ProS

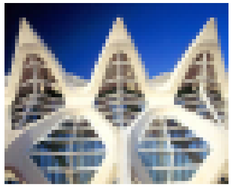
Centro de Investigación
en Métodos de
Producción de Software

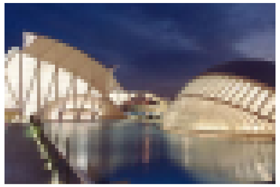
Conceptual Modeling of the Human Genome: Does it Really Worth?

Prof. Oscar Pastor
ProS Research Center
Technical University of Valencia, Spain

RESEARCH CHALLENGES ON INFORMATION SYSTEMS, RCIS 2009

UNIVERSIDAD
POLITECNICA
DE VALENCIA



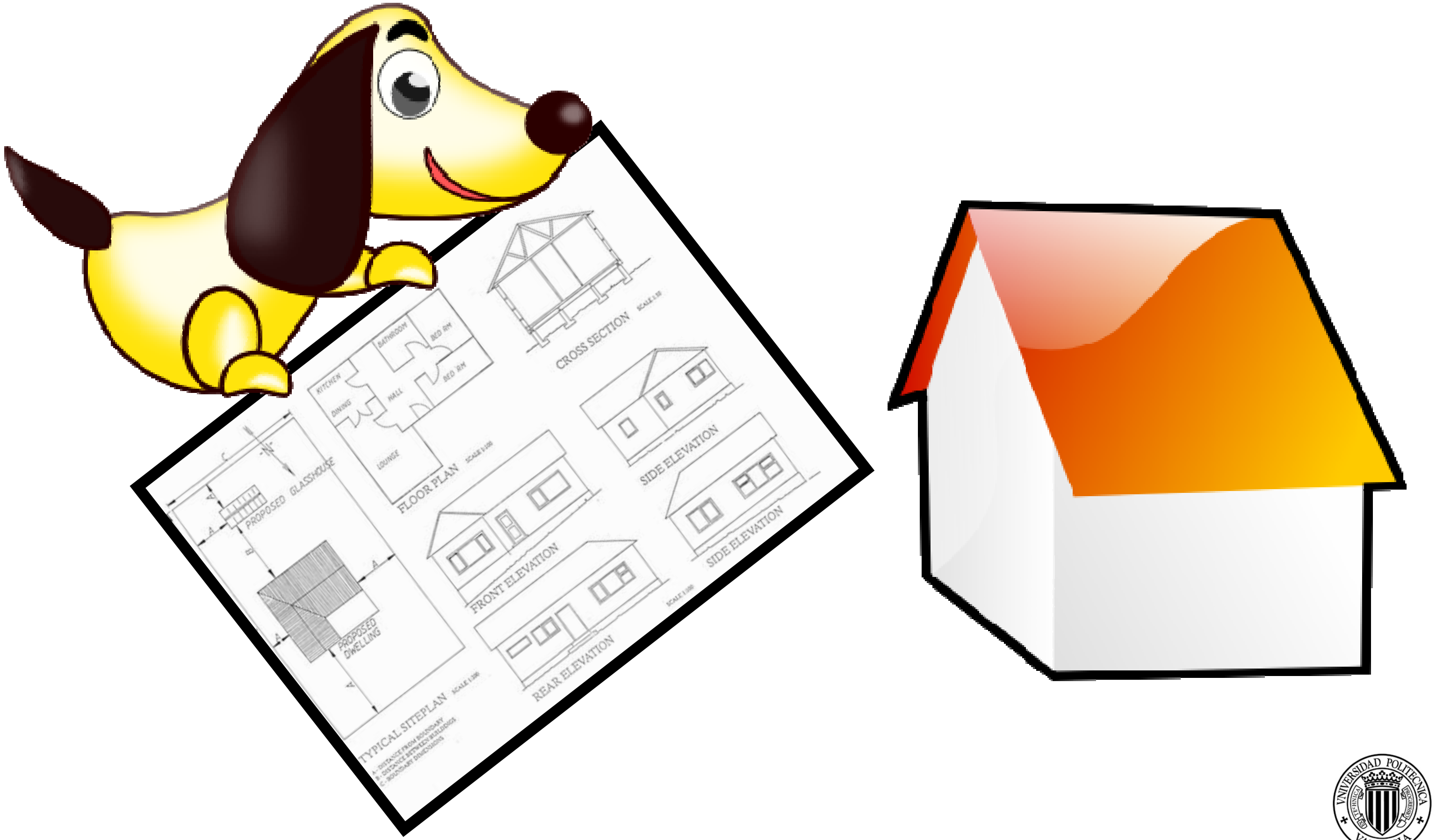


Agenda

1. Why a Keynote on CM and the Human Genome?
2. Problem Statement
3. The Role of Conceptual Modeling
4. The Present
5. The Short-Term Future
6. Understanding the Domain (Problem Space)
7. Building the ER Model / Data Base (Solution Space)
8. Conclusions

Experience in Conceptual Modeling

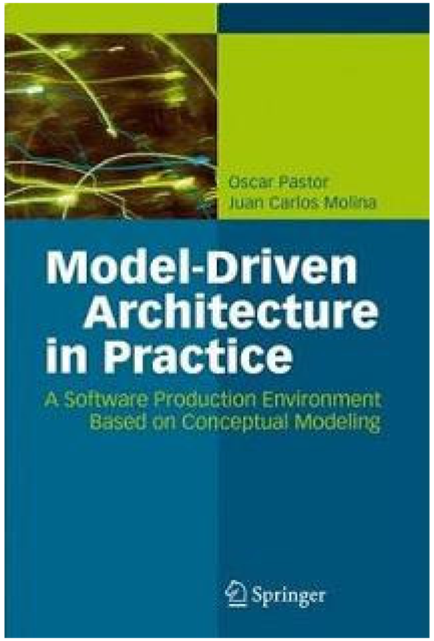
- You don’t need a plan to build your dog house



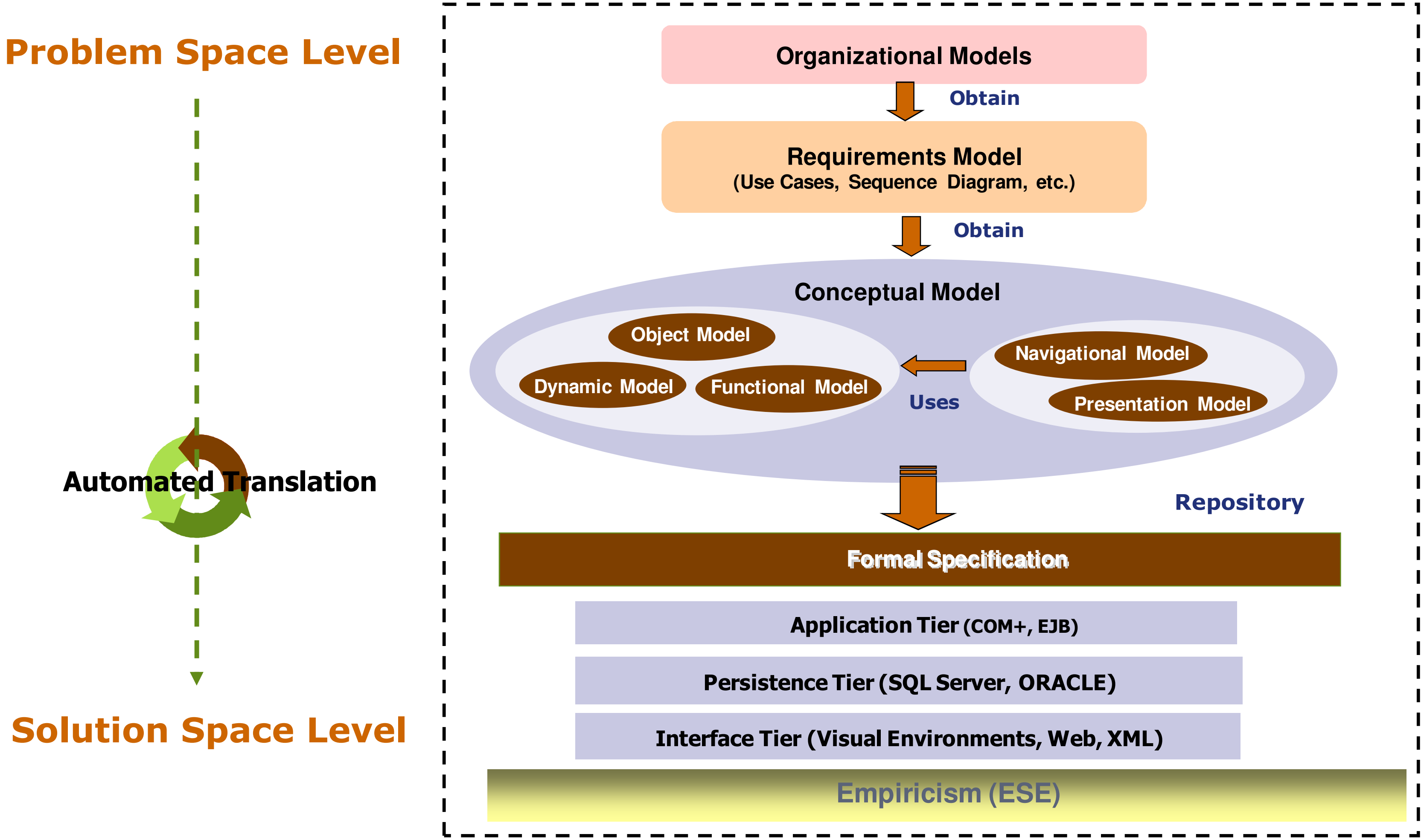
Experience in Conceptual Modeling

- We have been building
  - Traditional Information Systems
  - Web-based Information Systems
  - SOA-based systems
  - Pervasive Systems

  - … but, what is next?

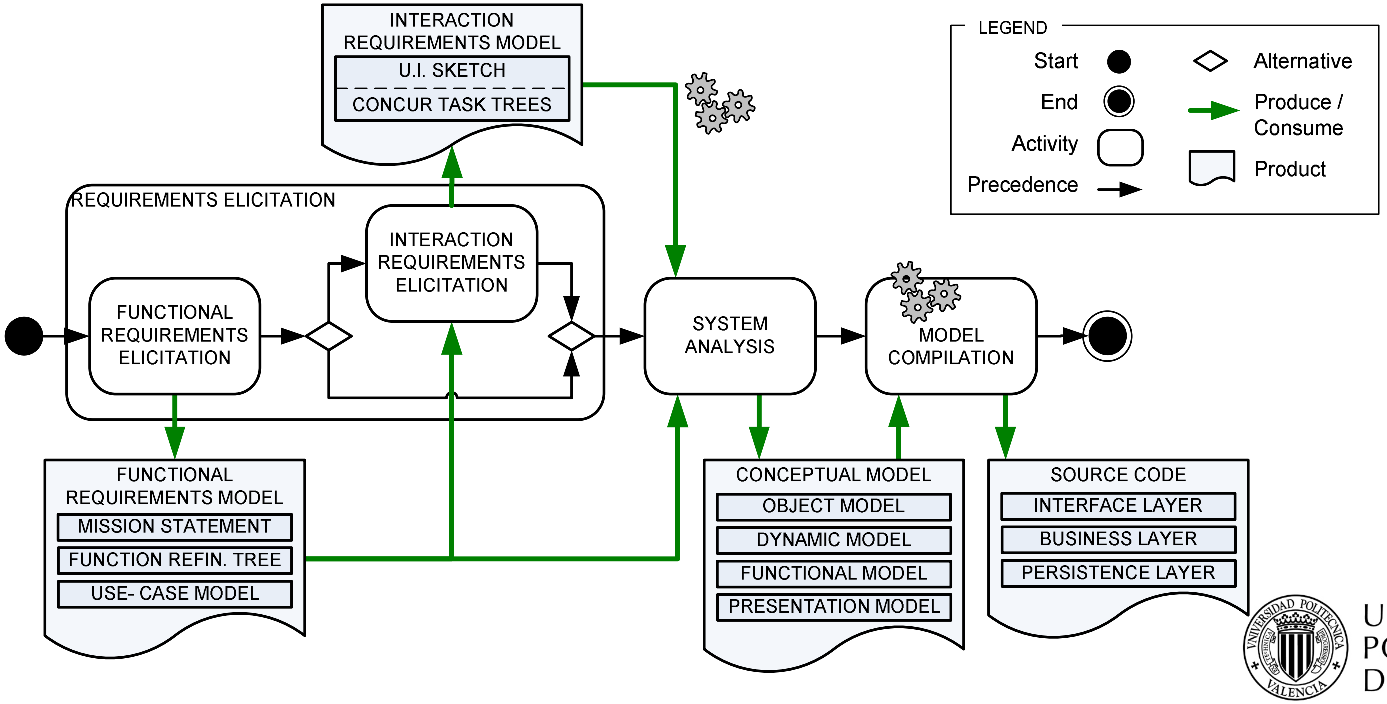

Oscar Pastor
Juan Carlos Molina
Model-Driven Architecture in Practice
A Software Production Environment Based on Conceptual Modeling
Springer

The OO-Method Approach



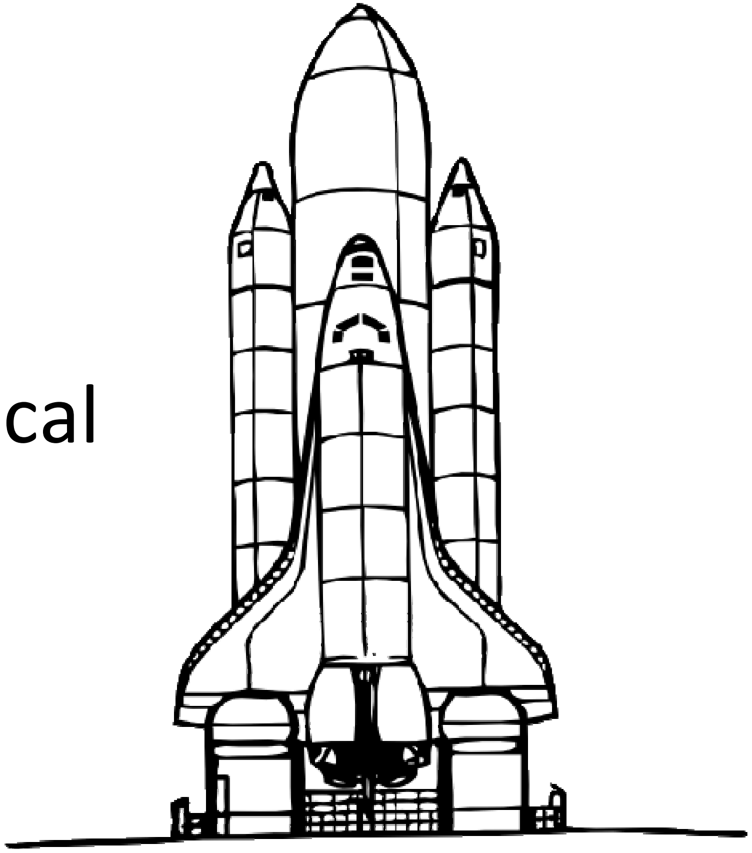
The OO-Method Approach

- **We try to clarify our software development process**
- Also, some gaps are being filled: an **Interaction Requirements Model** is being proposed, based on user-interface sketches that are supported a forest of task trees (ConcurTaskTrees notation)



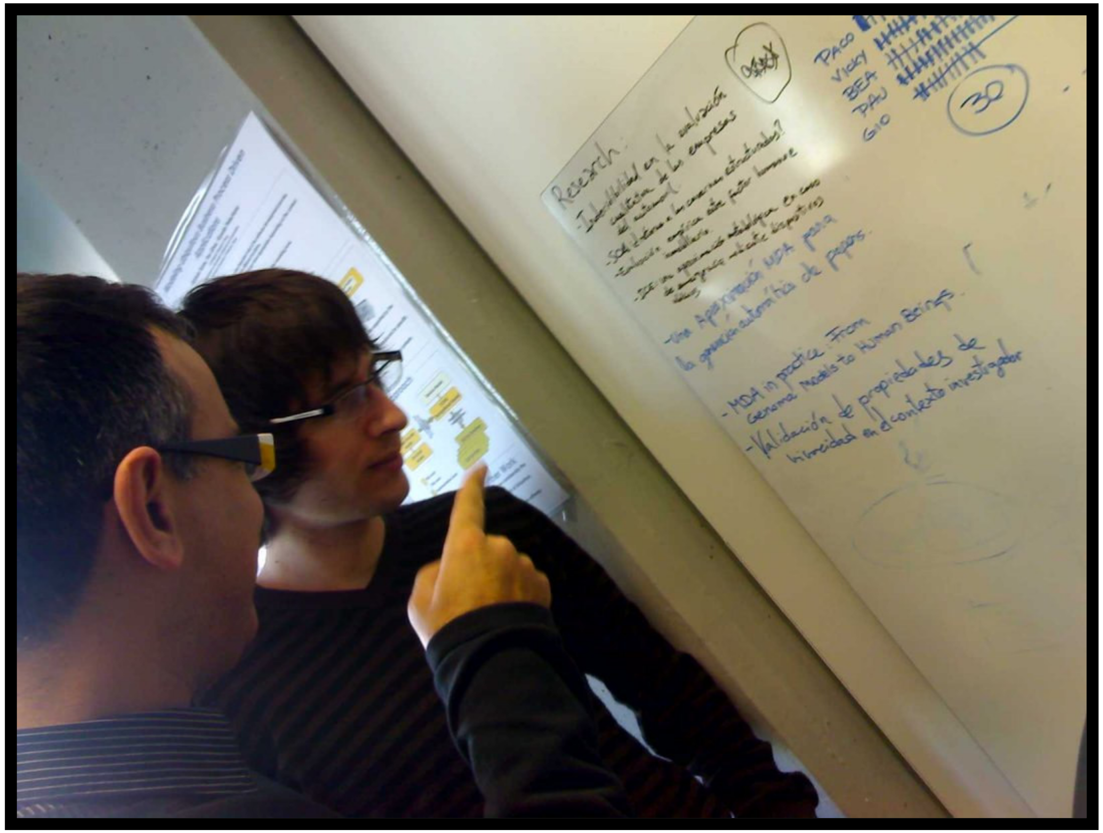
What is the most complex system you can imagine?

- Aircraft control?
- Weather prediction?
- Digital TV?
- Videogames?
- Web 2.0 socio-geographical mashups?

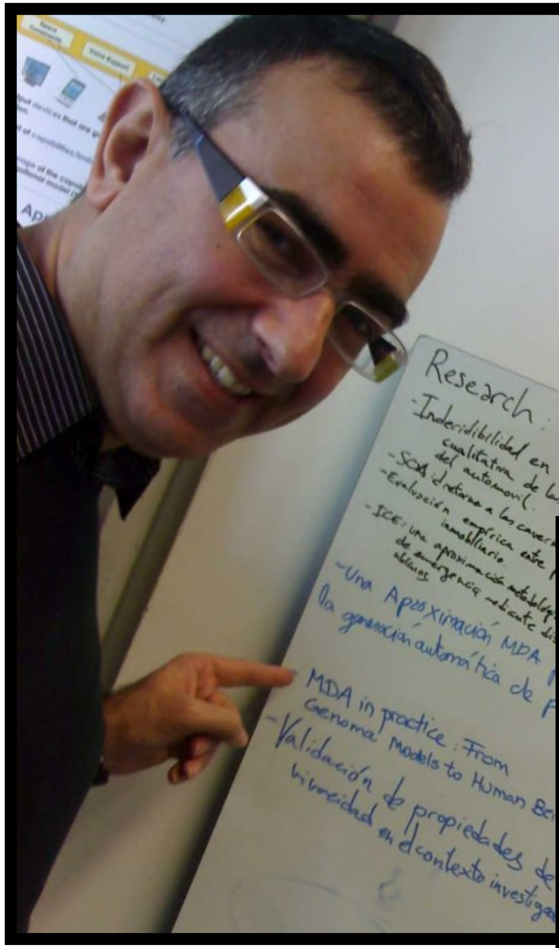


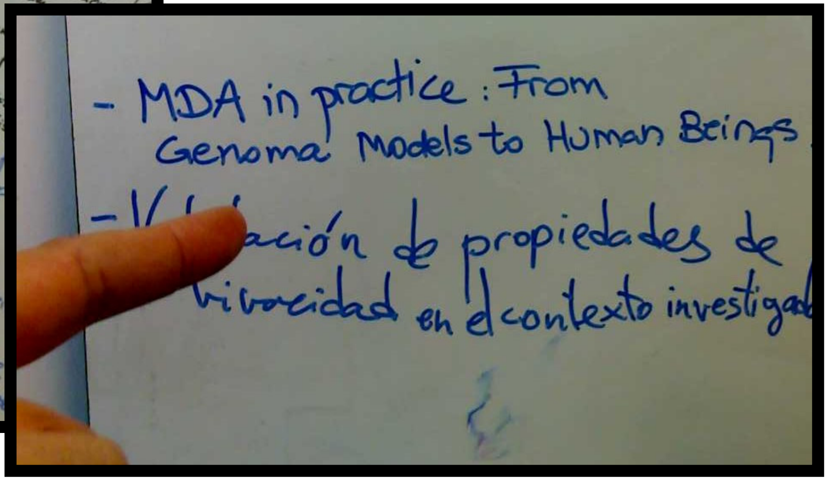
What is the most complex system you can imagine?

- Discussion started..


