

Genomics and Transcriptomics

Class 06 - Sequence Annotation



INSTRUCTOR:

Aureliano Bombarely
Department of Bioscience
Università degli Studi di Milano
aureliano.bombarely@unimi.it

Outline of Topics

1. Genome structural annotation.

1.1. Repeats.

1.2. Genes.

2. Genome functional annotation.

2.1. Functional descriptions.

2.2. Functional classification.



Outline of Topics

1. Genome structural annotation.

1.1. Repeats.

1.2. Genes.

2. Genome functional annotation.

2.1. Functional descriptions.

2.2. Functional classification.



1. Genome structural annotation.

Recommended review:

Review

Nature Reviews Genetics **13**, 329-342 (May 2012) | doi:10.1038/nrg3174

A beginner's guide to eukaryotic genome annotation

Mark Yandell & Daniel Ence

The falling cost of genome sequencing is having a marked impact on the research community with respect to which genomes are sequenced and how and where they are annotated. Genome annotation projects have generally become small-scale affairs that are often carried out by an individual laboratory. Although annotating a eukaryotic genome assembly is now within the reach of non-experts, it remains a challenging task. Here we provide an overview of the genome annotation process and the available tools and describe some best-practice approaches.



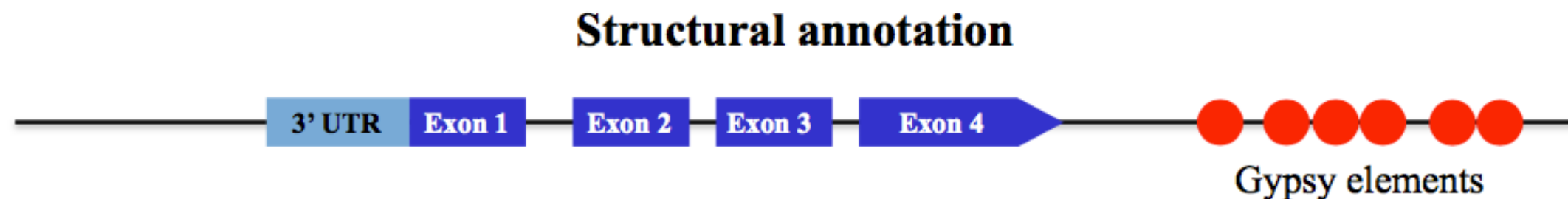
1. Genome structural annotation.

Structural annotation consists of the identification of genomic elements.

- ORFs and their localization
- gene structure
- coding regions
- location of regulatory motifs
- repeats

Functional annotation consists of attaching biological information to genomic elements.

- biochemical function
- biological function
- involved regulation and interactions
- expression



Functional annotation
kinase



1. Genome structural annotation.

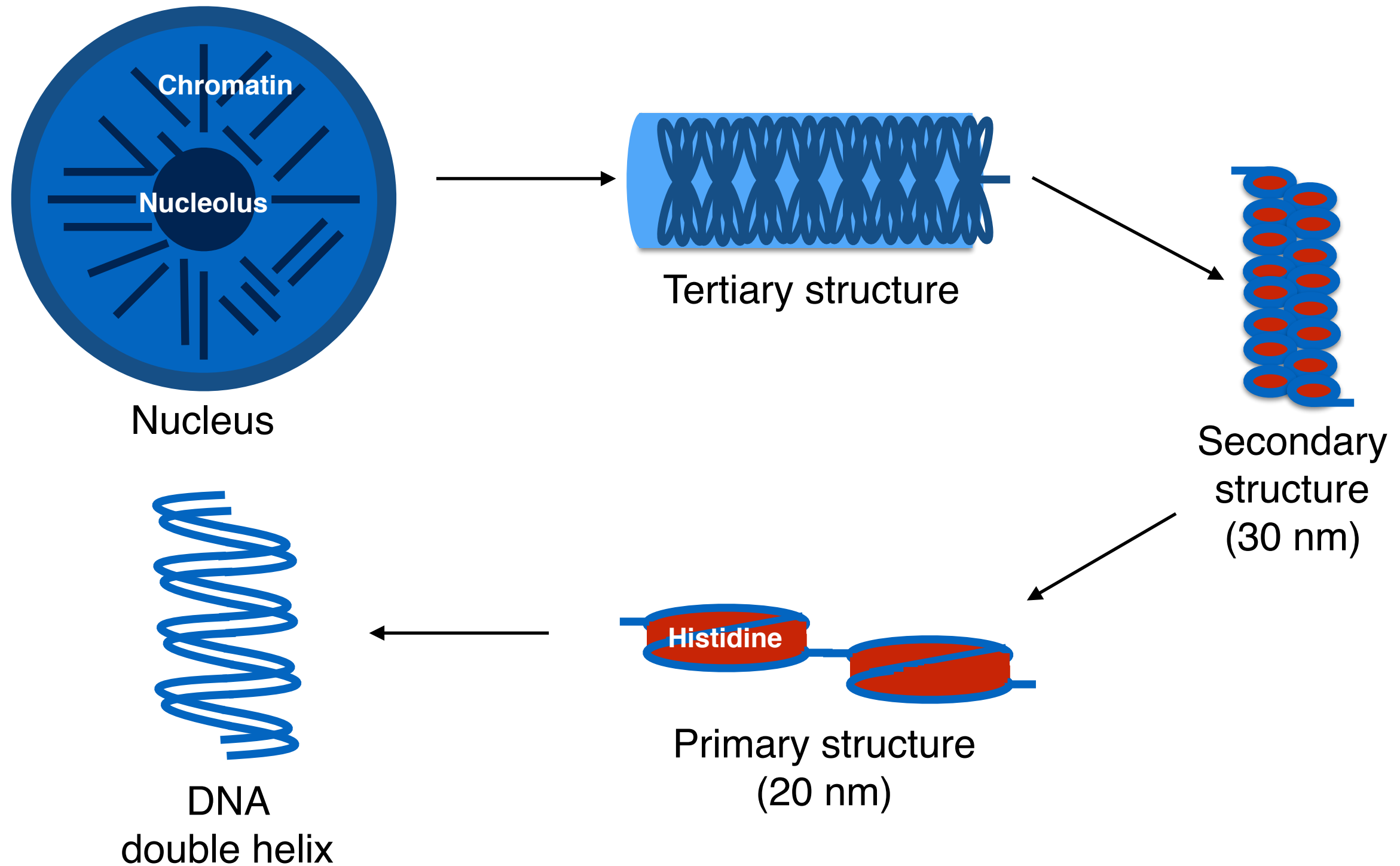
Two different types of annotations:

- **Automatic**, based in the pattern recognition through different algorithm such as Markov Chain models.
 - ◆ The quality depends of the quality of the predictors training. Usually it produces good quality for repetitive elements (e.g. hexokinases) and bad quality for singularities (e.g. single copy gene with complex intron-exon structure).
- **Manual**, based in the human supervised manual inspection of the data.
 - ◆ Specially good quality for functional annotations. Not feasible outside model genomes such as *Saccharomyces cerevisiae* and *Arabidopsis thaliana*.



