCTAAATTAACAATCTATAGCGAAATCATTCCTCCTTAACCTAGAATCTAAATCCAGAAGACGAATCGGCTAATCAAACGACTAACGAAGAAACAATCTG
AGGATTCGGGAAAAGAACTTTTACCTTCTCTGCGTCGCTTTCGCTTCTCTGAATTCAAATTAGTGCTCCGATCTCGAAAAGGGAAGATAATGAAAGTAATCTCA
CGTGTAGGGCTGTTGATATTTAATTTATAGGCAACGATGTCAAAAGCAAGTTGACGATCCTACGGTTGTTATCGCTCATCGGTGTGTTTCTCGTGTTTCTTAATC
GTACCGACTAAAATACCACAATATGTGTTTTCTCTCAGCCAGTCAACGGTTCAGATCTACCGTTTACACGTTTGACGAGGATCGCCAATCCTACGTGGTTGATTC
CTATTGGTTGAAT

- lista genelor din D41

>BnaC09g12820D assembled CDS
ATGGCTACTGGAGAAAACAGAACCGTGCAGGAAAATCTAAAGAAACACCTCGCAGTTTCAGTTCGAAACATTCAATGGAGTTATGGAATATTTTGGTCTATCT
CTGCTTCTCAACCAGGAGTGCTGGAATGGGGAGATGGATACTACAATGGAGACATTAAGACGAGGAAGACGATTCAAGCAGTGGAAGTCAAAGCTGACCAG
TTGGGTCTTGAGAGAAGCGATCAGCTTAGAGAGCTTTATGAATCTCTCTCGGTCGCGGAATCCTCAGCCTCCGGTGGCTCTCAGGTCAGTAGACGAGCTTCCG
CTACCGCTCTCTCTCCGGAAGATCTCACCGACACCGAGTGGTACTACCTAGTATGCATGTCTTTCGTCTTCAACATTGGTGAAGGAATTACCGGAGGAGCATTA
GGGAACGGAGAACCAATATGGCTATGTAACGCTCATACCGCCGACAGCAAAGTCTTTACTCGCTCTCTTCTCGCTAAAAGTGCTTCGCTTCTGACAGTAGTTTG
CTTCCCATTTCTTGGAGGAGTCCTT…….

- efectul genelor prezente in D41 (D41_efect_gene_cds.csv)

GeneID	chrN	startN	stopN	Lungime	Descriere proteina	Efect
BnaA01g00010D	chrA01	831	1437	92	signal peptidase subunit-12	F:peptidase activity; P:signal peptide processing; P:sphingoid biosynthetic process; C:integral to membrane; P:sterol biosynthetic process; C:signal peptidase complex
BnaA01g00020D	chrA01	1487	2436	169	tpa: protein kinase domain superfamily protein	F:molecular_function; P:biological_process
BnaA01g00030D	chrA01	2665	5455	614	protein scai-like	P:termination of G-protein coupled receptor signaling pathway
BnaA01g00040D	chrA01	8421	9623	368	kelch repeat-containing f-box family protein	F:molecular_function; P:biological_process

?