Research:
PACO
VICKY
BEA
PAU
GIO
30
- MDA in practice: From Genome models to Human Beings
- Validación de propiedades de hiwreidad en el contexto investigador
la generación automática de papers

We found it

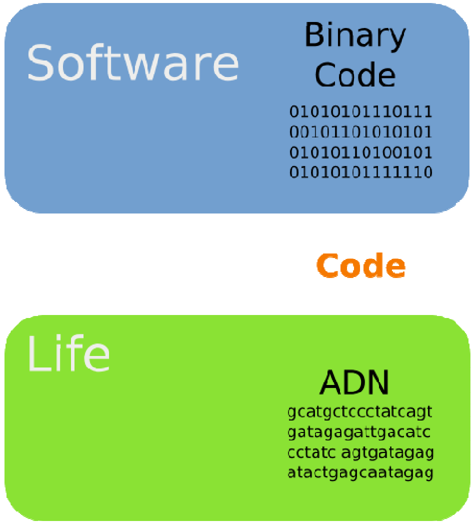

Research:
- Indecidibilidad en la evaluación cualitativa de los ... del automóvil
- SOA
- Evaluación empírica ... inmobiliario
- ICE: una aproximación ... de emergencia ...
- Una Aproximación MDA ... la generación automática de ...
- MDA in practice: From Genoma Models to Human Beings
- Validación de propiedades de vivacidad en el contexto investiga...


- MDA in practice: From Genoma Models to Human Beings
- V... ación de propiedades de vivacidad en el contexto investigad...

- Maybe, the answer is not so far from you…
- …it is **you**!!

A parallelism

- “A living organism is a *computer* or *machine* made up of genetic *circuits* in which DNA is the *software* that can be *hacked*." *— Drew Endy, MIT*



Software

Binary Code

01010101110111
00101101010101
01010110100101
01010101111110

Code

Life

ADN

gcatgctccctatcagt
gatagagattgacatc
cctatc agtgatagag
atactgagcaatagag

Building life

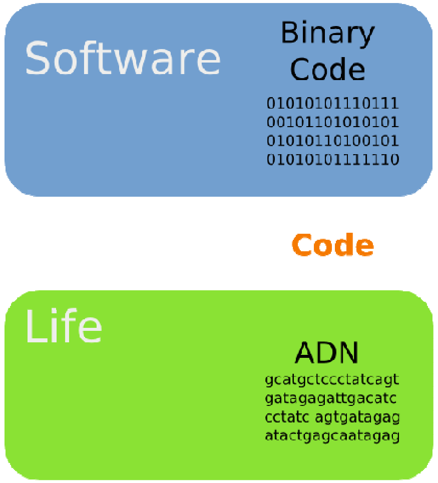
- Synthetic Biology can create new forms of life from scratch
  - A microbe that would help in fuel production
  - Biological films as a basis of new forms of lithography for assembling circuits
  - Cell division counters to prevent cancer
  - Re-designed seeds that the tree is programmed to grow into a house

…but, how is this *"software"* developed?

Building life

- “Using a laptop computer, published gene sequence information and mail-order synthetic DNA, just about anyone has the potential to construct genes or entire genomes from scratch." *— Drew Endy, MIT*

9
1-4
5-8
38
20-21
19
25-26
32
22-23
24
10-16
34
18
33
27
35
28-31
17
36-37
54
53
56-57
39
40
50-51
46-47
41-45
48
49
52
58
63
65
66
59
60
64
62
55
61


Software

Binary
Code

01010101110111
00101101010101
01010110100101
01010101111110

Code

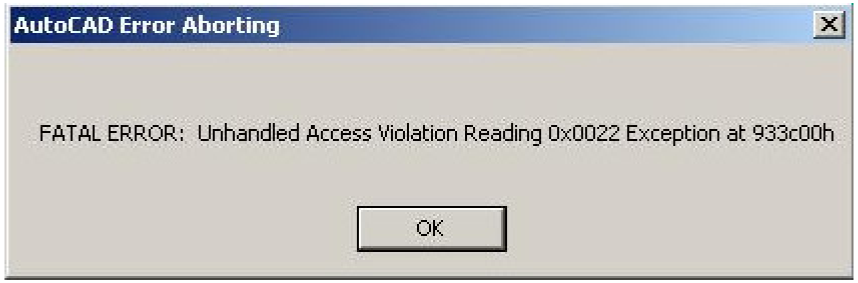
Life

ADN

gcatgctccctatcagt
gatagagattgacatc
cctatc agtgatagag
atactgagcaatagag

What about Software Quality?

- Handcraft development is error prone
  - …dangerous when talking about computers


AutoCAD Error Aborting

FATAL ERROR: Unhandled Access Violation Reading 0x0022 Exception at 933c00h

What about Software Quality?

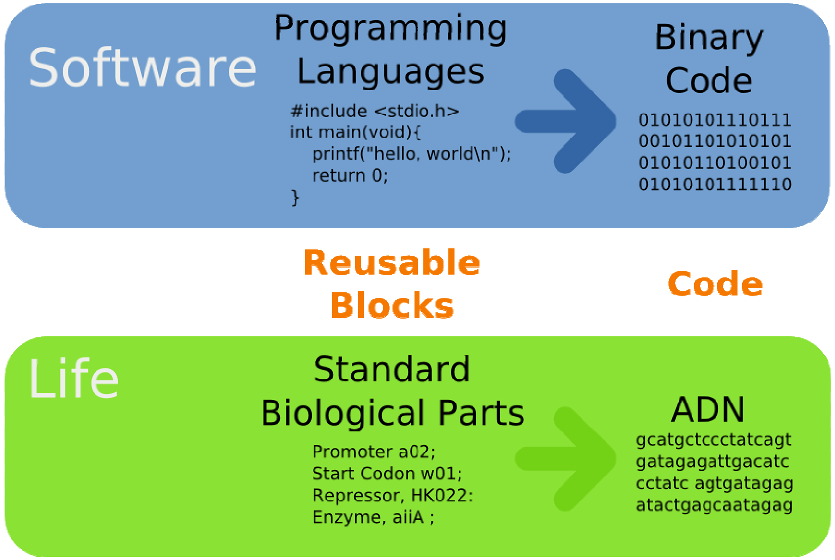
- Handcraft development is error prone
  - …lethal when dealing with life.

Abstraction as a solution

- Model Driven Development permits
  - Reason about the system prior to its construction
    - You can simulate the behavior to foresee the consequences of a system
  - Derivate the final system in an automatic way
    - Obtaining a consistent result

First step: Assembling

- First abstraction step
  - Standard Biological Parts



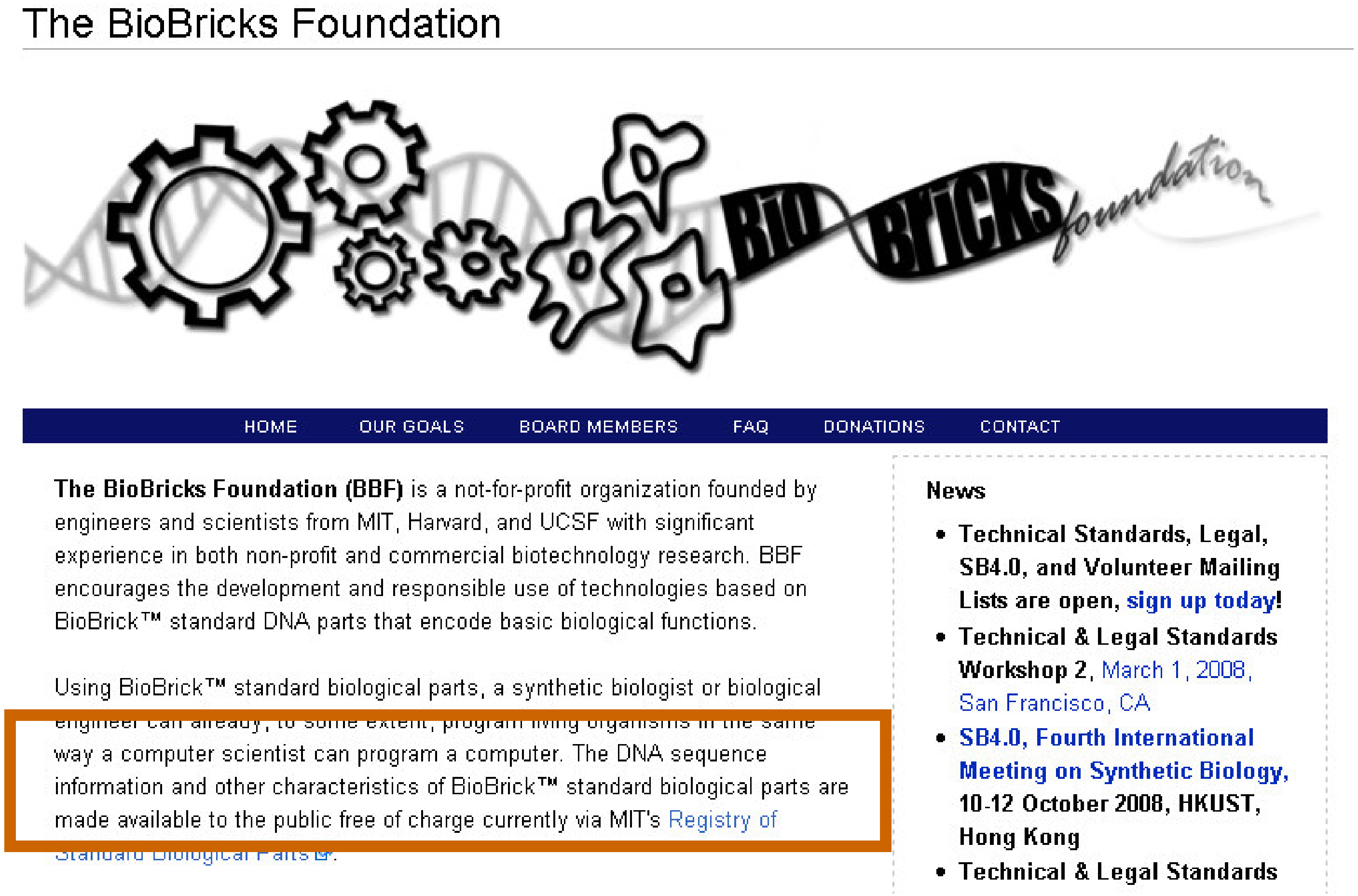
```
#include <stdio.h>
int main(void){
    printf("hello, world\n");
    return 0;
}
```

```
01010101110111
00101101010101
01010110100101
01010101111110
```

Promoter a02;
Start Codon w01;
Repressor, HK022:
Enzyme, aiiA ;

gcatgctccctatcagt
gatagagattgacatc
cctatc agtgatagag
atactgagcaatagag

BioBricks



The BioBricks Foundation

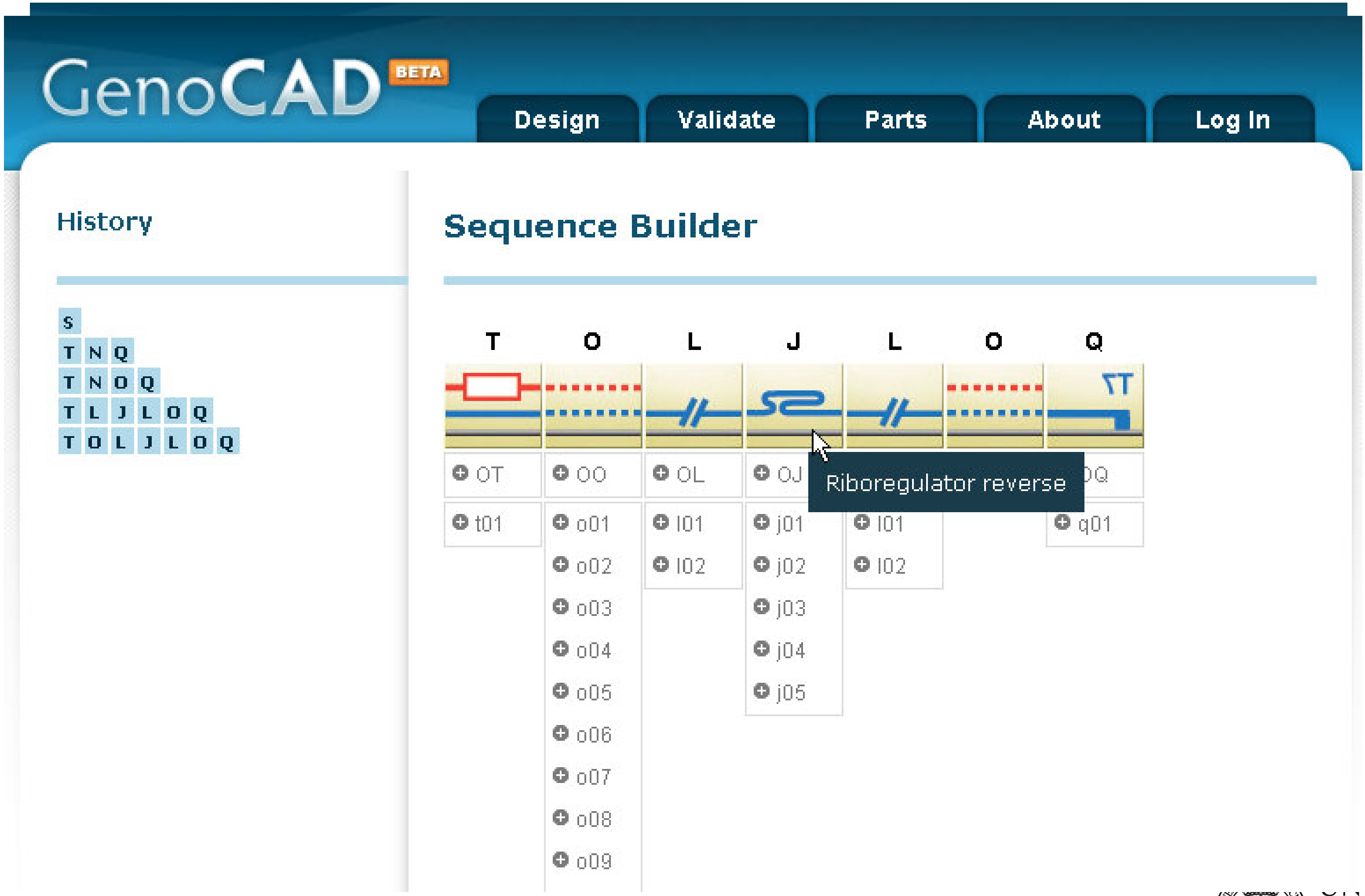
The BioBricks Foundation (BBF) is a not-for-profit organization founded by engineers and scientists from MIT, Harvard, and UCSF with significant experience in both non-profit and commercial biotechnology research. BBF encourages the development and responsible use of technologies based on BioBrick™ standard DNA parts that encode basic biological functions.

Using BioBrick™ standard biological parts, a synthetic biologist or biological engineer can already, to some extent, program living organisms in the same way a computer scientist can program a computer. The DNA sequence information and other characteristics of BioBrick™ standard biological parts are made available to the public free of charge currently via MIT's Registry of Standard Biological Parts.

News

- **Technical Standards, Legal, SB4.0, and Volunteer Mailing Lists are open, sign up today!**
- **Technical & Legal Standards Workshop 2**, March 1, 2008, San Francisco, CA
- **SB4.0, Fourth International Meeting on Synthetic Biology, 10-12 October 2008, HKUST, Hong Kong**
- **Technical & Legal Standards**

Electronics industry metaphors



History

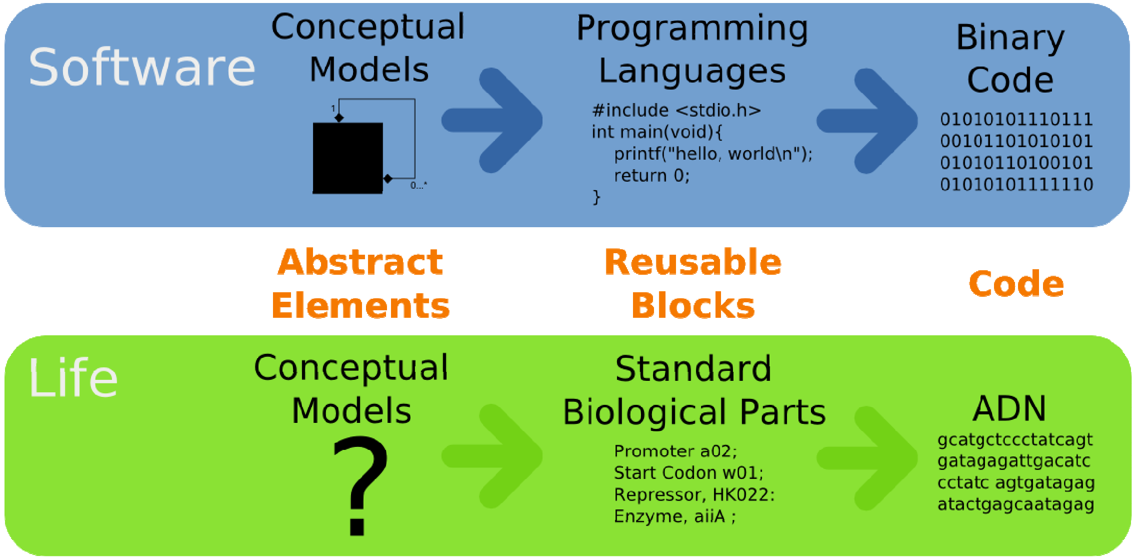
S
T N Q
T N O Q
T L J L O Q
T O L J L O Q

Sequence Builder

T	O	L	J	L	O	Q
OT	OO	OL	OJ			OQ
t01	o01	l01	j01	l01		q01
	o02	l02	j02	l02		
	o03		j03			
	o04		j04			
	o05		j05			
	o06					
	o07					
	o08					
	o09					

One step further: Modeling

- Conceptual models are needed for a systematic development of biological systems



```
#include <stdio.h>
int main(void){
    printf("hello, world\n");
    return 0;
}
```

```
01010101110111
00101101010101
01010110100101
01010101111110
```

```
Promoter a02;
Start Codon w01;
Repressor, HK022:
Enzyme, aiiA ;
```

```
gcatgctccctatcagt
gatagagattgacatc
cctatc agtgatagag
atactgagcaatagag
```

From Genome To Reality

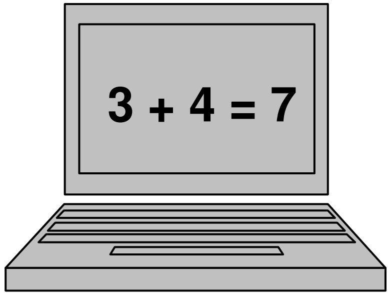
00010011	00000111	00000011	00001000

Physical Level

ADD $7 $3 $8

Instruction Level

Semantics: Add the values from the processor registers '3' and store the result in the register '8'



Representation Level

From Genome To Reality

AUG	GAA	CAC	GAC	GAG	UAA

Physical Level

START Glu His Asp Glu STOP

Semantics: Process a protein with the four selected aminoacids

Instruction Level

However, ¿Why?

Representation Level

One step further: Modeling

- Modeling benefits are needed for biological systems
  - Work at a higher abstraction level
    - Systems easy to specify
  - Reason about the system prior to construction
    - Foresee consequences in advance
    - Simulate, validate, etc.
  - Automate the development
    - In a systematic way

Conclusion

- With Conceptual Models targeted at digital elements, we can improve Information Systems Development
- With Conceptual Models targeted at life we can directly improve our living

Translational Research

- Movement of discoveries in basic research (the Bench) to application at the clinical level (the Bedside)
- A significant barrier: the lack of uniformly structured data across related biomedical domains
- A potential solution: Semantic Web Technologies

- Information ecosystem
  - Scientific literature
  - Experimental data
  - Summaries of knowledge of gene products
  - Diseases
  - Compounds
  - Informal scientific discourse and commentary in a variety of forums
- This data has been provided in numerous disconnected DBs –data silos-

- The lack of uniformly structured data affects many areas of biomedical research
  - Drug discovery
  - Systems biology
  - Individualized medicine
- …all of which rely heavily on integrating and interpreting data sets produced by different experimental methods at different levels of granularity

Example: Alzheimer's Disease (AD)

- Still no agreement on how it is caused, or where best to intervene to treat it or prevent it
- Recent hypothesis combines data from research in mouse genetics, cell biology, animal neuropsychology, protein biochemistry, neuropathology,… and other areas

Example: Huntington’s Disease (HD)

- Relatively simple genetic basis, and a model for autosomal dominant neurogenetic disorders proposed …

- But the mechanisms by which the disorder causes pathology still not understood, what creates profound difficulties with existing treatments.

How can the SW help biomedical research?

- Are Semantic Web Technologies the solution?
  - Thesauri, ontologies, rule systems, frame based representation systems,..
  - A query language (SPARQL)
  - RDF, OWL,…

Some potential advantages

- Global scope of identifiers
- RDFS and OWL are
  - Self-descriptive languages
  - Flexible, extendable and decentralized
- Ability to do inference, classification and consistency checking
  - A review of GO gave up to 10% of obsolete terms for gene annotations

Main objectives

- Identification of core vocabularies and ontologies to support effective access to knowledge and data
- Development of guidelines and best practices for unambiguously identifying resources such as docs and biological entities
- Development of strategies for linking to the information discussed in scientific pubs. from within those pubs.

The present…

- Applied Biosystems expects that the public availability of the human sequence data will help drive innovation and speed the development of new bioinformatics tools. These new tools are expected to enable researchers to interpret the meaning of the data that provide clues to better understand various aspects of health and disease.

- “To understand the extent and prevalence of structural variation in the human genome, which is still largely unknown, traditional sequencing methods are applied with good results, but much more needs to be discovered at a faster pace. The human paired-end data being released is of such depth that discovering smaller structural events at higher resolution becomes possible. The availability of this dataset in the public domain will accelerate our understanding of structural variation in normal and disease states, and open the door to a faster exploration of this type of genetic diversity across human populations.”