1. Genome structural annotation.

(Eukaryotic) Genome physical structure



Genome sequence structure

GCGTGCAGGACGATGACGCAGAAGCTGGCAGACGGATGCGAGCAGCAGCAGTGACGT

GACGACGGACGACGACGACGACGACGACGACGACGACGACGACGACGAGACGACGACGAA

Repeat

GACGACGACGACGTGACGCAGCAGACTGAT**TATACAGCTTGATATACGTACGGTATAA**

Promotor with TATA box

CGTGACGACGACTATAGCACACAGTGAAACGACAGTGACGAGCAGGTAGACGATGAC

GCAGCAAAACCATAGCA**ATGGCCGCATATTATGACGCAGAC**CGGACTGACGTGACGT

Gene with two exons

GACTTACGAGCATGCAGCAGTGCACGTGCAGTGACGTGACGTTTTTGGACGTAGCAGT

Identification of genomic elements



1. Genome structural annotation.

Identification of genomic elements

1. Sequence homology with known elements (e.g. read from RNA-Seq, known transposable element...).
2. Sequence identification based in sequence patterns (e.g. ATG, GAx where x is 20 or more).



Outline of Topics

1. Genome structural annotation.

1.1. Repeats.

1.2. Genes.

2. Genome functional annotation.

2.1. Functional descriptions.

2.2. Functional classification.



1.1. Repeats.

Genomic repetitive element.

Repeated sequences (repetitive elements, or repeats) are ***patterns of nucleic acids (DNA or RNA) that occur in multiple copies throughout the genome***. The functions and descriptions of these sequences are currently being characterized by scientists. Repetitive DNA was first detected because of its rapid reassociation kinetics.

-Wikipedia (repeated sequences)



Types of repetitive elements

- **Tandem repeats.**

- Tandem gene paralogs (e.g R-genes).
- rDNA.
- Satellite DNA (e.g. centromeres and telomeres).

- **Dispersed repeats.**

- Gene paralogs (e.g. DHARs).
- tDNA
- Transposons
 - Class I (Retrotransposons): LINE, SINE and LTR transposons.
 - Class II: DNA transposons

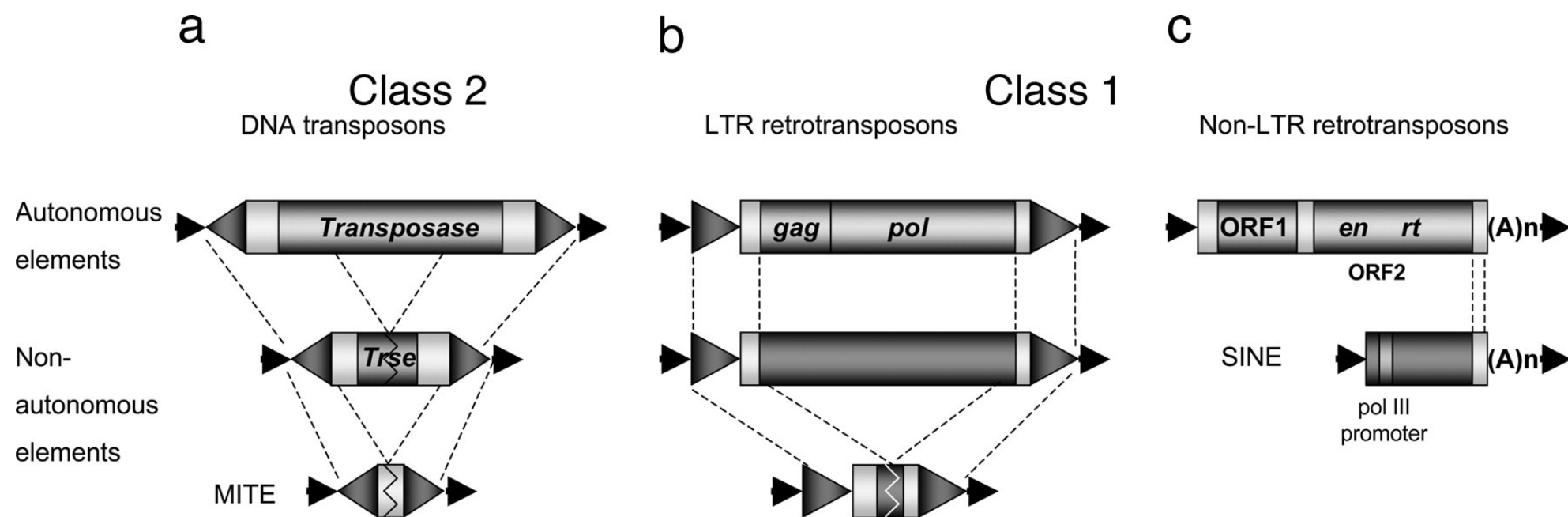


1.1. Repeats.

Transposons (or transposable element)

A transposable element (TE or transposon) ***is a DNA sequence that can change its position within the genome, sometimes creating or reversing mutations and altering the cell's genome size.*** Transposition often results in duplication of the TE. Barbara McClintock's discovery of these jumping genes earned her a Nobel prize in 1983.[1]

-Wikipedia (transposable element)