The caos of the genome data

- Currently there are **tons of data** from the genome publicly available
- Some of these databases are **free available** on the Web because owners doesn't know how to find relevant information
- Each database is defined with an specific schema, data format, identifications, etc.
- The **integration** of the different sources is a very difficult task

Example: Looking for information about the NF1 Gene

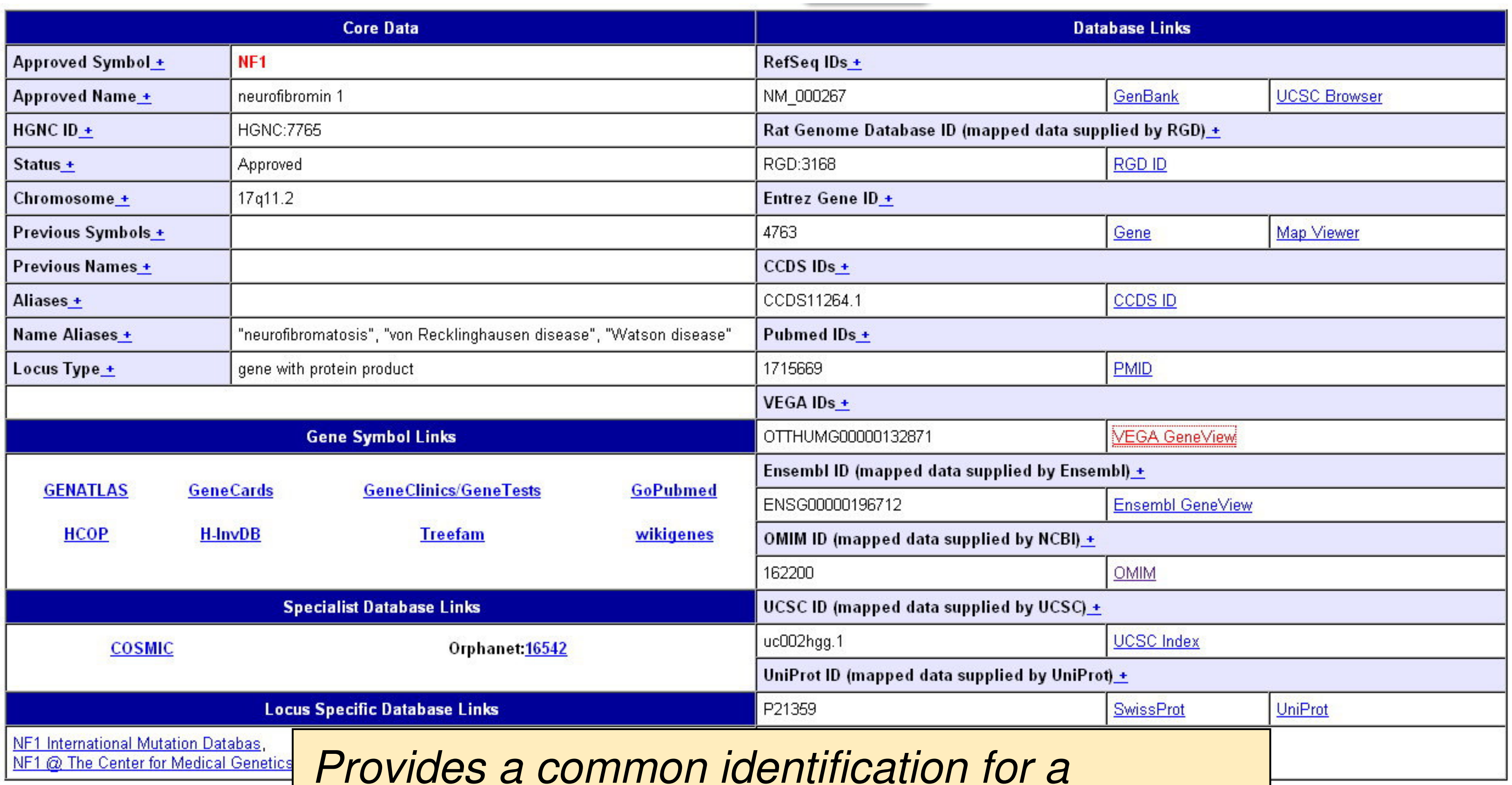
- A genomic laboratory must perform an analysis to determine in the subject suffers from Neurofibromatosis

- Currently the genetic analyst must manually search in the different databases to elaborate the report

- As a first research exercise, we have been looking for information about the NF1 Gene that provokes the Neurofibromatosis disease

- Several databases have been consulted to understand how the data is stored and retrieved

HUGO



Core Data	
Approved Symbol +	NF1
Approved Name +	neurofibromin 1
HGNC ID +	HGNC:7765
Status +	Approved
Chromosome +	17q11.2
Previous Symbols +	
Previous Names +	
Aliases +	
Name Aliases +	"neurofibromatosis", "von Recklinghausen disease", "Watson disease"
Locus Type +	gene with protein product

Gene Symbol Links

GENATLAS	GeneCards	GeneClinics/GeneTests	GoPubmed
HCOP	H-InvDB	Treefam	wikigenes

Specialist Database Links

COSMIC Orphanet:16542

Locus Specific Database Links

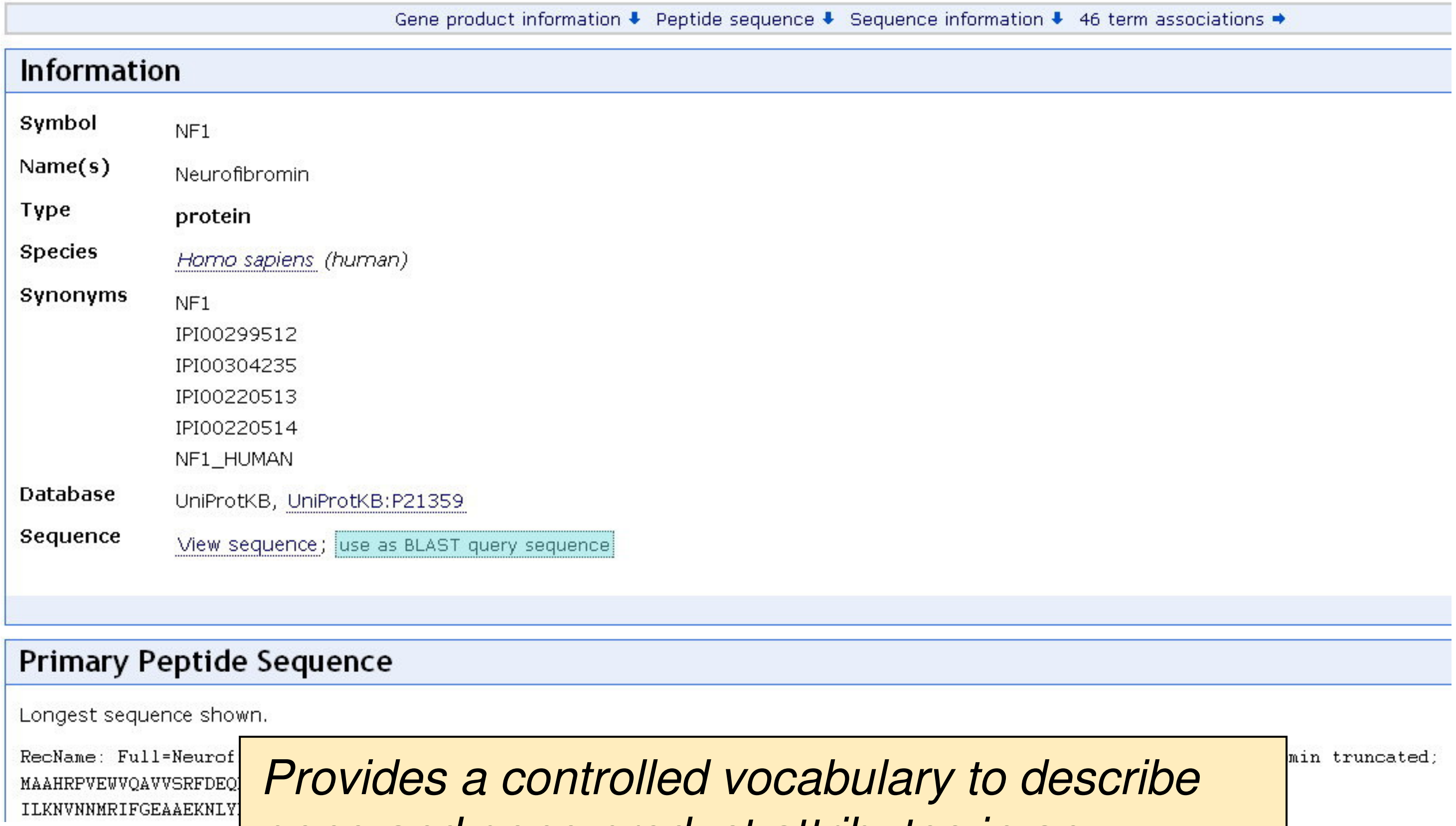
NF1 International Mutation Databas,
NF1 @ The Center for Medical Genetics

Database Links		
RefSeq IDs +		
NM_000267	GenBank	UCSC Browser
Rat Genome Database ID (mapped data supplied by RGD) +		
RGD:3168	RGD ID	
Entrez Gene ID +		
4763	Gene	Map Viewer
CCDS IDs +		
CCDS11264.1	CCDS ID	
Pubmed IDs +		
1715669	PMID	
VEGA IDs +		
OTTHUMG00000132871	VEGA GeneView	
Ensembl ID (mapped data supplied by Ensembl) +		
ENSG00000196712	Ensembl GeneView	
OMIM ID (mapped data supplied by NCBI) +		
162200	OMIM	
UCSC ID (mapped data supplied by UCSC) +		
uc002hgg.1	UCSC Index	
UniProt ID (mapped data supplied by UniProt) +		
P21359	SwissProt	UniProt

Provides a common identification for a particular gene and the different alias used in another databases

Gene Ontology

NF1



Gene product information ⬇ Peptide sequence ⬇ Sequence information ⬇ 46 term associations ➡

Information

Symbol	NF1
Name(s)	Neurofibromin
Type	**protein**
Species	*Homo sapiens* (human)
Synonyms	NF1 IPI00299512 IPI00304235 IPI00220513 IPI00220514 NF1_HUMAN
Database	UniProtKB, UniProtKB:P21359
Sequence	View sequence; use as BLAST query sequence

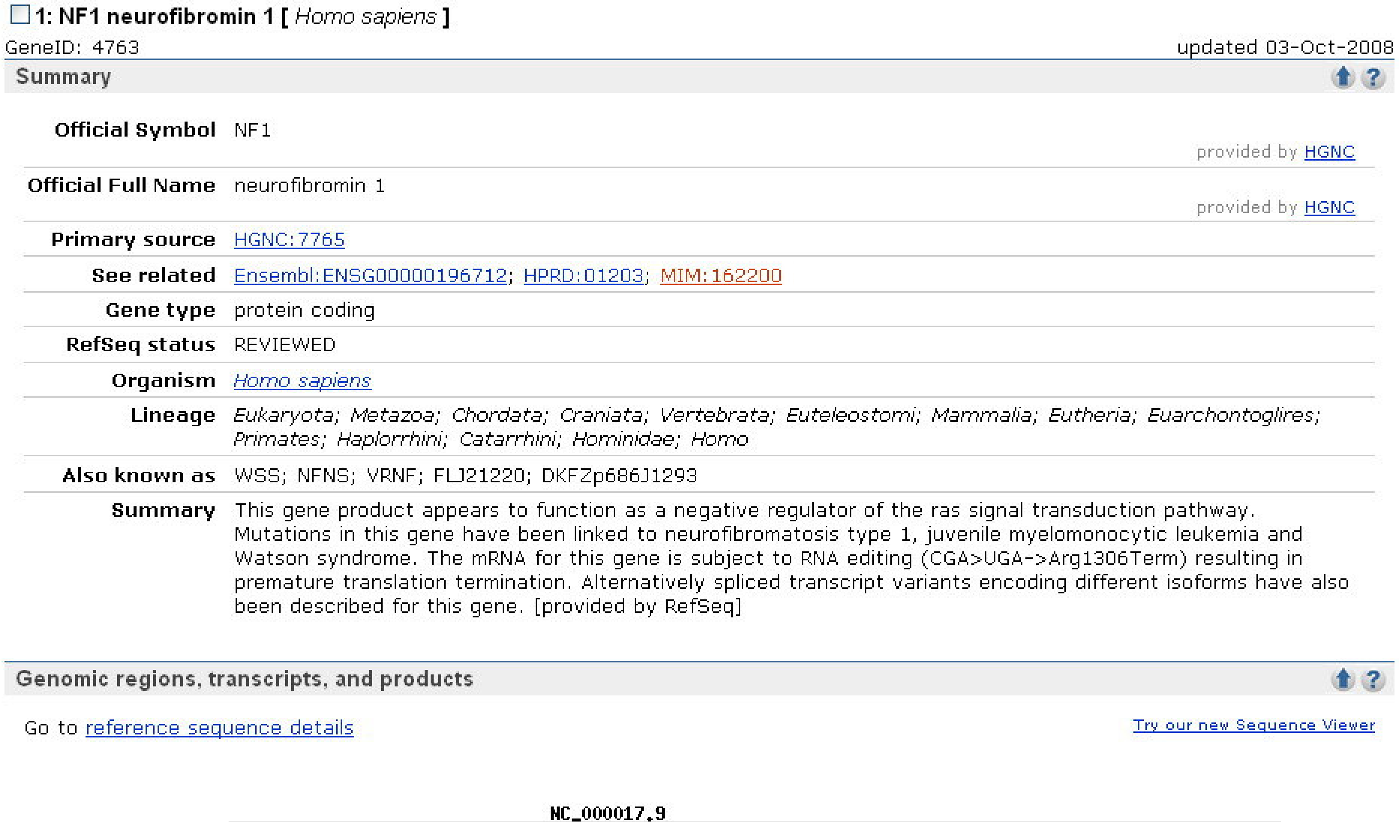
Primary Peptide Sequence

Longest sequence shown.

```
RecName: Full=Neurof                                                                              min truncated;
MAAHRPVEWVQAVVSRFDEQ
ILKNVNNMRIFGEAAEKNLY
```

Provides a controlled vocabulary to describe gene and gene product attributes in any organism. Useful to find relationships with a particular genomic term

Entrez GENE



□1: **NF1 neurofibromin 1 [*Homo sapiens*]**

GeneID: 4763 updated 03-Oct-2008

Summary

Official Symbol	NF1 provided by HGNC
Official Full Name	neurofibromin 1 provided by HGNC
Primary source	HGNC:7765
See related	Ensembl:ENSG00000196712; HPRD:01203; MIM:162200
Gene type	protein coding
RefSeq status	REVIEWED
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as	WSS; NFNS; VRNF; FLJ21220; DKFZp686J1293
Summary	This gene product appears to function as a negative regulator of the ras signal transduction pathway. Mutations in this gene have been linked to neurofibromatosis type 1, juvenile myelomonocytic leukemia and Watson syndrome. The mRNA for this gene is subject to RNA editing (CGA>UGA->Arg1306Term) resulting in premature translation termination. Alternatively spliced transcript variants encoding different isoforms have also been described for this gene. [provided by RefSeq]

Genomic regions, transcripts, and products

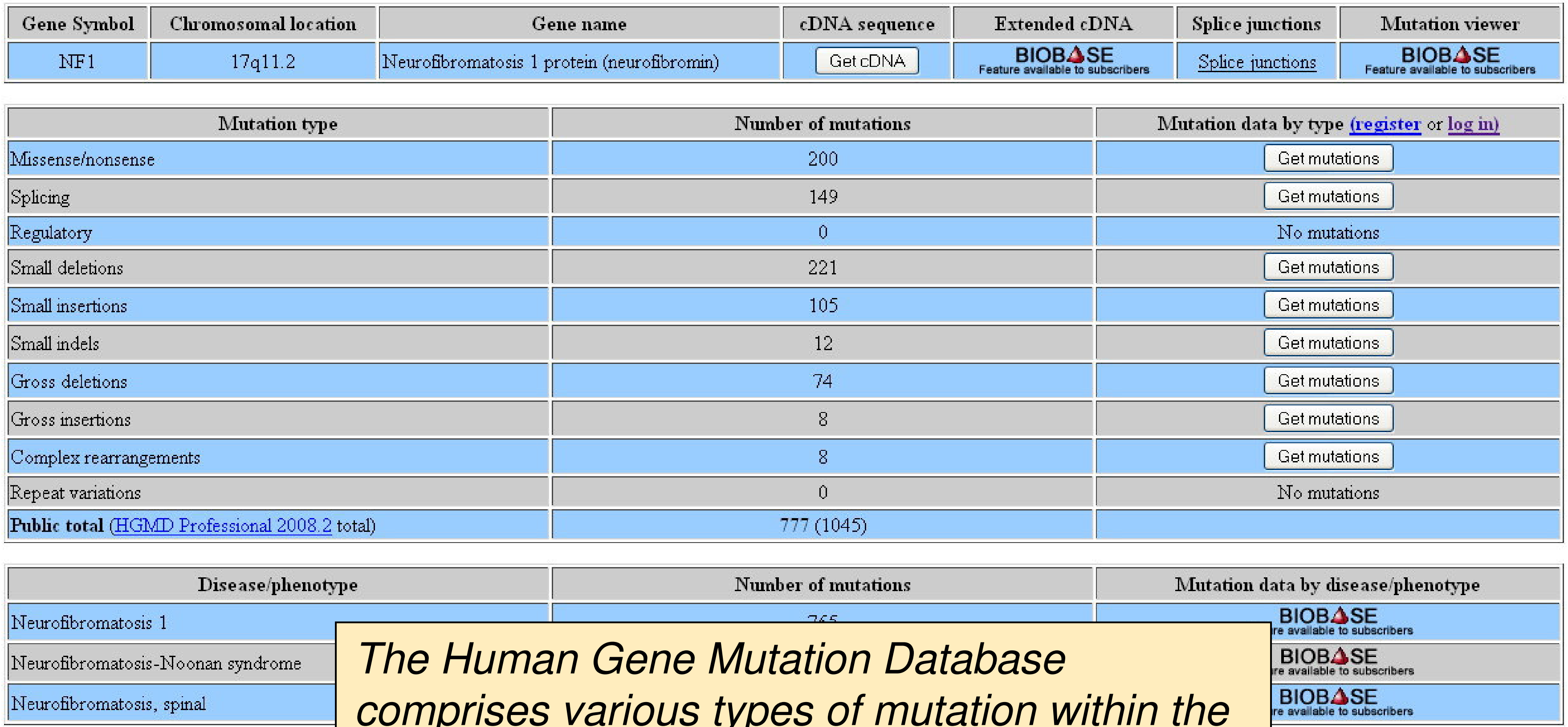
Go to reference sequence details

Try our new Sequence Viewer

NC_000017.9

Entrez Gene provides a unified query environment for *genes* provided by the NCBI. It can be considered ad the “facto” standard database to find information about a gene

HGMD



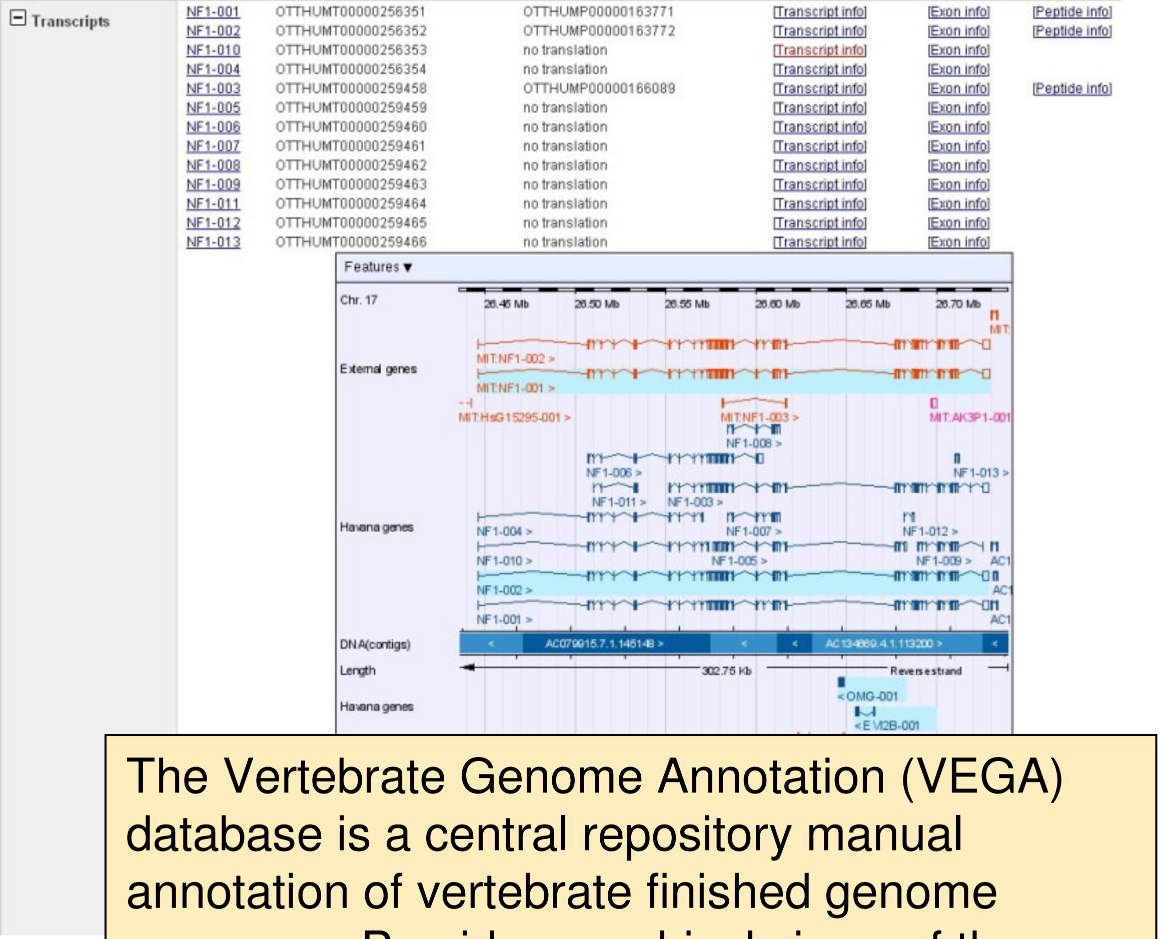
Gene Symbol	Chromosomal location	Gene name	cDNA sequence	Extended cDNA	Splice junctions	Mutation viewer
NF1	17q11.2	Neurofibromatosis 1 protein (neurofibromin)	Get cDNA	BIOBASE Feature available to subscribers	Splice junctions	BIOBASE Feature available to subscribers