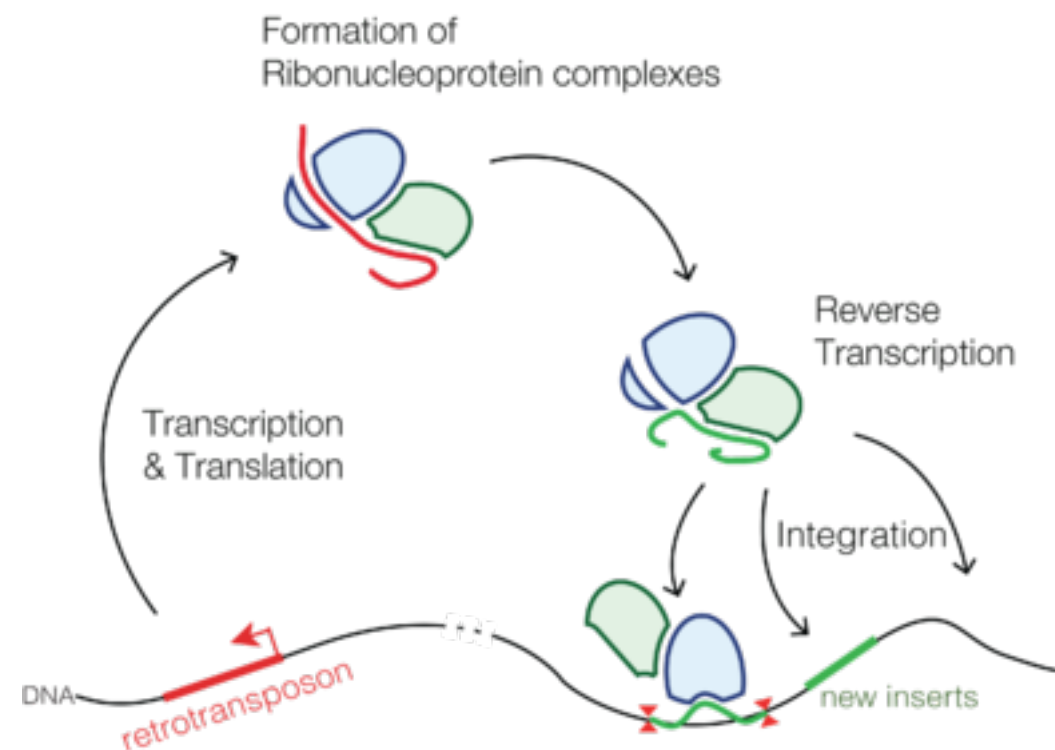


1.1. Repeats.

Class I: Retrotransposons

Retrotransposons are transposable DNA elements (transposons) that ***employ retroviral-like reverse transcription during the process of transposition***: retrotransposon DNA is first transcribed into an RNA template, then reverse transcribed into DNA, and then inserted into a new genomic site.

-NCI Thesaurus



Class I: Retrotransposons

- **Long Terminal Repeat (LTR) Transposons**, have a LTR that ranges from 100 to 5,000 bp. They are subdivided in:
 - ***Ty1 Copia***. They are divided in five lineages (Sirevirus/Maximus, Oryco/Ivana, Retrofit/Ale, TORK and Bianca)
 - ***Ty3-Gypsy***. Some lineages are Errantiviruses, Chromoviruses and Metaviruses.
 - ***Endogenous Retroviruses (ERV)***.
- ***Non-LTR Transposons***.
 - ***Long Interspersed Nuclear Elements (LINE)***.
 - ***Short Interspersed Nuclear Elements (SINE)***.



1.1. Repeats.

Class II: DNA Transposons

“They usually have a simple structure with a short terminal inverted repeat (TIR) ($\approx 10\text{--}40$ bp, but can be up to ≈ 200 bp) and a single gene encoding the transposase. Transposase binds in a sequence-specific manner to the ends of its encoding element and to the ends of nonautonomous family members. Once bound, transposase initiates a cut-and-paste reaction whereby the element is excised from the donor site (generating an “empty site”) and inserted into a new site in the genome”.

Class II TE Plant superfamilies:

- TIR order
 - CACTA.
 - Mutator.
 - hAT
 - Tc1/Mariner
 - PIF/Harbinger
- Helitron order



1.1. Repeats.

Software used to identify repeats

PROGRAM	TYPE	APPROACH	CITATION
RepeatMasker	Library based.	Search by homology	Smit et al. 1996
PLOTREP (Censor)	Library based.	Search by homology	Toth et al. 2006
LTR_STRUCT	Library based.	Search for LTR Transposons	McCarthy and McDonald 2003
Greedier	Library based.	Search by homology. Nested elements	Li et al. 2008
RTAnalyzer	Signature based	LINEs, Alus and retrogenes using Blast.	Lucier et al. 2007
FINDMITE	Signature based	Search for MITEs	Tu 2001
HelitronFinder	Signature based	Helitron specific	Du et al. 2008
RECON	Ab-initio	Multiple alignments to detect families	Bao and Eddy, 2002
PILER	Ab-initio	Multiple alignments to detect families	Edgar and Myers, 2005
RepeatScout	Ab-initio	Multiple alignments to detect families	Prices et al. 2005
RepeatFinder/REPuter	Ab-initio	Multiple alignments to detect families	Volfovsky et al. 2001
RepeatModeler	Pipeline	Uses RepeatScout/TRF and RECON	Unpublished
RepeatRunner	Pipeline	Uses PILER, RepeatMasker and BlastX	Smith et al. 2007



1.1. Repeats.

Example of annotated repeats in the petunia genomes

		<i>P. axillaris</i> v1.6.2		<i>P. inflata</i> v1.0.1	
Type	Repeat	Size (Mb)	% Genome	Size (Mb)	% Genome
Low Complexity	Simple Repeat	26.79	0.90	23.70	0.80
	Nucleotide Rich/Satellites	4.73	0.15	4.45	0.15
CLASS II: DNA transposons	CMC	5.71	0.19	5.06	0.17
	Harbinger	1.35	0.04	1.31	0.04
	hAT	4.28	0.14	4.22	0.14
	Maverick	0.03	<0.01	0.04	<0.01
	MULE	4.82	0.16	5.24	0.18
	TcMar	3.87	0.13	4.01	0.13
	Other	1.11	0.04	1.18	0.04
CLASS I: Non LTR transposons	LINE	12.80	0.43	14.07	0.47
	SINE	1.40	0.05	1.67	0.06
CLASS I: LTR transposons	LTR/Ty1 Copia	62.07	2.09	57.83	1.94
	LTR/Ty3 Gypsy	109.37	3.68	103.75	3.49
	Other LTR	1.07	0.04	1.07	0.04
RNA	rRNA	0.12	<0.01	0.11	<0.01
	tRNA	0.03	<0.01	0.03	<0.01
	snRNA	0.01	<0.01	0.01	<0.01
Unknown/Unspecified		565.84	44.84	542.24	42.03
		806.74	64.02	771.27	59.79



Outline of Topics

1. Genome structural annotation.

1.1. Repeats.

1.2. Genes.

2. Genome functional annotation.

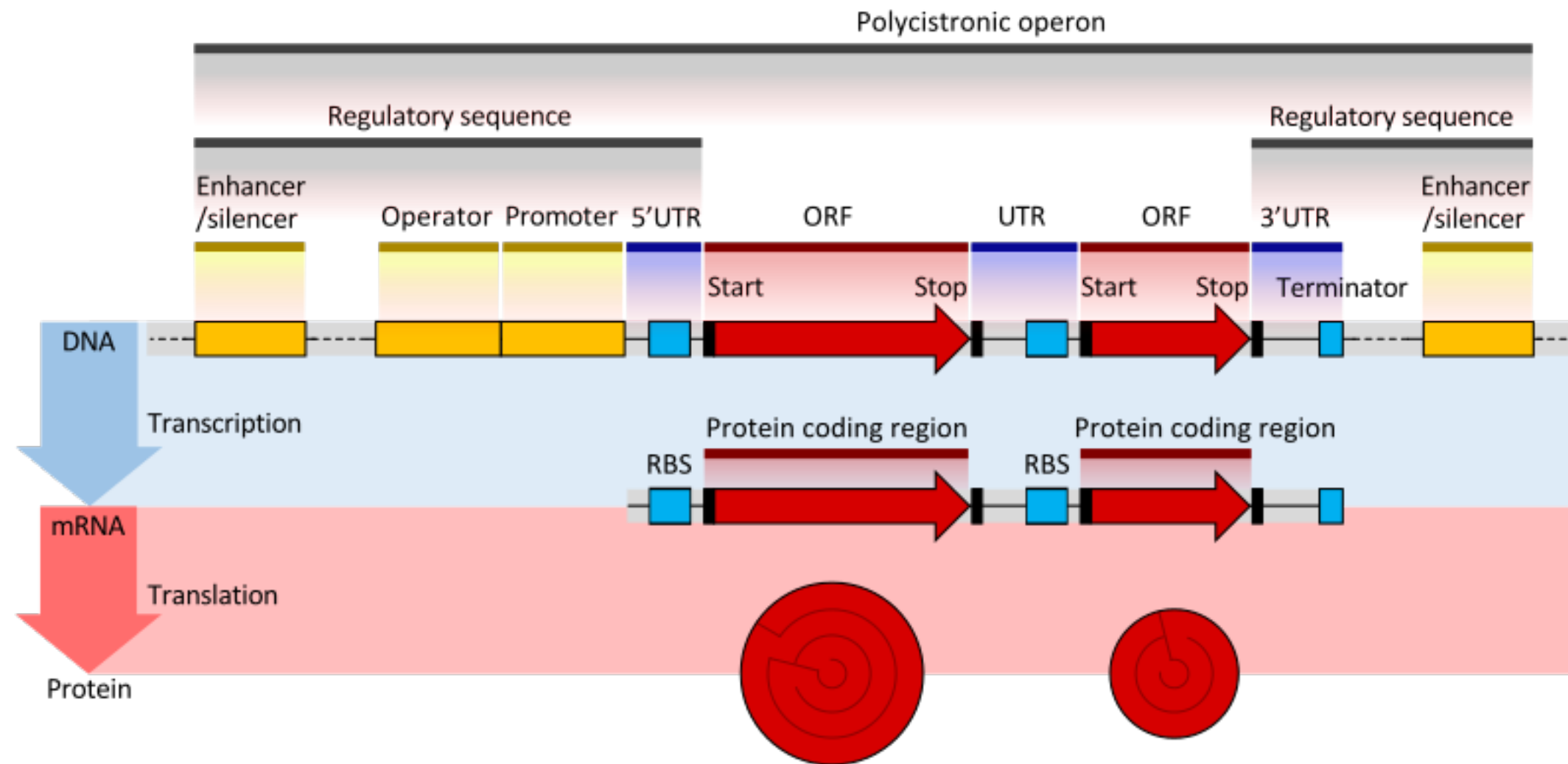
2.1. Functional descriptions.

2.2. Functional classification.



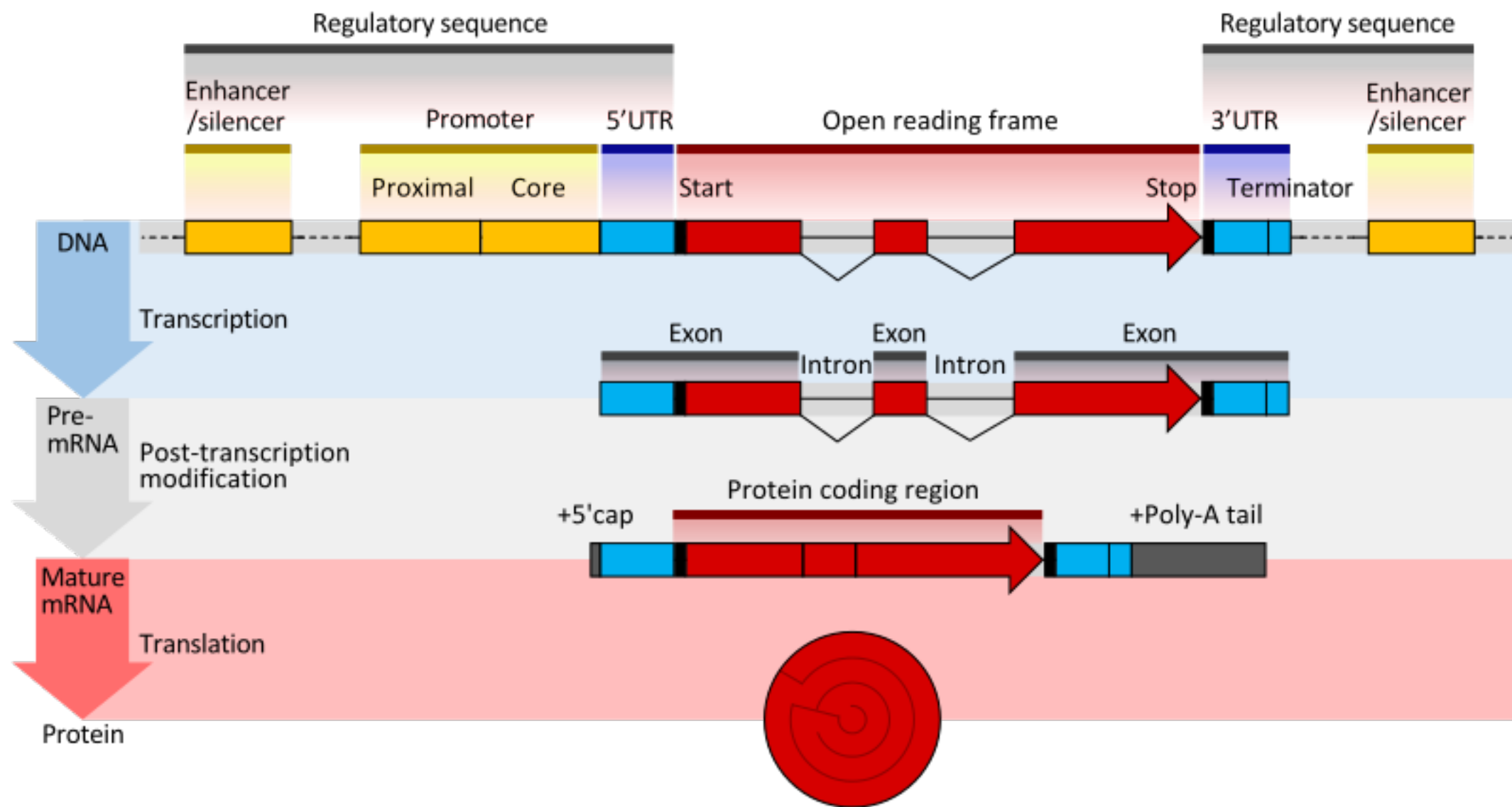
1.2. Genes.