Mutation type	Number of mutations	Mutation data by type (register or log in)
Missense/nonsense	200	Get mutations
Splicing	149	Get mutations
Regulatory	0	No mutations
Small deletions	221	Get mutations
Small insertions	105	Get mutations
Small indels	12	Get mutations
Gross deletions	74	Get mutations
Gross insertions	8	Get mutations
Complex rearrangements	8	Get mutations
Repeat variations	0	No mutations
Public total (HGMD Professional 2008.2 total)	777 (1045)	

Disease/phenotype	Number of mutations	Mutation data by disease/phenotype
Neurofibromatosis 1	765	BIOBASE re available to subscribers
Neurofibromatosis-Noonan syndrome		BIOBASE re available to subscribers
Neurofibromatosis, spinal		BIOBASE re available to subscribers

The Human Gene Mutation Database comprises various types of mutation within the coding regions, splicing and regulatory regions of human nuclear genes causing inherited disease

VEGA



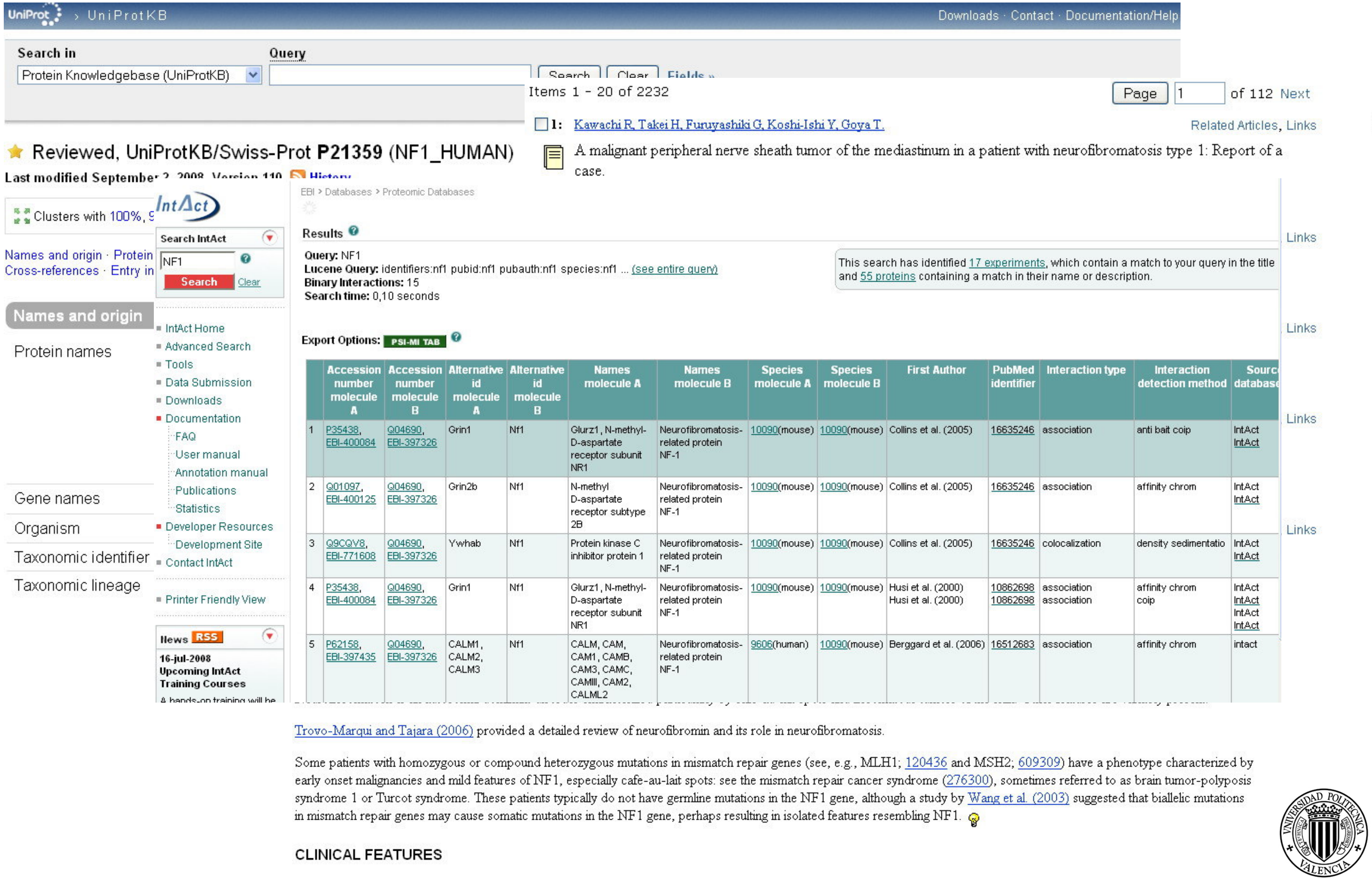
⊟ Transcripts

NF1-001	OTTHUMT00000256351	OTTHUMP00000163771	[Transcript info]	[Exon info]	[Peptide info]
NF1-002	OTTHUMT00000256352	OTTHUMP00000163772	[Transcript info]	[Exon info]	[Peptide info]
NF1-010	OTTHUMT00000256353	no translation	[Transcript info]	[Exon info]	
NF1-004	OTTHUMT00000256354	no translation	[Transcript info]	[Exon info]	
NF1-003	OTTHUMT00000259458	OTTHUMP00000166089	[Transcript info]	[Exon info]	[Peptide info]
NF1-005	OTTHUMT00000259459	no translation	[Transcript info]	[Exon info]	
NF1-006	OTTHUMT00000259460	no translation	[Transcript info]	[Exon info]	
NF1-007	OTTHUMT00000259461	no translation	[Transcript info]	[Exon info]	
NF1-008	OTTHUMT00000259462	no translation	[Transcript info]	[Exon info]	
NF1-009	OTTHUMT00000259463	no translation	[Transcript info]	[Exon info]	
NF1-011	OTTHUMT00000259464	no translation	[Transcript info]	[Exon info]	
NF1-012	OTTHUMT00000259465	no translation	[Transcript info]	[Exon info]	
NF1-013	OTTHUMT00000259466	no translation	[Transcript info]	[Exon info]	

The Vertebrate Genome Annotation (VEGA) database is a central repository manual annotation of vertebrate finished genome sequence. Provides graphical views of the different gene transcripts

Centro de Investigación en Métodos de Producción de Software

And more…



UniProt › UniProtKB
Downloads · Contact · Documentation/Help
Search in
Query
Protein Knowledgebase (UniProtKB)
Search
Clear
Fields »
Page
1
of 112 Next

Items 1 - 20 of 2232

□ 1: Kawachi R, Takei H, Furuyashiki G, Koshi-Ishi Y, Goya T. Related Articles, Links

A malignant peripheral nerve sheath tumor of the mediastinum in a patient with neurofibromatosis type 1: Report of a case.

Reviewed, UniProtKB/Swiss-Prot **P21359** (NF1_HUMAN)

Last modified September 2, 2008. Version 110 History

Clusters with 100%, 9

Names and origin · Protein
Cross-references · Entry in

Names and origin

Protein names

Gene names

Organism

Taxonomic identifier

Taxonomic lineage

16-jul-2008
Upcoming IntAct Training Courses
A hands-on training will be

EBI › Databases › Proteomic Databases

Results

Query: NF1
Lucene Query: identifiers:nf1 pubid:nf1 pubauth:nf1 species:nf1 ... (see entire query)
Binary Interactions: 15
Search time: 0,10 seconds

This search has identified 17 experiments, which contain a match to your query in the title and 55 proteins containing a match in their name or description.

Export Options: PSI-MI TAB

	Accession number molecule A	Accession number molecule B	Alternative id molecule A	Alternative id molecule B	Names molecule A	Names molecule B	Species molecule A	Species molecule B	First Author	PubMed identifier	Interaction type	Interaction detection method	Sourc databas
1	P35438, EBI-400084	Q04690, EBI-397326	Grin1	Nf1	Glurz1, N-methyl-D-aspartate receptor subunit NR1	Neurofibromatosis-related protein NF-1	10090(mouse)	10090(mouse)	Collins et al. (2005)	16635246	association	anti bait coip	IntAct IntAct
2	Q01097, EBI-400125	Q04690, EBI-397326	Grin2b	Nf1	N-methyl D-aspartate receptor subtype 2B	Neurofibromatosis-related protein NF-1	10090(mouse)	10090(mouse)	Collins et al. (2005)	16635246	association	affinity chrom	IntAct IntAct
3	Q9CQV8, EBI-771608	Q04690, EBI-397326	Ywhab	Nf1	Protein kinase C inhibitor protein 1	Neurofibromatosis-related protein NF-1	10090(mouse)	10090(mouse)	Collins et al. (2005)	16635246	colocalization	density sedimentatio	IntAct IntAct
4	P35438, EBI-400084	Q04690, EBI-397326	Grin1	Nf1	Glurz1, N-methyl-D-aspartate receptor subunit NR1	Neurofibromatosis-related protein NF-1	10090(mouse)	10090(mouse)	Husi et al. (2000) Husi et al. (2000)	10862698 10862698	association association	affinity chrom coip	IntAct IntAct IntAct IntAct
5	P62158, EBI-397435	Q04690, EBI-397326	CALM1, CALM2, CALM3	Nf1	CALM, CAM, CAM1, CAMB, CAM3, CAMC, CAMIII, CAM2, CALML2	Neurofibromatosis-related protein NF-1	9606(human)	10090(mouse)	Berggard et al. (2006)	16512683	association	affinity chrom	intact

Links

Links

Links

Links

Trovo-Marqui and Tajara (2006) provided a detailed review of neurofibromin and its role in neurofibromatosis.

Some patients with homozygous or compound heterozygous mutations in mismatch repair genes (see, e.g., MLH1; 120436 and MSH2; 609309) have a phenotype characterized by early onset malignancies and mild features of NF1, especially cafe-au-lait spots: see the mismatch repair cancer syndrome (276300), sometimes referred to as brain tumor-polyposis syndrome 1 or Turcot syndrome. These patients typically do not have germline mutations in the NF1 gene, although a study by Wang et al. (2003) suggested that biallelic mutations in mismatch repair genes may cause somatic mutations in the NF1 gene, perhaps resulting in isolated features resembling NF1.

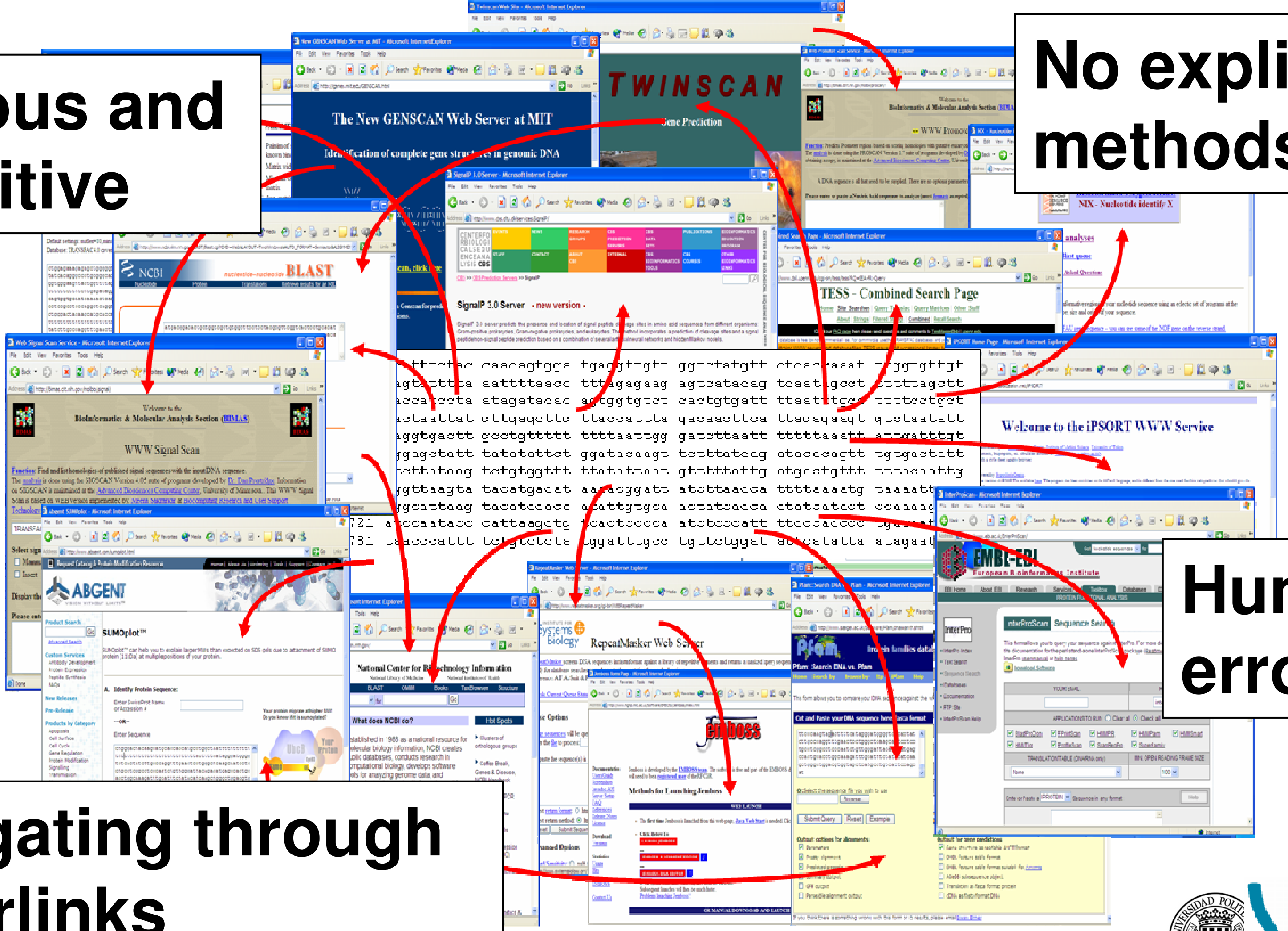
CLINICAL FEATURES

Recent Activity

- NF1 (2232 results)
- NF1 (2232 results)
- Methylglyoxal mediates p38 in human endothelia
- NF! (40868 results)
- NF! (40868 results)

UNIVERSIDAD POLITECNICA DE VALENCIA

Manual Methods of data analysis


The New GENSCAN Web Server at MIT
Identification of complete gene structures in genomic DNA
TWINSCAN
Gene Prediction
NCBI
BLAST
SignalP 3.0 Server
TESS - Combined Search Page
WWW Signal Scan
Bioinformatics & Molecular Analysis Section (BIMAS)
Welcome to the iPSORT WWW Service
ABGENT
SUMOplot™
National Center for Biotechnology Information
RepeatMasker Web Server
Pfam
EMBL-EBI
InterProScan
emboss

Tedious and repetitive

No explicit methods

Human error

Navigating through hyperlinks

Drawbacks observed

- Different identifications (ids) for the same disease gene
- The data is available on the Web but databases cannot always be directly queried
- The position (locus) of a particular gene depends on the genome sequenced
- Data is changing continuously
- High amount of information not well structured
- To provide a quality report about a gene disease several databases not interconnected must be manually consulted

The short-term future

- The problem is getting worse !!!!!
- The DNA Sequencing hardware is evolving dramatically
- In next years, we will be able to sequence a complete human genome faster and cheaper

What Would You Do With a $10K Genome?
Sequence the Whole Genome
View the Whole Transcriptome
Unlock the Cancer Genome
Resequence All My Biologically Relevant Genes
Discover Novel Transcripts and RNAs

The short-term future

- However, currently there is no software available to deal with the new challenges

- Software is required to:
  - Automatically find the mutations from a sequenced sample and store the new ones detected
  - Compare the genome of different subjects in order to determine all the differences between them
  - Trace the pathway from the genome code to the final phenotype of the individuals

- Conceptual modeling is required to produce quality software in this emerging domain

Our Solution: Conceptual Modelling

- **Main goal:** provide Conceptual Models to represent the genome in order to enhance the Model-driven development of Biogenetic software
- The gene ontology is a useful resource to define a taxonomy but not to guide the software implementation
- The first step is to provide a common **E-R model** that will be able to support the genomic data complexity
- First approaches has been proposed by N.W. Paton et. Al[1], S.Ram [2], C.Tao and D.Embley [3]

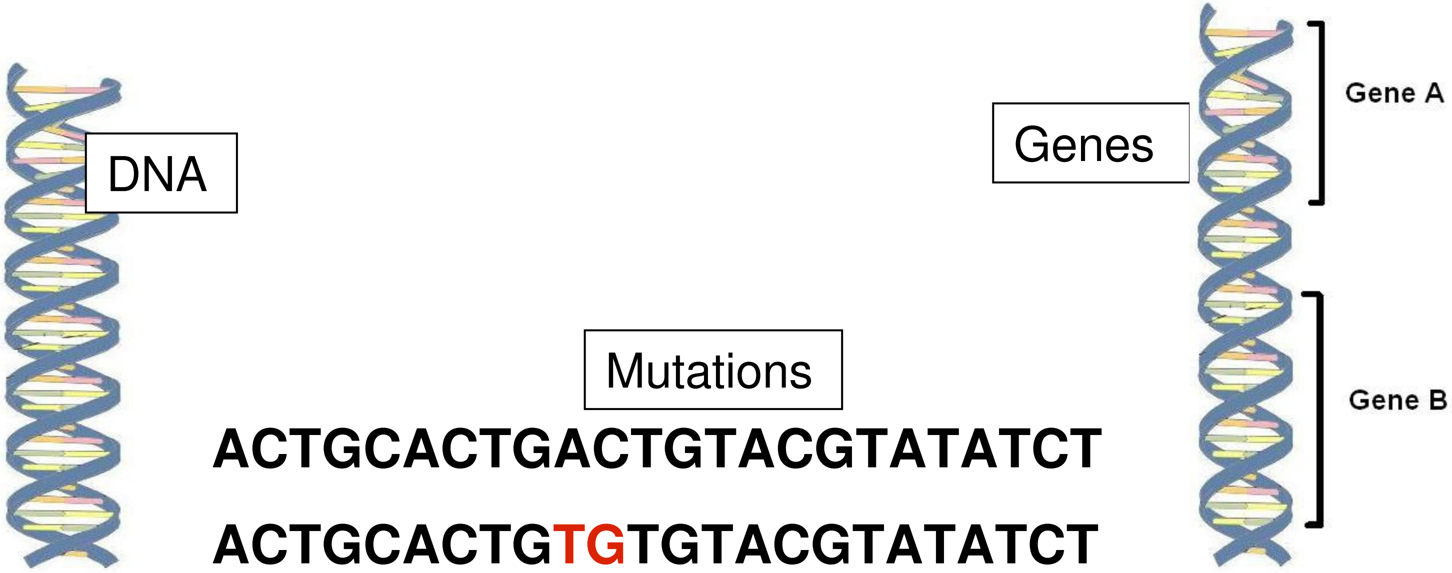
[1] *N. W. Paton, S. A. Khan, A. Hayes, F. Moussouni, A. Brass, K. Eilbeck, C. A. Goble, S. J. Hubbard, and S. G. Oliver, "Conceptual modelling of Genomic Information," Bioinformatics, vol. 16, pp. 548-557, 2000.*

[2] Ram,S.: Toward Semantic Interoperability of Heterogeneous Biological Data Sources.CAiSE 2005 : 32-32

[3] *Tao,C.; Embley,D.: Seed-Based Generation of Personalized Bio-ontologies for Information Extraction. ER Workshops 2007: 74-84*

Genotype

The entire genetic identity of an individual that **does not show** any outward characteristics, *e.g.* Genes, mutations



DNA

Genes

Mutations

ACTGCACTGACTGTACGTATATCT

ACTGCACTGTGTGTACGTATATCT

Source: Paul Fisher -UMIST

Phenotype

(harder to characterise)

The observable expression of gene’s producing **notable characteristics** in an individual, *e.g.* Hair or eye colour, body mass, resistance to disease

Brown

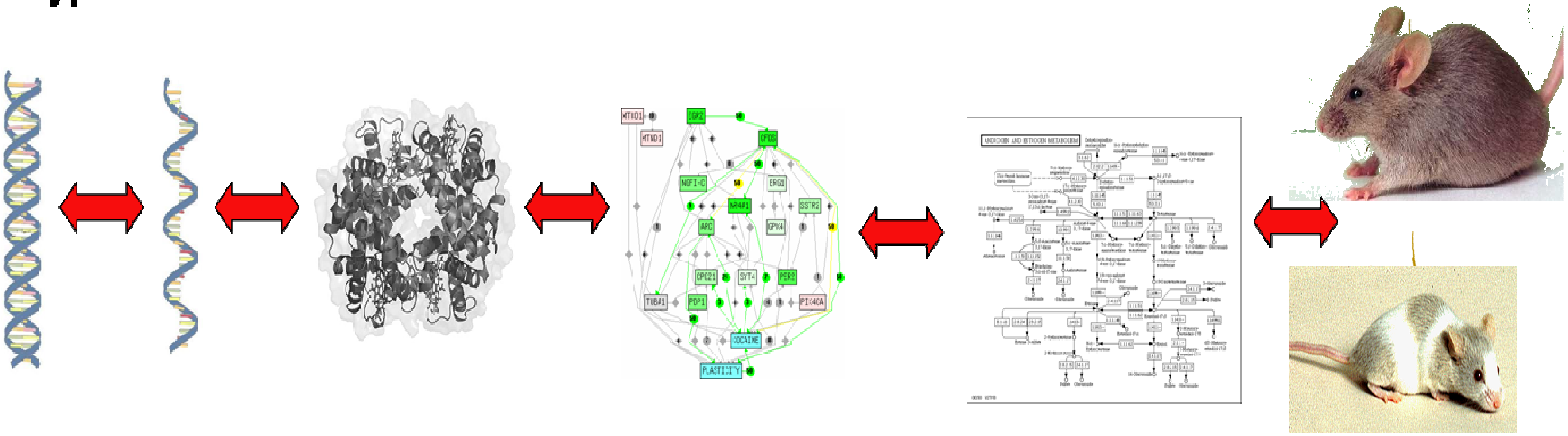
vs.