Structure of a Prokaryotic Gene



1.2. Genes.

Structure of an Eukaryotic Gene



Gene structural annotation

- ***Ab-initio* gene prediction.**
 - Rely in mathematical models to determine intron-exon structure.
 - Do not external evidence (e.g. ESTs).
 - Do not report untranslated regions (UTRs).
 - Accuracy intron-exon structure < 60%.
- **Evidence-driven gene prediction.**
 - ESTs, RNA-Seq and know protein data need to be aligned.
 - Good accuracy.
 - Computationally intensive.



1.2. Genes.

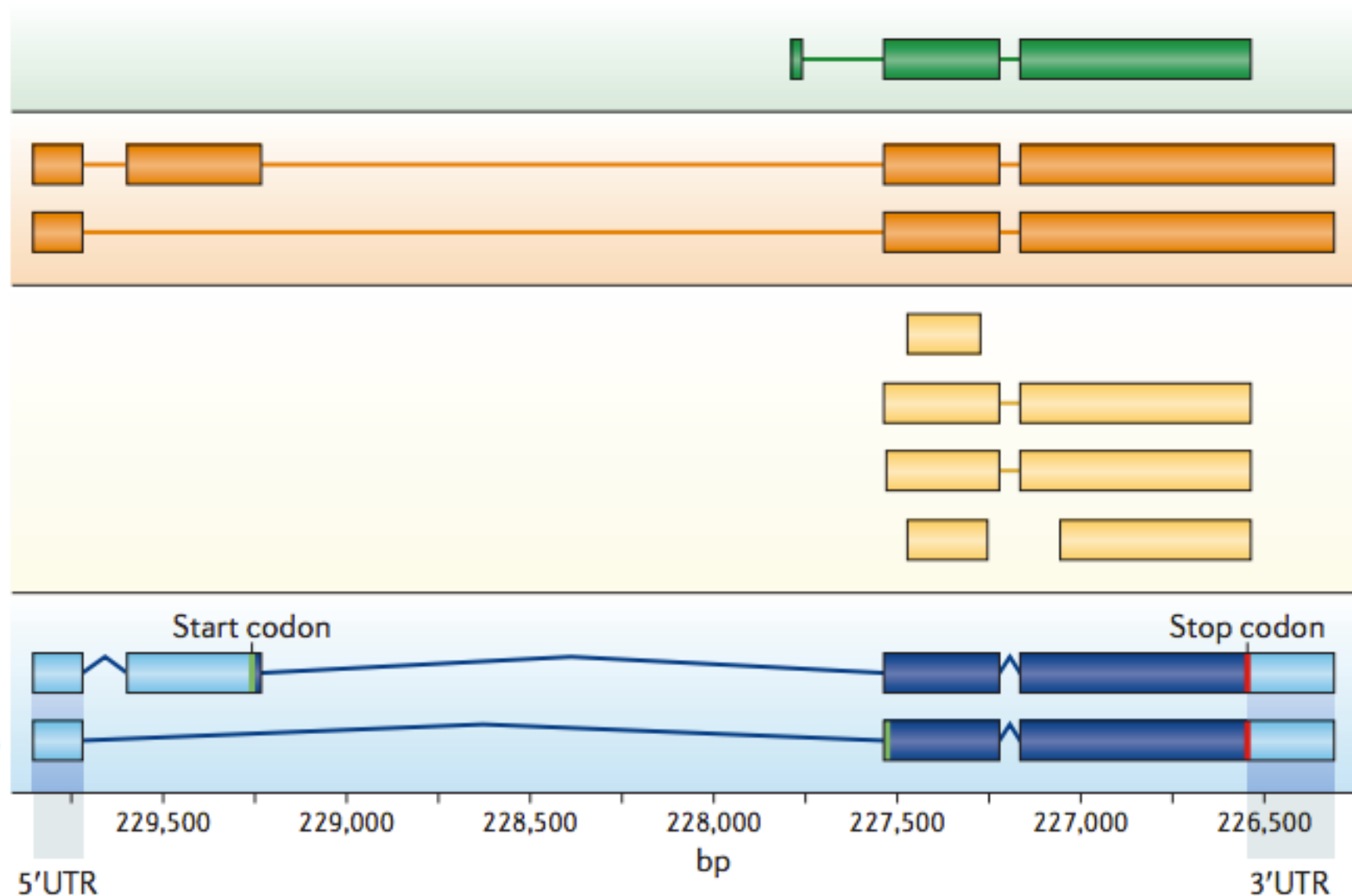
Gene structural annotation

Gene prediction
(SNAP)

mRNA or EST evidence
(Exonerate)

Protein evidence
(BLASTX)

Gene annotation resulting
from synthesizing all
available evidence
(two alternative splice forms)



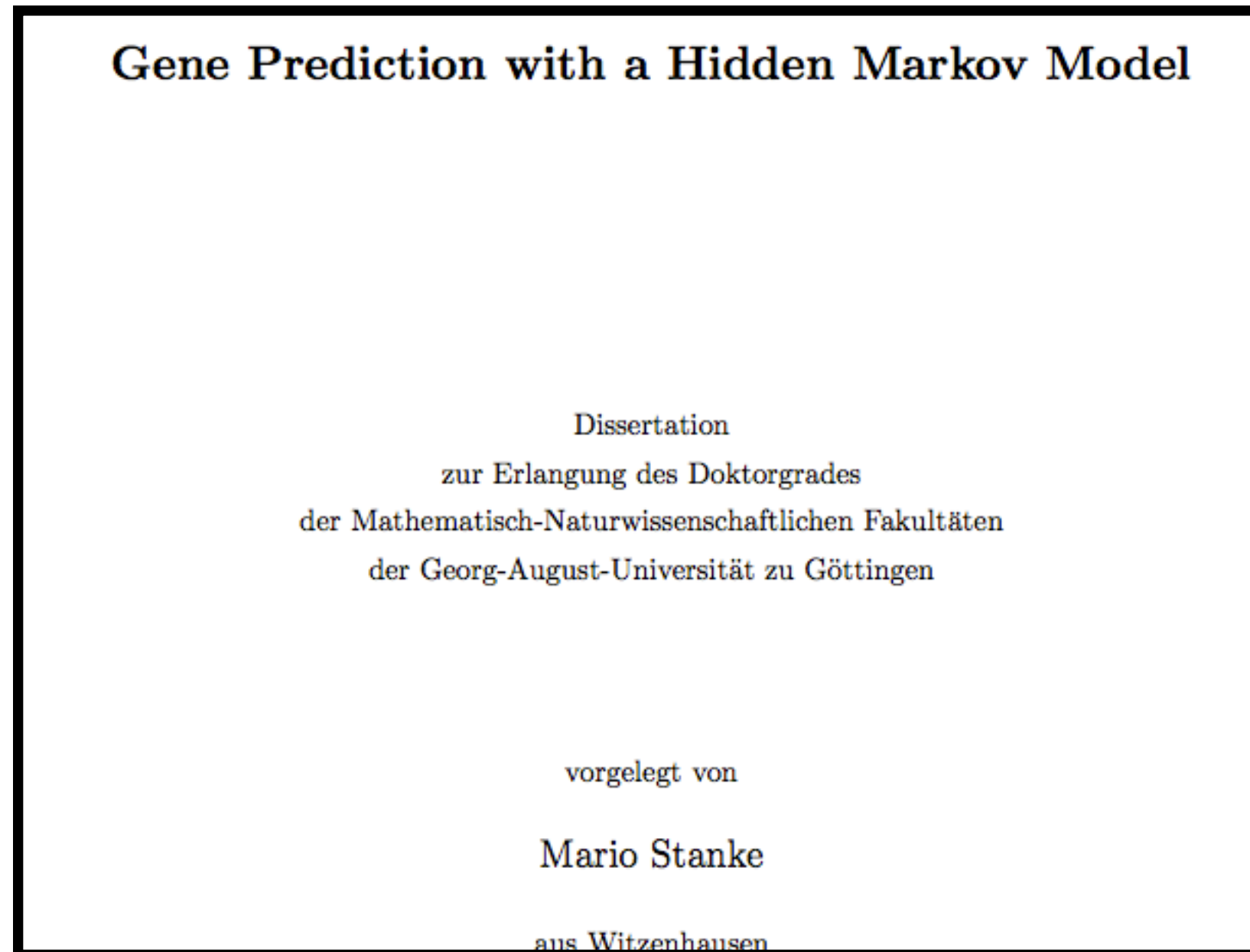
1.2. Genes.

Software used for gene prediction

PROGRAM	TYPE	APPROACH	CITATION
Augustus	Ab-Initio/Evidence	Generalized Hidden Markov Models (HMM)	Stanke et al. 2006
Eugene	Ab-Initio/Evidence	HMM + Evidence alignment	Foissac et al. 2008
FGENESH	Ab-initio	HMM	Solovyev et al. 2006
GeneMark	Ab-initio	HMM + Unsupervised training	Ter-Hovhannisyan et al. 2008
GENSCAN	Ab-initio	Fourier transformation	Burge and Karlin, 1998
SNAP	Ab-initio	Semi-HMM	Korf, 2004
Exonerate	Evidence	Sequence alignment	Slater and Birney, 2005
GeneWise	Evidence	Sequence alignment	Birney et al. 2004
GenomeScan	Evidence	Sequence alignment	Yeh et al. 2001
Tophat/Cufflinks	Evidence	Based on RNA-Seq alignments	Trapnell et al. 2012
MAKER	Pipeline	Integrative approach with different programs	Holt and Yandell, 2011
PASA	Pipeline	Integrative approach with different programs	Haas et al. 2003
Ensembl	Pipeline	Integrative approach with different programs	Ashurst et al. 2005



More interested in the HMM for gene prediction ...



1.2. Genes.

Example of annotated genes in the petunia genomes

	<i>P. axillaris</i> v1.6.2	<i>P. inflata</i> v1.0.1	<i>S. lycopersicum</i> ITAG 2.40	<i>S. tuberosum</i> DM_V4.03
Genes Models	35,812	39,408	34,725	39,027
mRNA	32,928	36,697	34,725	56,212
tRNA	2,884	2,711	0	0
exons/gene model	4.85	4.78	4.61	4.24
five_prime_UTRs	6,573	6,822	13,548	0*
three_prime_UTRs	11,006	11,050	15,343	0*
average mRNA length (bp)	1,172	1,141	1,209	1,415
average protein length (Aa)	393	385	344	302
Gene space size (Mb)	140	152	110	118