White and Brown

Source: Paul Fisher -UMIST

Genotype

Phenotype



DNA

RNA

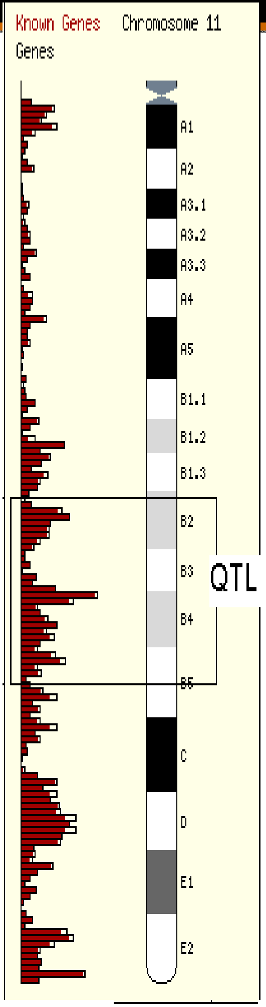
Protein

Protein-Protein interaction

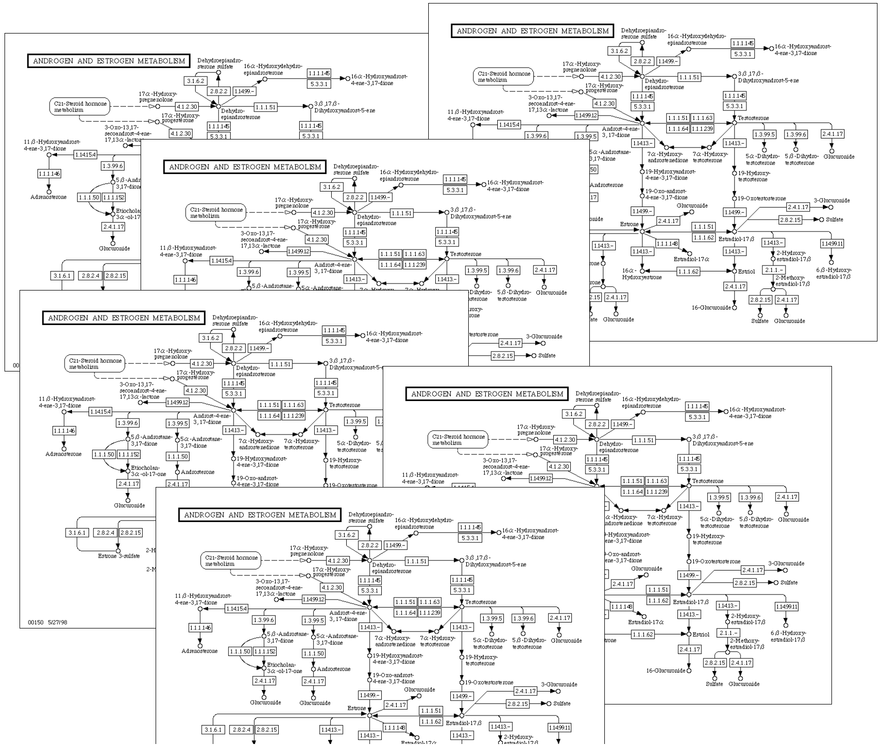
Pathway

Trait

Source: Paul Fisher -UMIST



200

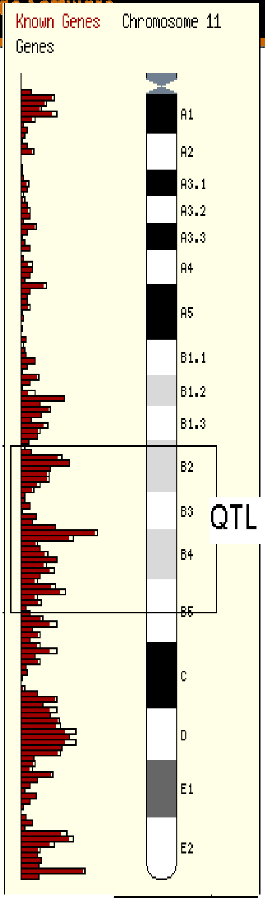


?

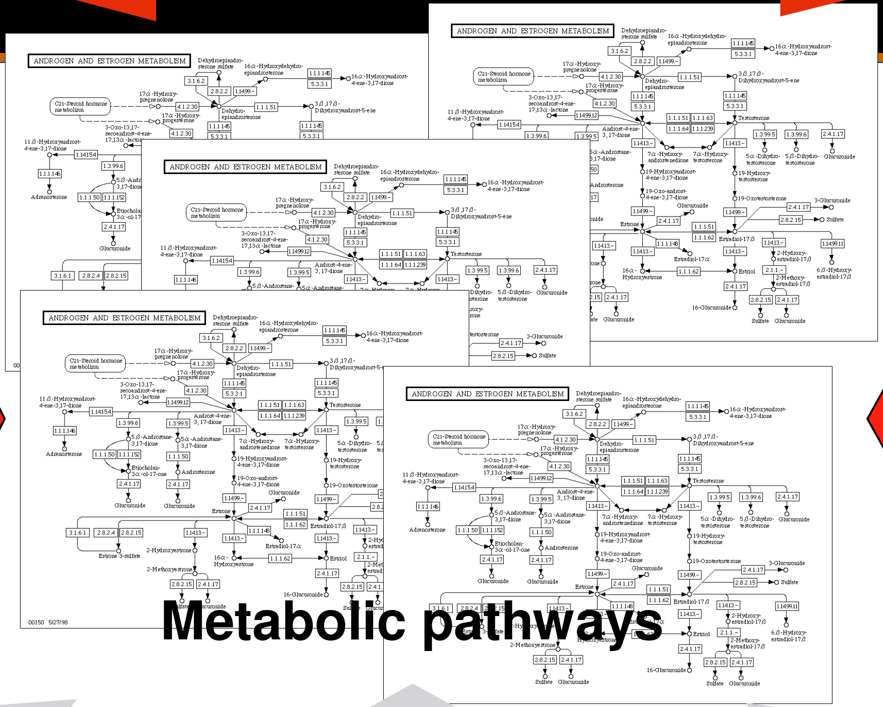
What processes to investigate?

Source: Paul Fisher -UMIST

ProS Centro de Investigación en Métodos de Producción de Software

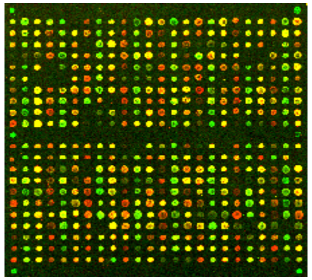


200



Metabolic pathways

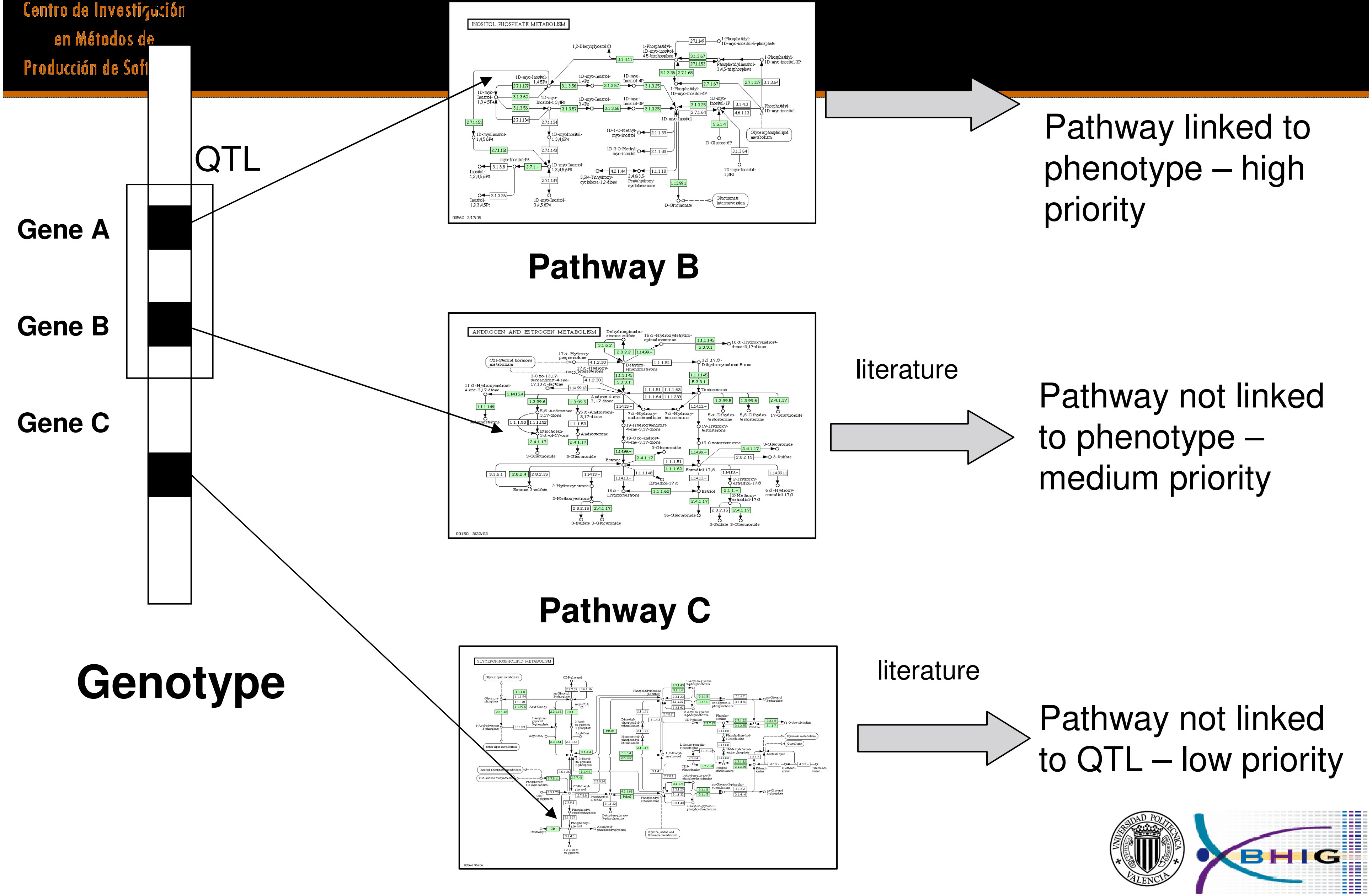
?



Microarray + QTL

Genes captured in microarray experiment and present in QTL (**Quantitative Trait Loci**) region

Phenotypic response investigated using microarray in form of expressed genes or evidence provided through QTL mapping

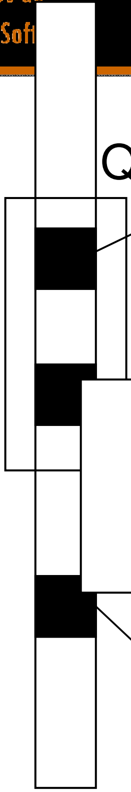

QTL
Gene A
Gene B
Gene C
INOSITOL PHOSPHATE METABOLISM
Pathway B
ANDROGEN AND ESTROGEN METABOLISM
Pathway C
GLYCEROPHOSPHOLIPID METABOLISM
literature
literature

Pathway linked to phenotype – high priority

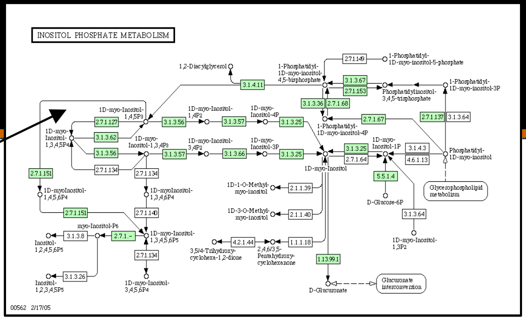
Pathway not linked to phenotype – medium priority

Pathway not linked to QTL – low priority

Genotype



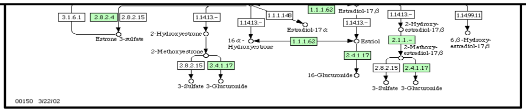
Genotype



Pathway B

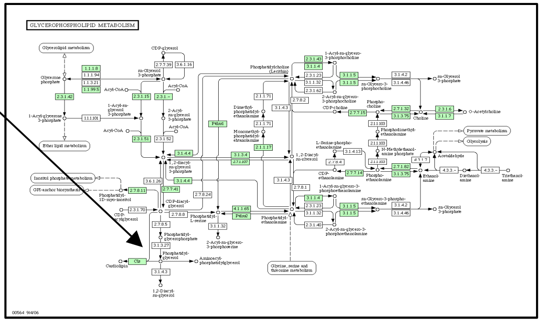
Pathway linked to phenotype – high priority

DONE MANUALLY



…ked
medium priority

Pathway C

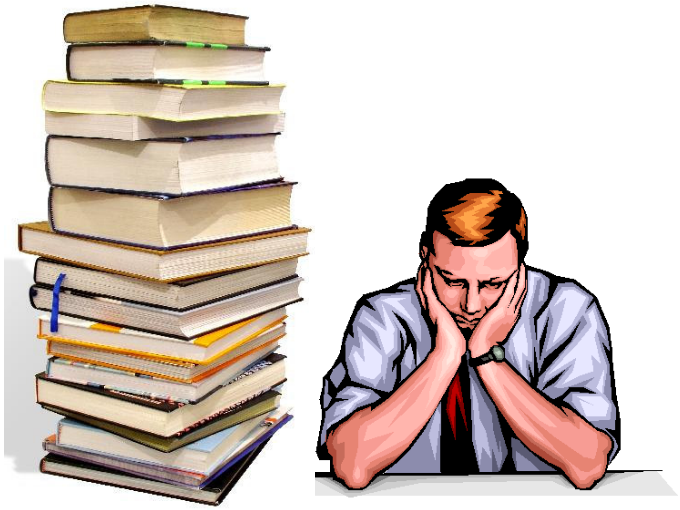


literature

Pathway not linked to QTL – low priority

It can’t be that hard, right?

- PubMed contains ~17,787,763 journals to date
- Manually searching is tedious and frustrating
- Can be hard finding the links



Source: Paul Fisher -UMIST

Computers can help with data gathering and information extraction – that’s their job !!!

Understanding the Domain (the Problem Space)

- Life as we know it is specified by the Genomes of the myriad organisms with which we share the planet.
- The nuclear genome comprises 3,2 G nucleotides of DNA, divided into 24 linear mollecules, the shortest 50M nucleotides, the longest 260M, each contained in a different chromosome.
- These 24 chromosomes consist of 22 autosomes and the two sex chromosomes, X and Y
- Some 35.000 genes are present in the human nuclear genome.

Understanding the Domain (the Problem Space)

	Size Mb	Num genes	RefSeq RNA	ESTs
Oryctolagus cuniculus (rabbit)	3500	20.000	----	32.000
Homo sapiens (human)	3000	35.000	40.000	8.100.000
Macaca mulatta (monkay)	3000	28.000	43.000	58.000
Pan troglodytes (chimpanzee)	3000	25.000	57.000	16.000
Bos taurus (cow)	3000	25.000	28.000	1.300.000
Felis catus (cat)	3000	18.000	317	186.000
Rattus novergicus (rat)	2800	29.000	37.000	812.000
Sus scrofa (pig)	2800	--	1.423	1.300.000
Canis familiaris (dog)	2400	24.000	33.000	365.000
Mus musculus (mouse)	2500	29.000	40.000	4.745.000
Danio rerio (pez zebra)	1700	25.000	37.000	1.345.000
Xenopus tropicalis (frog)	1700	19.000	27.000	1.112.000
Gallus gallus (cockerel)	1200	17.000	19.000	599.000
Apis mellifera (bee)	200	--	9.000	78.000
Drosophila melagonaster (fly)	132	15.000	20.000	388.000
Caenorhabditis elegans (worm)	97	27.000	28.000	346.000

Understanding the Domain (the Problem Space)

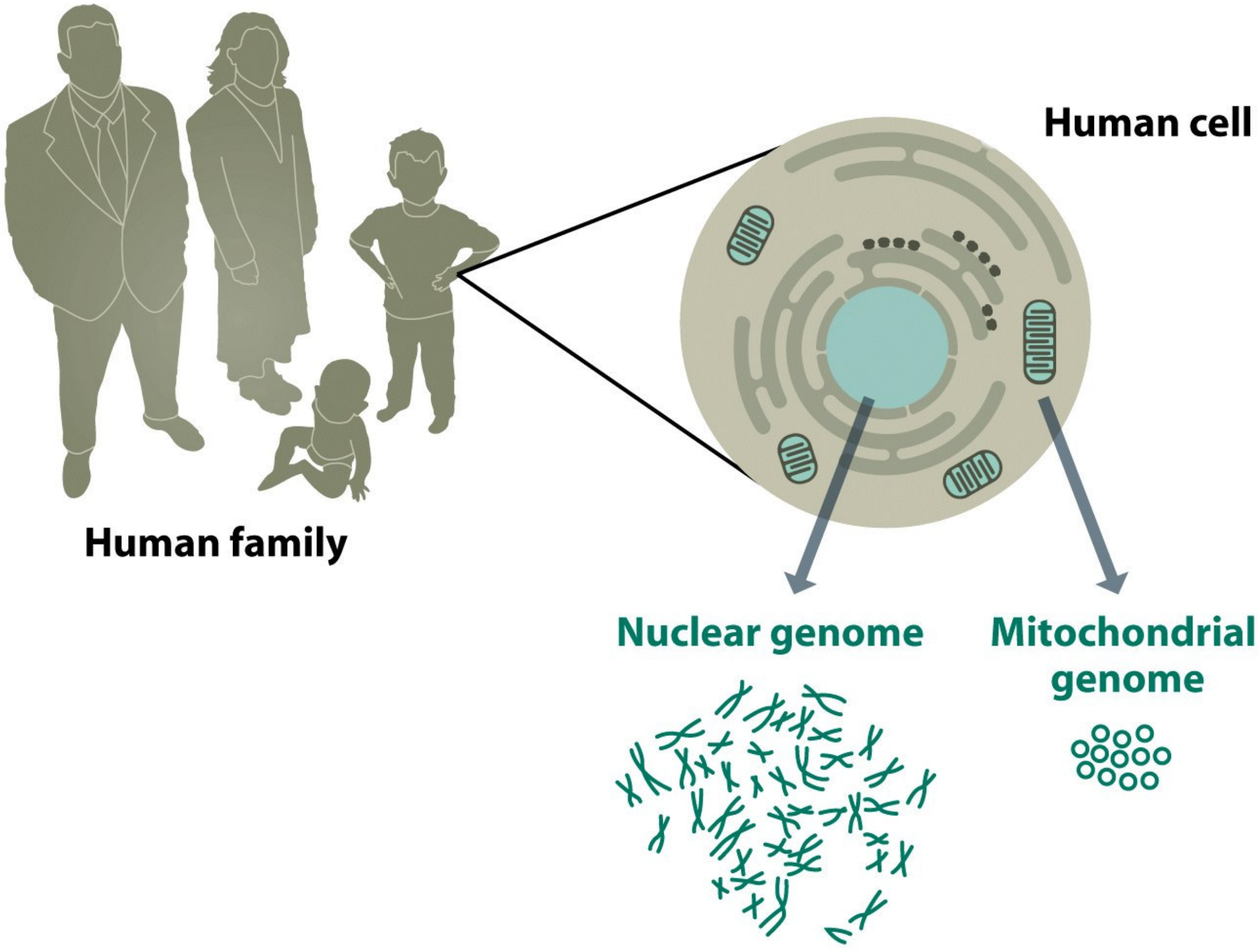
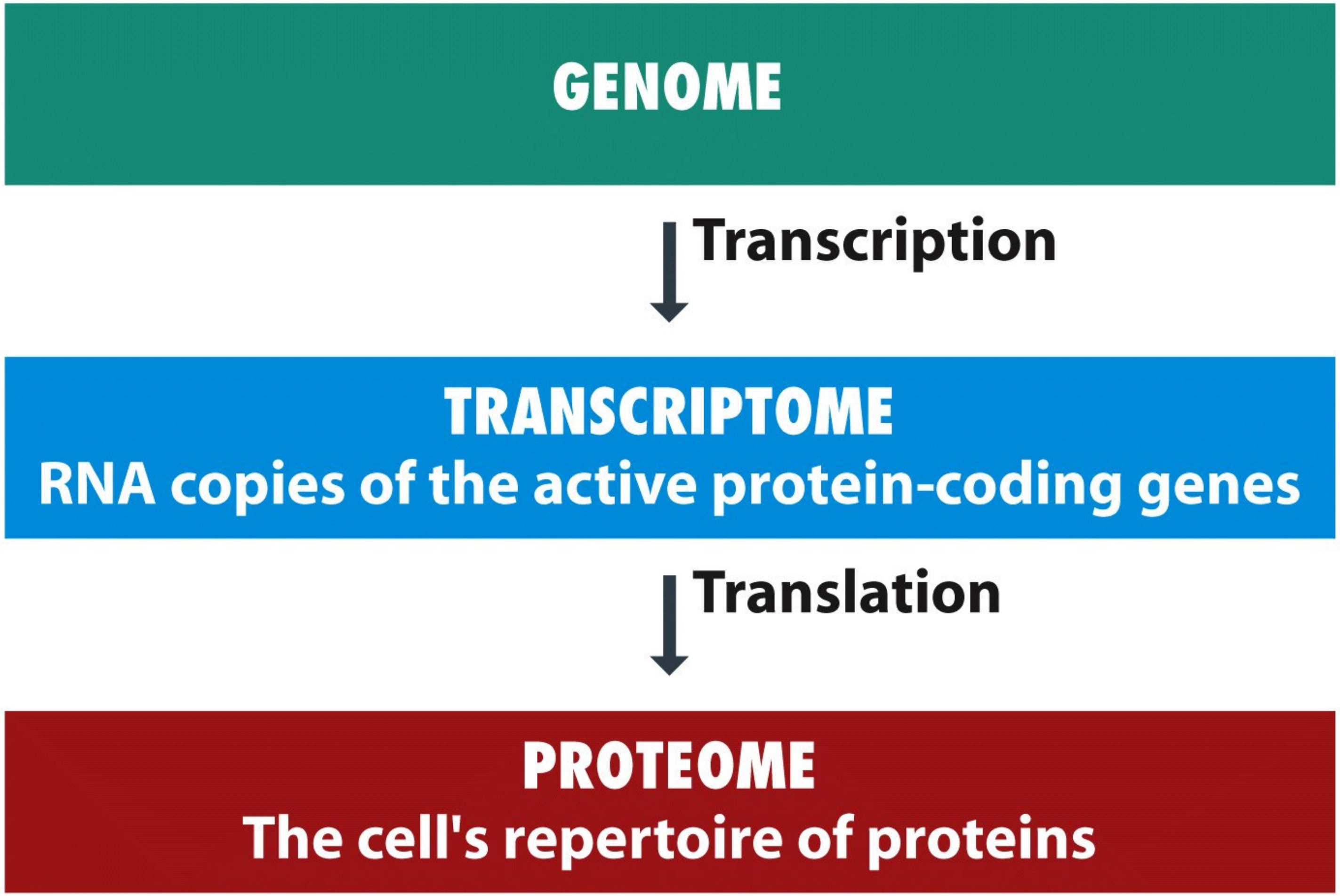


Figure 1.1 *Genomes 3* (© Garland Science 2007)

Understanding the Domain (the Problem Space)

- The genome is a store of biological information but on its own it is unable to release that information to the cell
- Each of the 10^{13} cells in the adult human body has its own copies of the genome
- Genome expression
  - Transcription: individual genes are copied into RNA molecules
  - Translation: proteins synthesized by translation of the individual RNA molecules present in the transcriptome.

Understanding the Domain (the Problem Space)



GENOME

Transcription

TRANSCRIPTOME
RNA copies of the active protein-coding genes

Translation

PROTEOME
The cell's repertoire of proteins

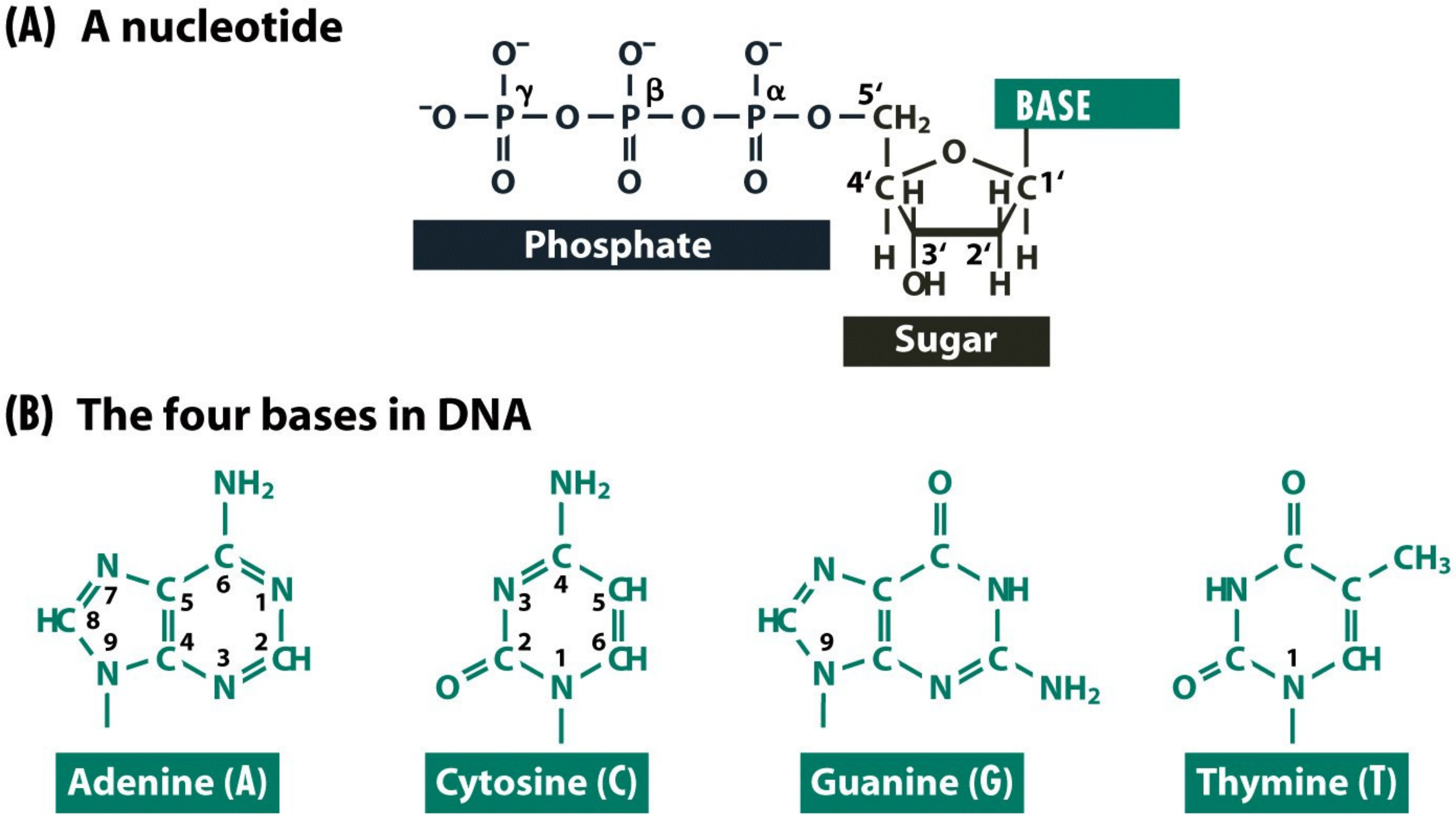
Figure 1.2 *Genomes 3* (© Garland Science 2007)

Understanding the Domain (the Problem Space)

- Genes are made of DNA
- DNA is a linear, unbranched polymer in which the monomeric subunits are four chemically distinct nucleotides than can be linked in any order and in chains containing even millions of units in lenght

Centro de Investigación
en Métodos de
Producción de Software

Understanding the Domain (the Problem Space)



(A) A nucleotide

(B) The four bases in DNA

Figure 1.4 *Genomes 3* (© Garland Science 2007)

UNIVERSIDAD
POLITECNICA
DE VALENCIA

Centro de Investigación en Métodos de Producción de Software

Understanding the Domain (the Problem Space)

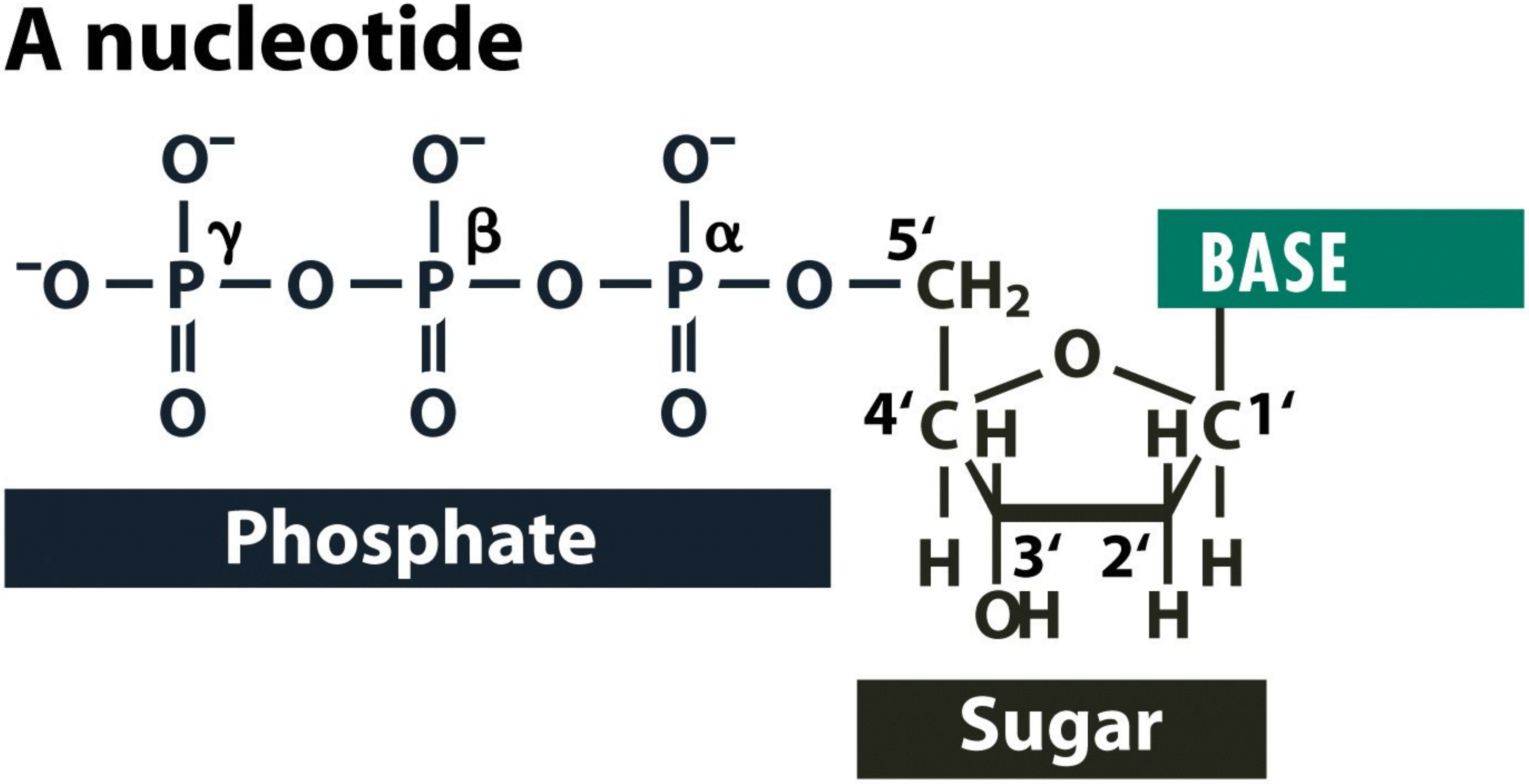


Figure 1.4a *Genomes 3* (© Garland Science 2007)

UNIVERSIDAD POLITECNICA DE VALENCIA

Understanding the Domain (the Problem Space)

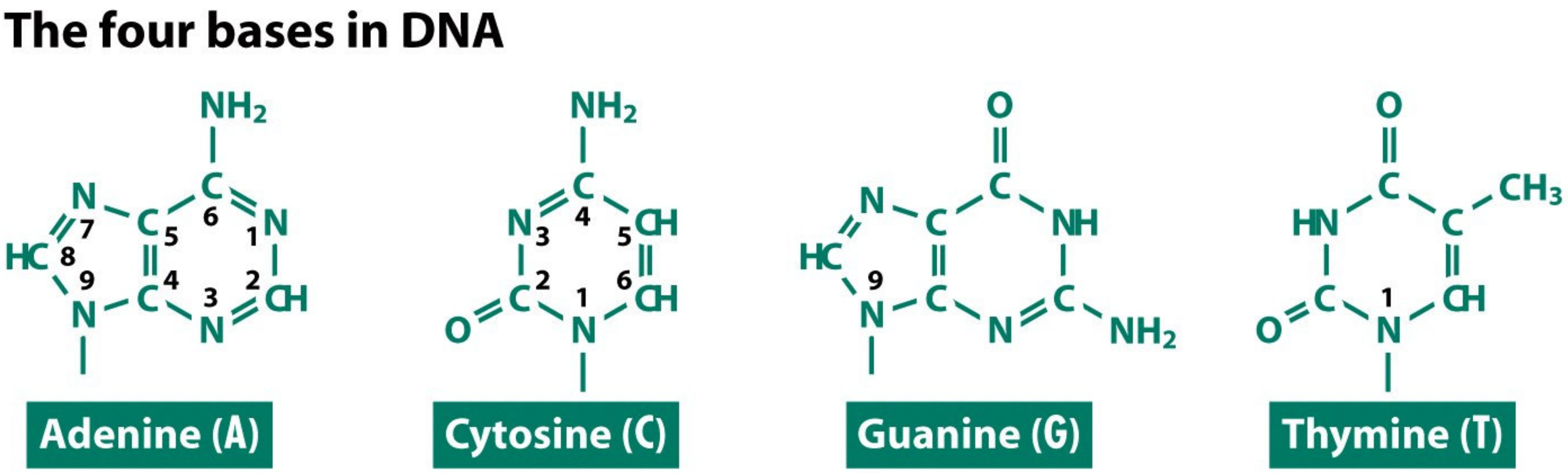


Figure 1.4b *Genomes 3* (© Garland Science 2007)

Understanding the Domain (the Problem Space)

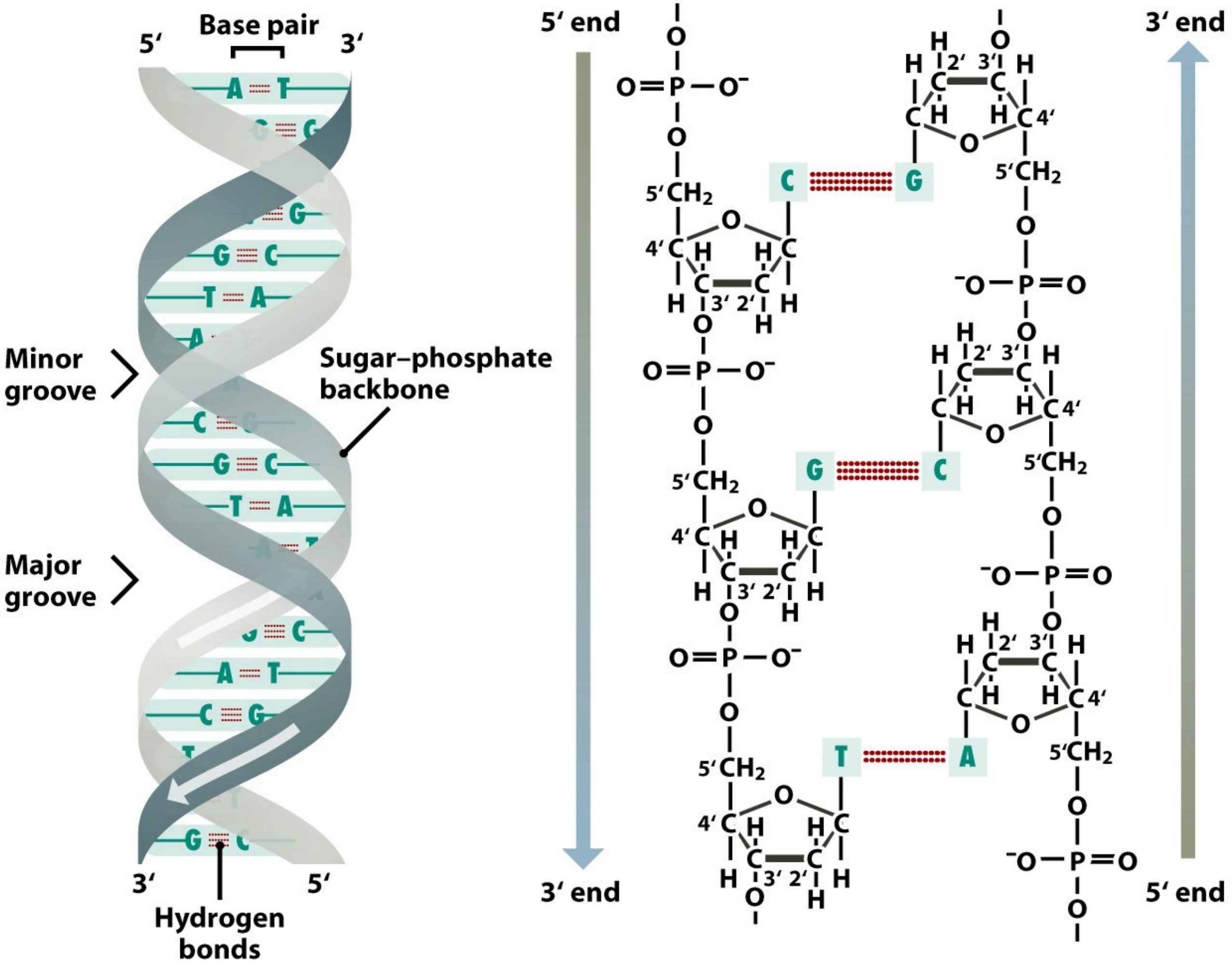


Figure 1.8a *Genomes 3* (© Garland Science 2007)

Centro de Investigación en Métodos de Producción de Software

Understanding the Domain (the Problem Space)

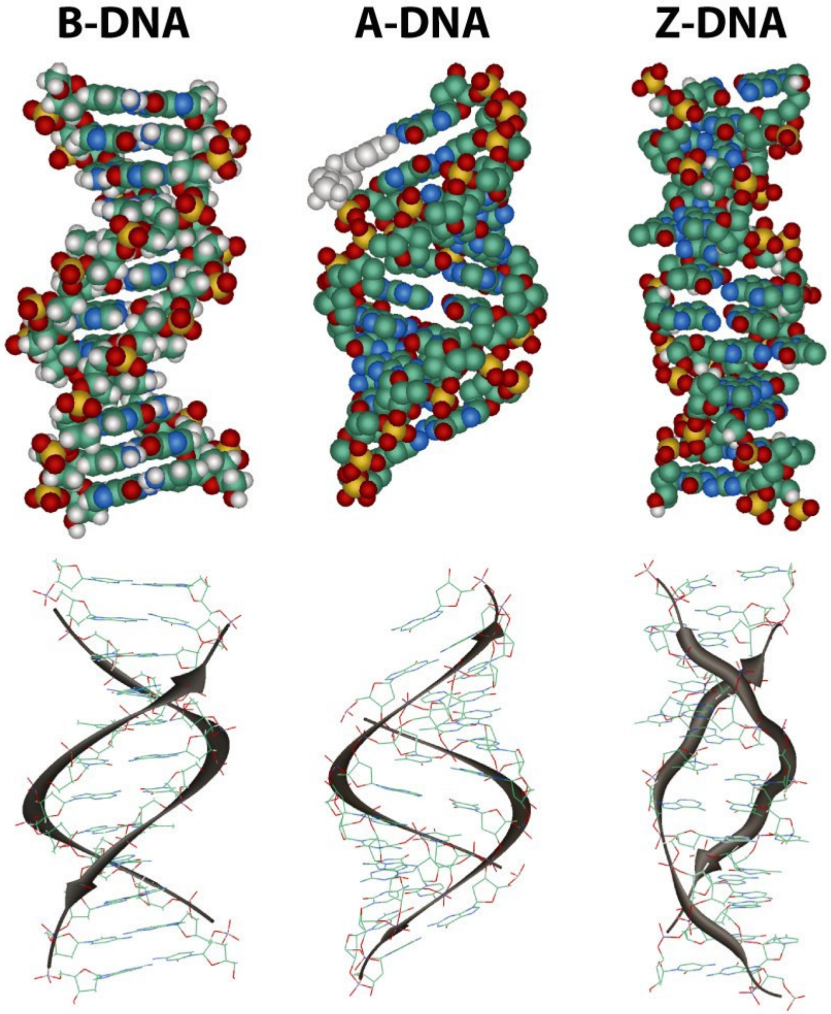
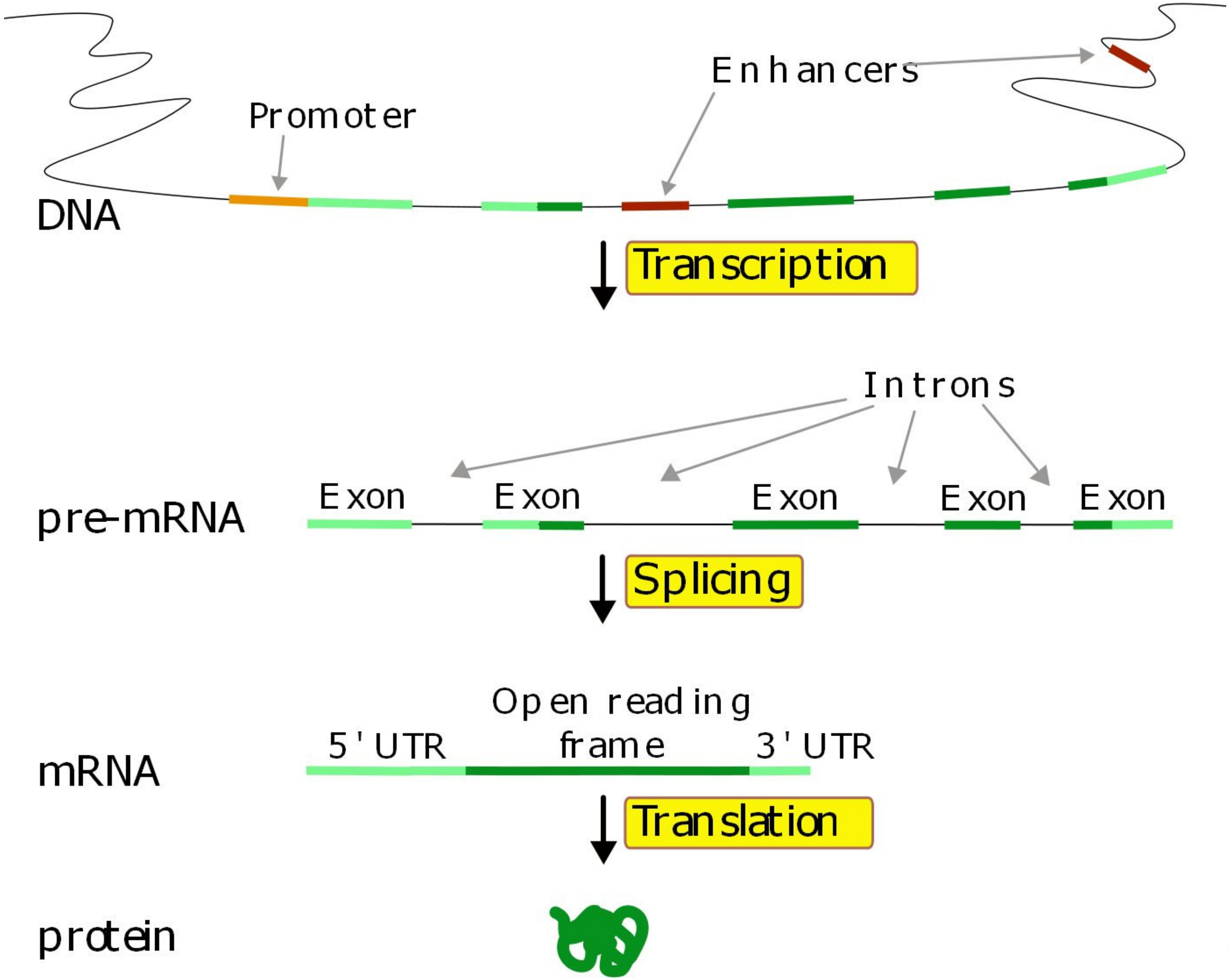