Outline of Topics

1. Genome structural annotation.

1.1. Repeats.

1.2. Genes.

2. Genome functional annotation.

2.1. Functional descriptions.

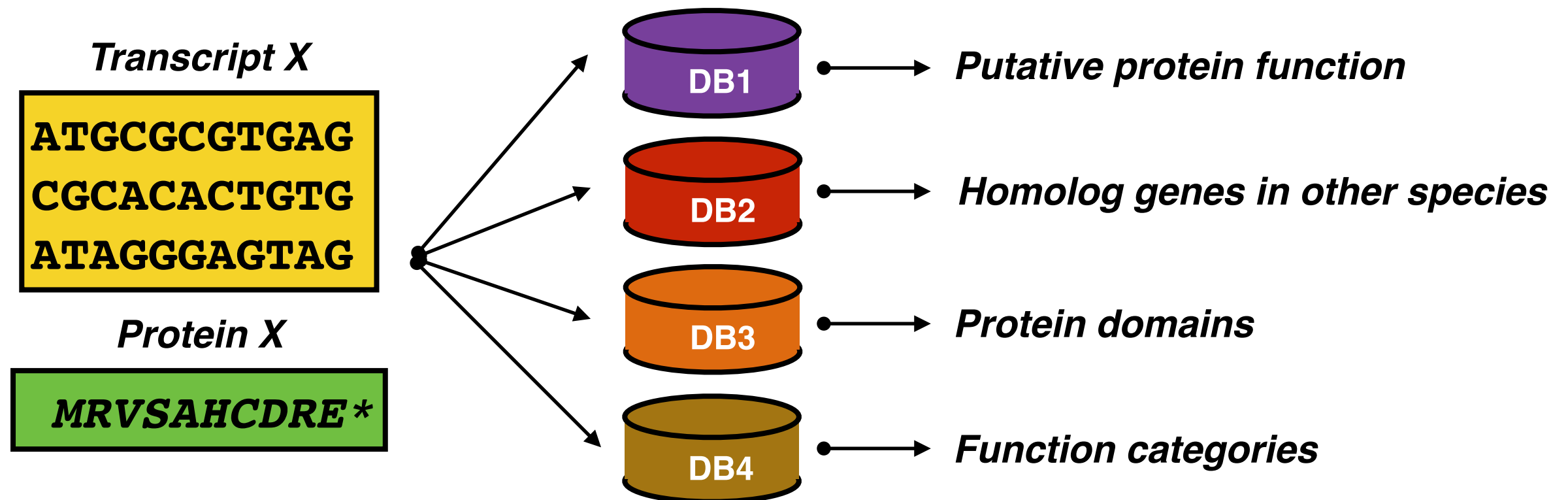
2.2. Functional classification.



2. Genome functional annotation.

Genome functional annotation

Attaching of the biological annotation to the gene models.



Outline of Topics

1. Genome structural annotation.

1.1. Repeats.

1.2. Genes.

2. Genome functional annotation.

2.1. Functional descriptions.

2.2. Functional classification.



Genome functional annotation

Gene functional description association is usually performed through sequence homology search with protein databases with attached functions.

- Whole protein sequence homology, using BlastP or BlastX with some specific cutoffs (e.g. maximum e-value $1e-10$)
- Protein domain homology, using InterProScan.



Genome functional annotation: Whole sequence homology

Protein sequence databases:

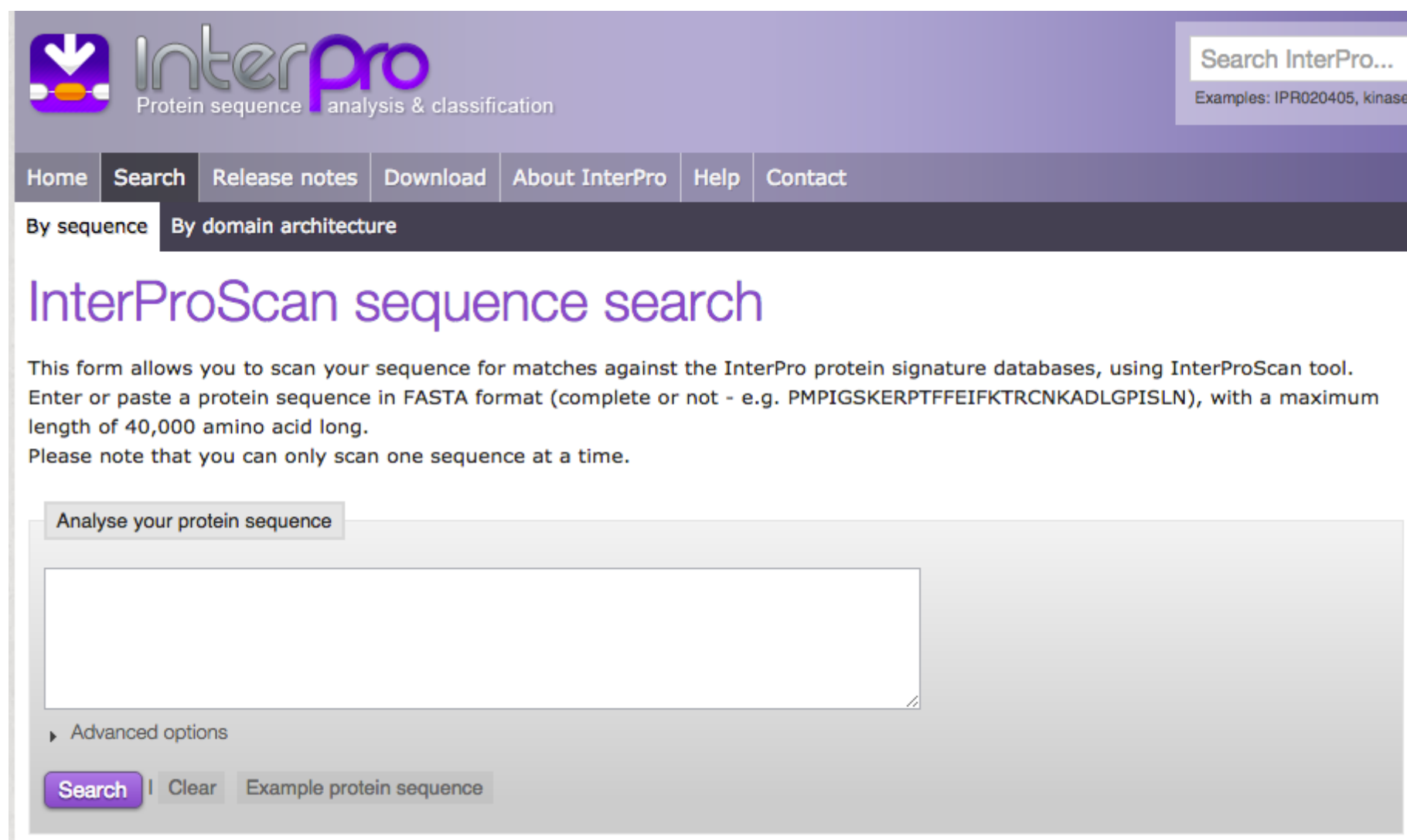
1. Protein manually curated database
2. Protein electronically annotated repository
3. Species/Genus/Family specific databases
4. Function specific databases
5. Tool specific databases



2.1. Functional descriptions.

Genome functional annotation: Protein domain homology

The tool used for the protein domain search is InterProScan. It combines different tools such as BlastP (sequence homology), SignalP (signal peptide prediction)...