Figure 1.9 *Genomes 3* (© Garland Science 2007)

UNIVERSIDAD POLITECNICA DE VALENCIA

Understanding the Domain (the Problem Space)



Understanding the Domain (the Problem Space)

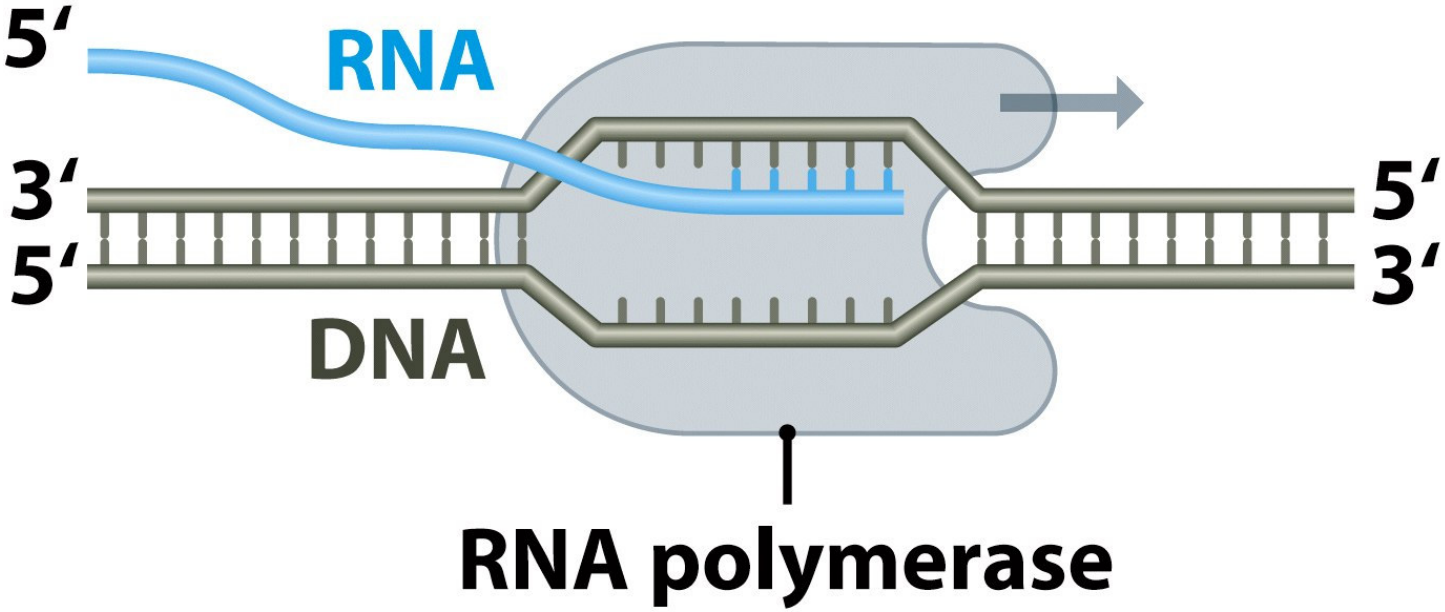


Figure 12.2 *Genomes 3* (© Garland Science 2007)

Understanding the Domain (the Problem Space)

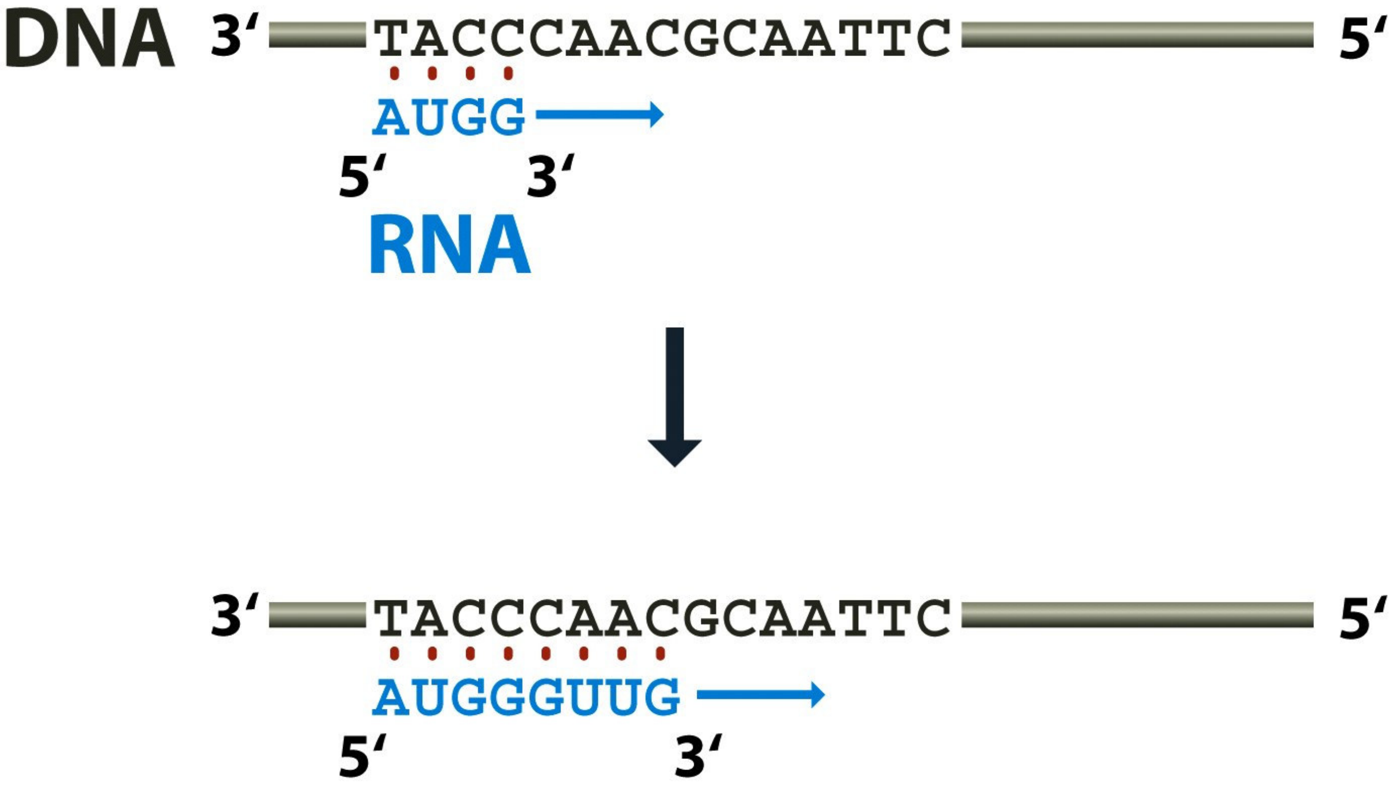


Figure 1.11 *Genomes 3* (© Garland Science 2007)

ProS
Centro de Investigación en Métodos de Producción de Software

Understanding the Domain (the Problem Space)

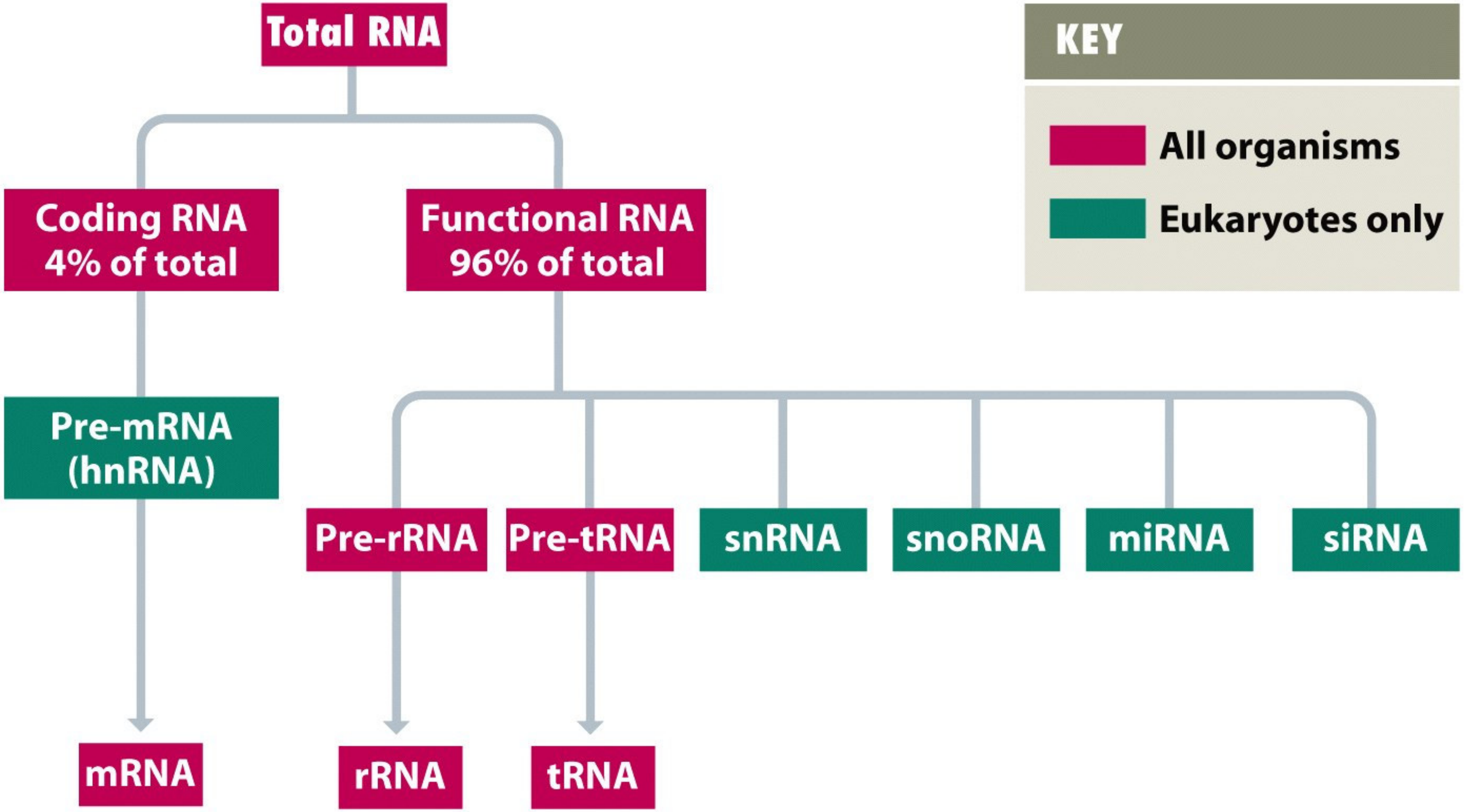


Figure 1.12 *Genomes 3* (© Garland Science 2007)

UNIVERSIDAD POLITECNICA DE VALENCIA

Understanding the Domain (the Problem Space)

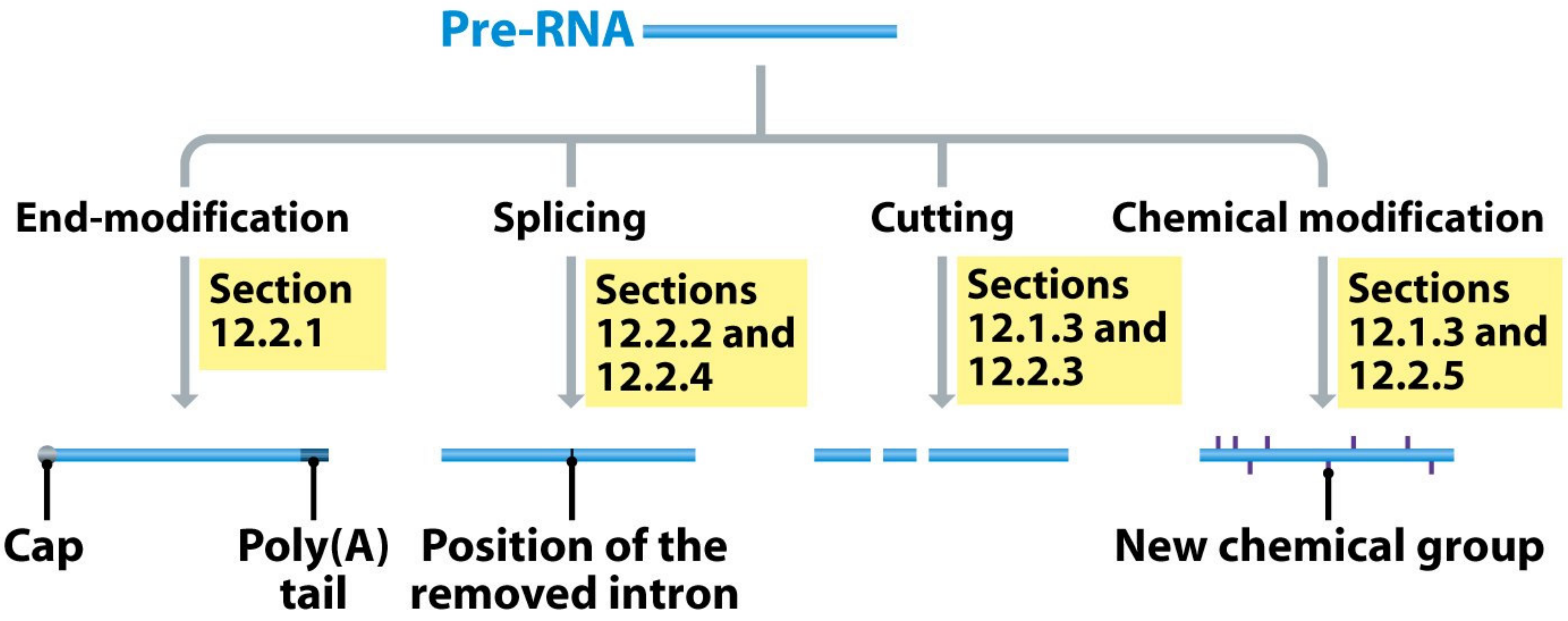

Pre-RNA
End-modification
Section 12.2.1
Splicing
Sections 12.2.2 and 12.2.4
Cutting
Sections 12.1.3 and 12.2.3
Chemical modification
Sections 12.1.3 and 12.2.5
Cap
Poly(A) tail
Position of the removed intron
New chemical group

Figure 1.13 *Genomes 3* (© Garland Science 2007)

Centro de Investigación en Métodos de Producción de Software

Understanding the Domain (the Problem Space)

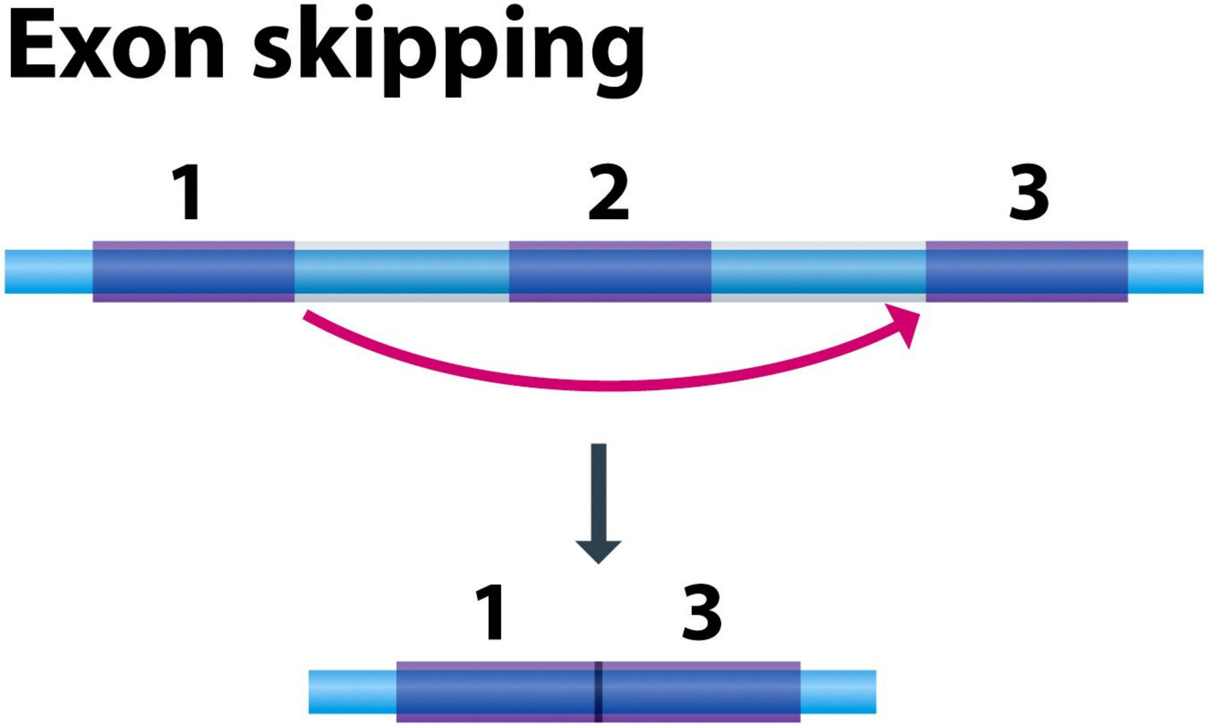


Figure 12.28a *Genomes 3* (© Garland Science 2007)

UNIVERSIDAD POLITECNICA DE VALENCIA

Centro de Investigación en Métodos de Producción de Software

Understanding the Domain (the Problem Space)

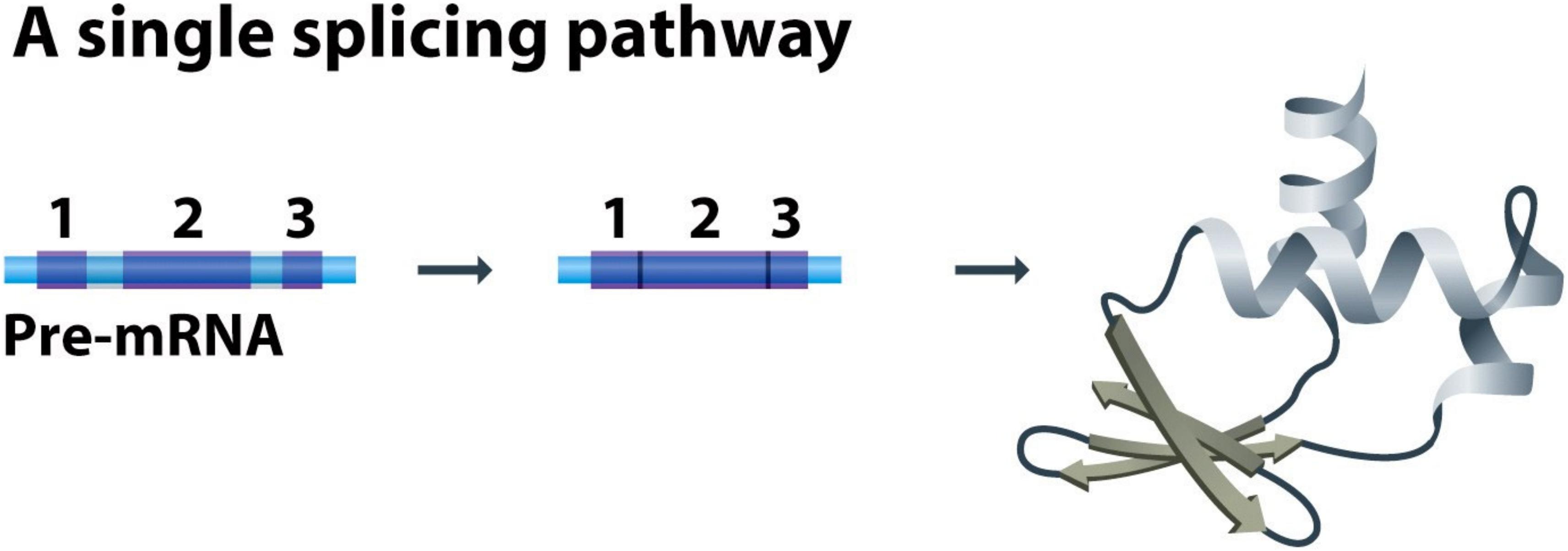


Figure 12.32a *Genomes 3* (© Garland Science 2007)