The screenshot shows the InterProScan web interface. At the top, there is a purple header with the InterPro logo and the text "Protein sequence analysis & classification". To the right of the header is a search bar labeled "Search InterPro..." with examples: "IPR020405, kinase,". Below the header is a navigation menu with links: Home, Search, Release notes, Download, About InterPro, Help, and Contact. Under the "Search" link, there are two tabs: "By sequence" (selected) and "By domain architecture". The main heading is "InterProScan sequence search". Below this, a paragraph explains: "This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool. Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKERPTFFEIFKTRCNKADLGPISLN), with a maximum length of 40,000 amino acid long. Please note that you can only scan one sequence at a time." Below this text is a large text input field labeled "Analyse your protein sequence". At the bottom, there is a section for "Advanced options" with a "Search" button, a "Clear" button, and a link to "Example protein sequence".



2.1. Functional descriptions.

The following databases make up the InterPro Consortium:

	<p>PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs. PROSITE is base at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland.</p>
	<p>HAMAP stands for High-quality Automated and Manual Annotation of Proteins. HAMAP profiles are manually created by expert curators. They identify proteins that are part of well-conserved proteins families or subfamilies. HAMAP is based at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.</p>
	<p>Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Pfam is based at the Wellcome Trust Sanger Institute, Hinxton, UK.</p>
	<p>PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family or domain. PRINTS is based at the University of Manchester, UK.</p>
	<p>ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches. ProDom is based at PRABI Villeurbanne, France.</p>
	<p>SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. SMART is based at at EMBL, Heidelberg, Germany.</p>
	<p>TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. TIGRFAMs is based at the J. Craig Venter Institute, Rockville, MD, US.</p>
	<p>PIRSF protein classification system is a network with multiple levels of sequence diversity from superfamilies to subfamilies that reflects the evolutionary relationship of full-length proteins and domains. PIRSF is based at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, US.</p>
	<p>SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY is based at the University of Bristol, UK.</p>
	<p>CATH-Gene3D database describes protein families and domain architectures in complete genomes. Protein families are formed using a Markov clustering algorithm, followed by multi-linkage clustering according to sequence identity. Mapping of predicted structure and sequence domains is undertaken using hidden Markov models libraries representing CATH and Pfam domains. CATH-Gene3D is based at University College, London, UK.</p>
	<p>PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise. These subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function, as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences. PANTHER is based at at University of Southern California, CA, US.</p>



Outline of Topics

1. Genome structural annotation.

1.1. Repeats.

1.2. Genes.

2. Genome functional annotation.

2.1. Functional descriptions.

2.2. Functional classification.



2.2. Functional classification.

Functional classification is a way to assign a protein function to a specific category such as “transcription factor”. So far, the methodology more extensively used is the **Gene Ontology**.

Gene ontologies:

Structured controlled vocabularies (ontologies) that describe **gene products** in terms of their associated

- **biological processes**,
- **cellular components** and
- **molecular functions**

in a species-independent manner



2.2. Functional classification.

Biological processes.

Recognized series of events or molecular functions. A process is a collection of molecular events with a defined beginning and end. E.g. Carotenoid biosynthesis.

Cellular components.

Describes locations, at the levels of subcellular structures and macromolecular complexes. E.g. nucleus.

Molecular functions

Describes activities, such as catalytic or binding activities, that occur at the molecular level. E.g. Phosphatase



2.2. Functional classification.

<http://geneontology.org/>

Gene Ontology Consortium

Home
Documentation ▾
Downloads ▾
User stories ▾
Community ▾
Tools ▾
About ▾
Contact us

Search GO data

terms and gene products

Search

Enrichment analysis

Your gene IDs here...

Search

Submit

Gene Ontology Consortium

CytoScape EM

Enrichment Map Cytoscape Plugin

Highlighted GO term

Representing "phases" in GO biological process

The GOC has recently introduced a new term [biological phase \(GO:0044848\)](#), as a direct subclass of biological process. This class represents a distinct period or stage during which biological processes can occur.

[more](#)

Random FAQs

- [What is an OWL file?](#)



2.2. Functional classification.

Additionally there are pipelines to integrate several approaches (e.g. Blast sequence homology and protein domain searches). The most popular ones are:

- ***Blast2GO*** ().
 - Intuitive GUI.
 - Slow and poorly configurable.
- ***Mercator/Mapman*** (<http://mapman.gabipd.org/web/guest/applications>)
 - Metabolic pathways oriented.
 - Nice visualization by boxes.
- ***AHRD*** ().
 - Quick. Score system for the quality of the annotation.
 - Limited to old Blast annotations (Blastall).



2.2. Functional classification.

The Plant Cell, Vol. 28: 1759–1768, August 2016, www.plantcell.org © 2016 American Society of Plant Biologists. All rights reserved.

COMMENTARY

Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences^{OPEN}

Elisabeth Veeckman,^{a,b} Tom Ruttink,^{a,b} and Klaas Vandepoele^{b,c,d,1}

^aInstitute for Agricultural and Fisheries Research, Plant Sciences Unit, Growth and Development, B-9090 Melle, Belgium

^cDepartment of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium

^dDepartment of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Ghent, Belgium

^bBioinformatics Institute Ghent, Ghent University, B-9052 Ghent, Belgium

ORCID IDs: 0000-0003-0510-6317 (E.V.); 0000-0002-1012-9399 (T.R.); 0000-0003-4790-2725 (K.V.)