UNIVERSIDAD POLITECNICA DE VALENCIA

Understanding the Domain (the Problem Space)

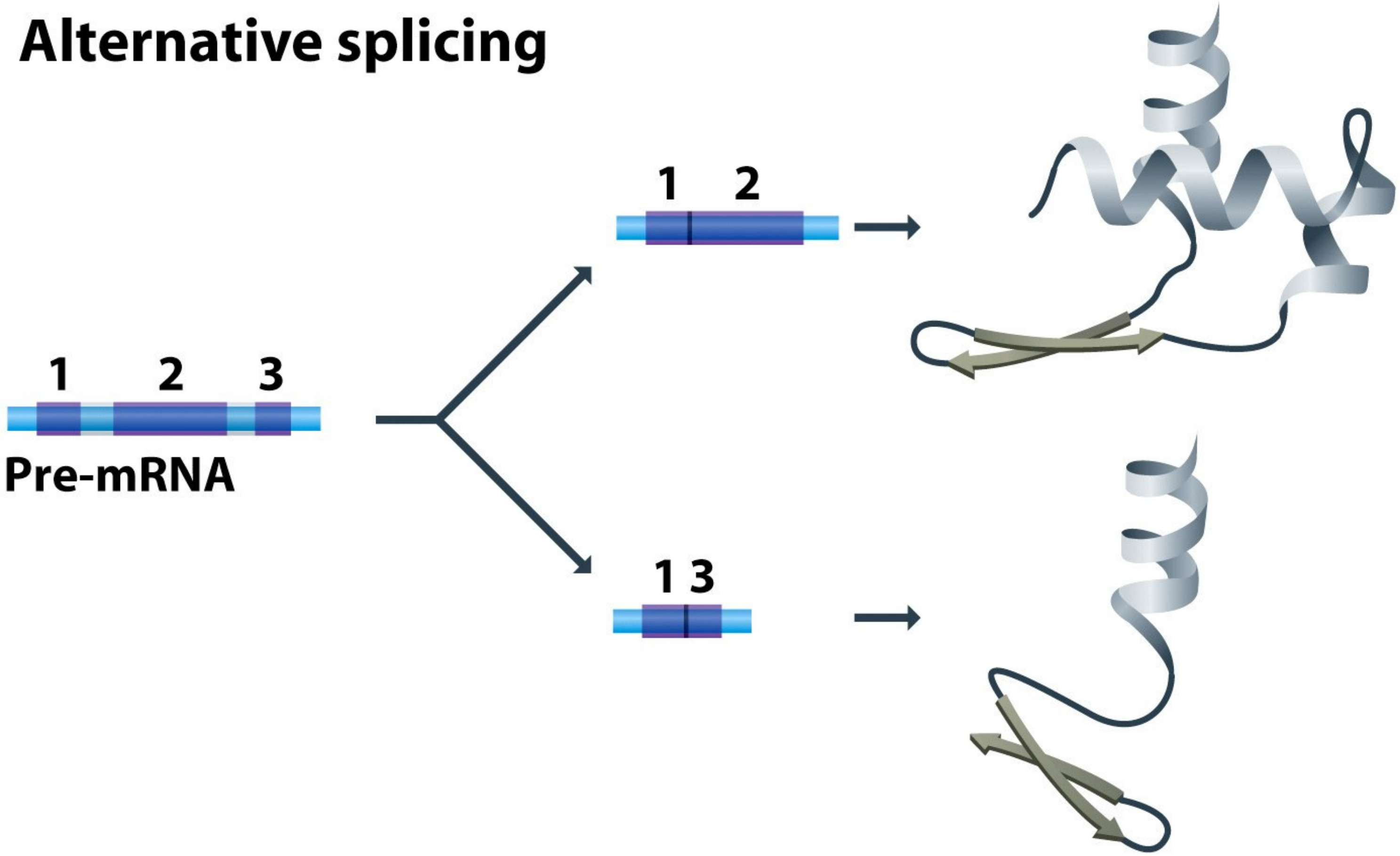


Figure 12.32b *Genomes 3* (© Garland Science 2007)

From transcriptome to proteome

- The flow of information from DNA to RNA by transcription does not provide any conceptual difficulty
- The second phase of genome expression is less easy to understand
- mRNA molecules of the transcriptome direct synthesis of proteins
- Existence of an adaptor molecule –tRNA- that forms a bridge between the mRNA and the polypeptide being synthesized

From transcriptome to proteome

- Genetic code: how the nucleotide sequence of an mRNA is translated into the aminoacid sequence of a protein
- Proteins are made up from a set of 20 aminoacids
- Different sequences of amino acids result in different combinations of chemical reactivities
- Codon: codeword comprising three nucleotides
- Two-letter code is not enought, three-letter code provides 64 potential codons
- Code degeneracy
- Punctuation codons

From transcriptome to proteome

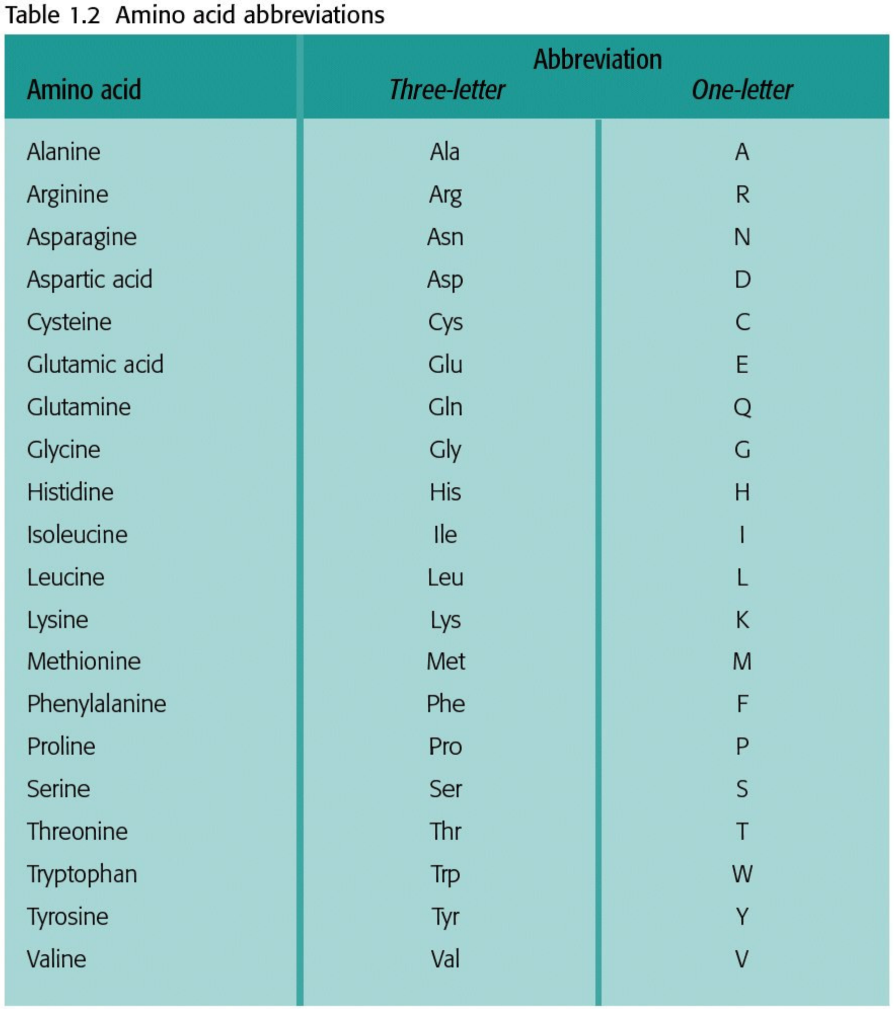
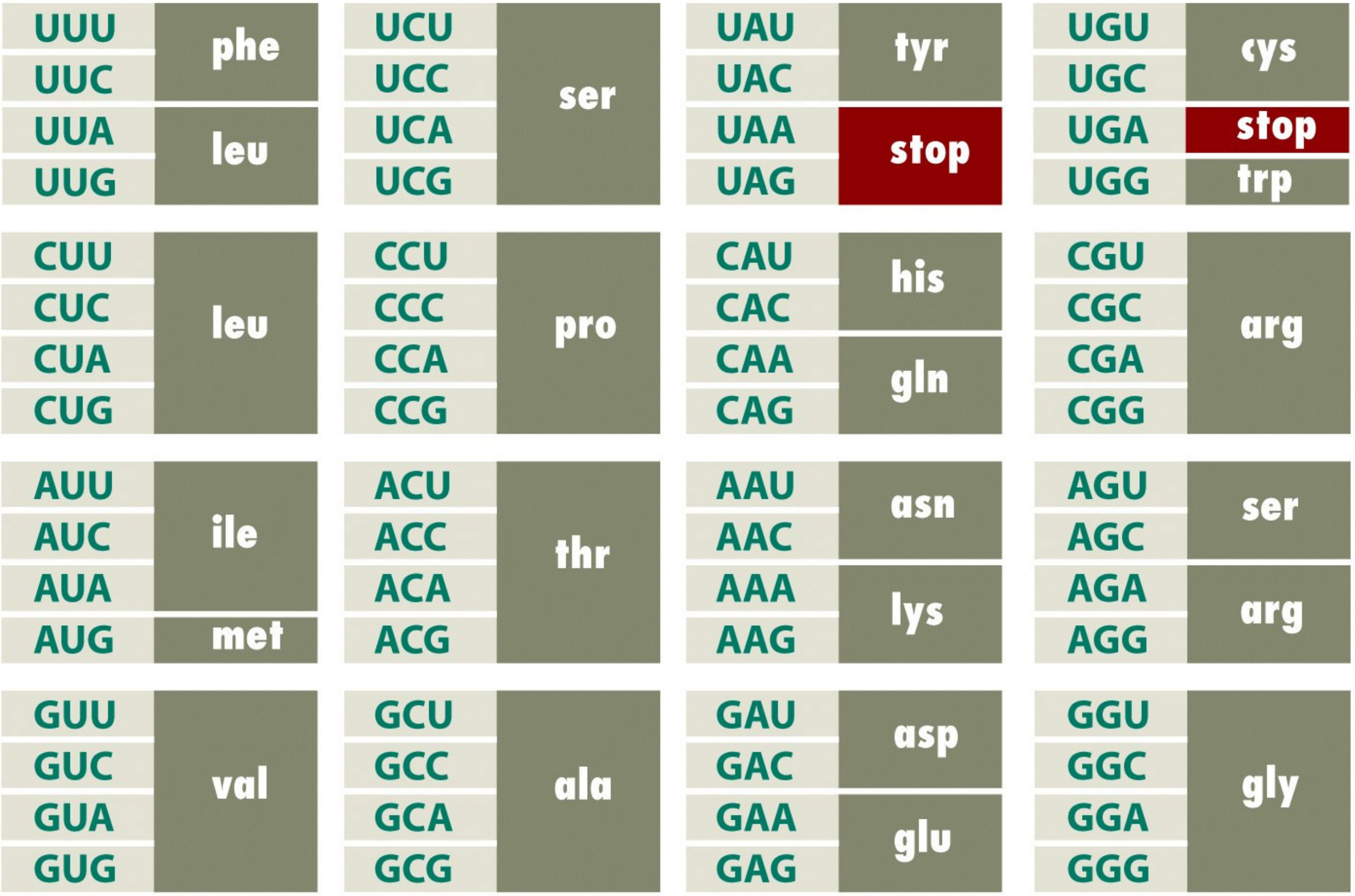


Table 1.2 Amino acid abbreviations

Amino acid	Abbreviation	
	Three-letter	*One-letter*
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Table 1.2 *Genomes 3* (© Garland Science 2007)

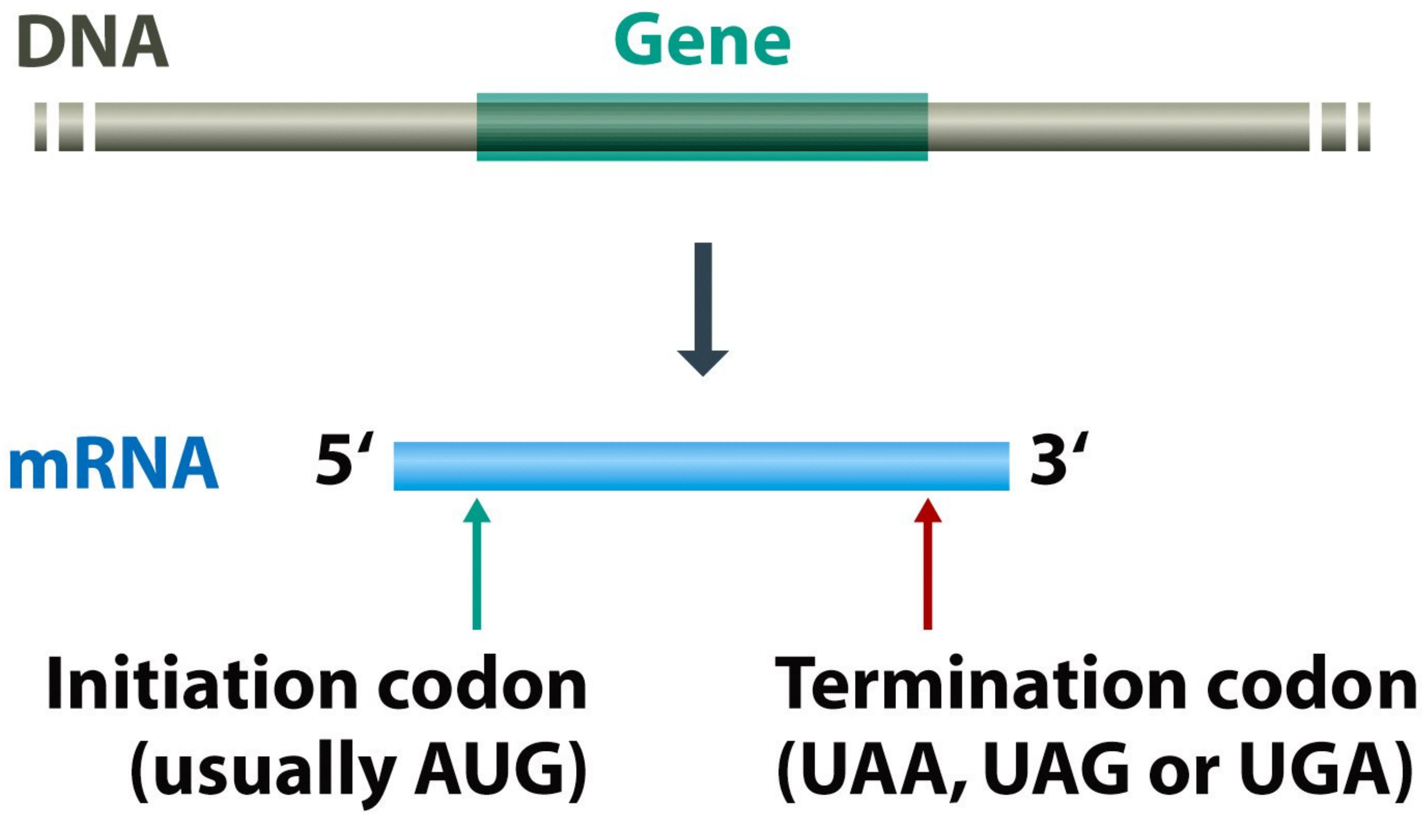
From transcriptome to proteome



UUU	phe	UCU	ser	UAU	tyr	UGU	cys
UUC		UCC		UAC		UGC	
UUA	leu	UCA		UAA	stop	UGA	stop
UUG		UCG		UAG		UGG	trp
CUU	leu	CCU	pro	CAU	his	CGU	arg
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	gln	CGA	
CUG		CCG		CAG		CGG	
AUU	ile	ACU	thr	AAU	asn	AGU	ser
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	lys	AGA	arg
AUG	met	ACG		AAG		AGG	
GUU	val	GCU	ala	GAU	asp	GGU	gly
GUC		GCC		GAC		GGC	
GUA		GCA		GAA	glu	GGA	
GUG		GCG		GAG		GGG	

Figure 1.20 *Genomes 3* (© Garland Science 2007)

From transcriptome to proteome


DNA

Gene

mRNA

5′

3′

Initiation codon (usually AUG)

Termination codon (UAA, UAG or UGA)

Figure 1.21 *Genomes 3* (© Garland Science 2007)

From transcriptome to proteome

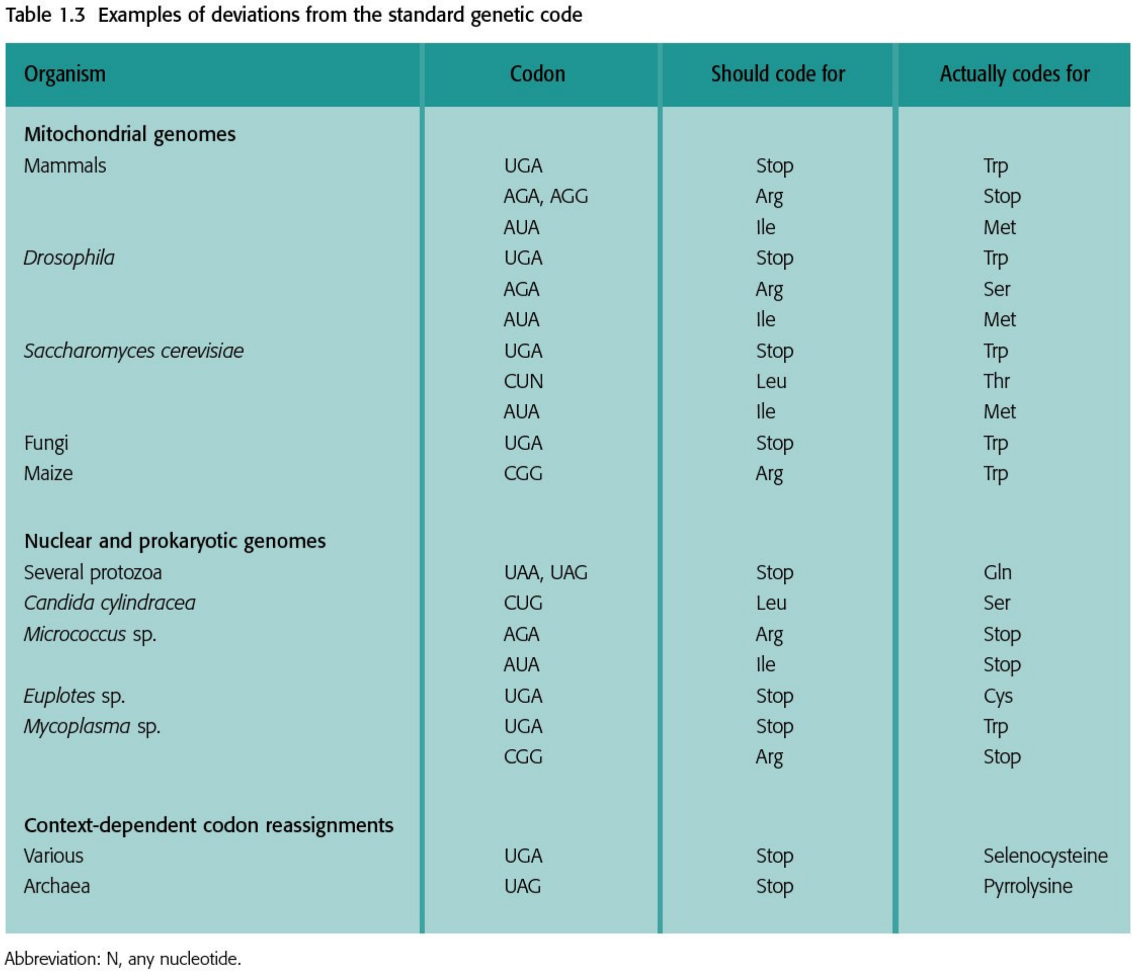


Table 1.3 Examples of deviations from the standard genetic code

Organism	Codon	Should code for	Actually codes for
Mitochondrial genomes			
Mammals	UGA	Stop	Trp
	AGA, AGG	Arg	Stop
	AUA	Ile	Met
Drosophila	UGA	Stop	Trp
	AGA	Arg	Ser
	AUA	Ile	Met
Saccharomyces cerevisiae	UGA	Stop	Trp
	CUN	Leu	Thr
	AUA	Ile	Met
Fungi	UGA	Stop	Trp
Maize	CGG	Arg	Trp
Nuclear and prokaryotic genomes			
Several protozoa	UAA, UAG	Stop	Gln
Candida cylindracea	CUG	Leu	Ser
Micrococcus sp.	AGA	Arg	Stop
	AUA	Ile	Stop
Euplotes sp.	UGA	Stop	Cys
Mycoplasma sp.	UGA	Stop	Trp
	CGG	Arg	Stop
Context-dependent codon reassignments			
Various	UGA	Stop	Selenocysteine
Archaea	UAG	Stop	Pyrrolysine

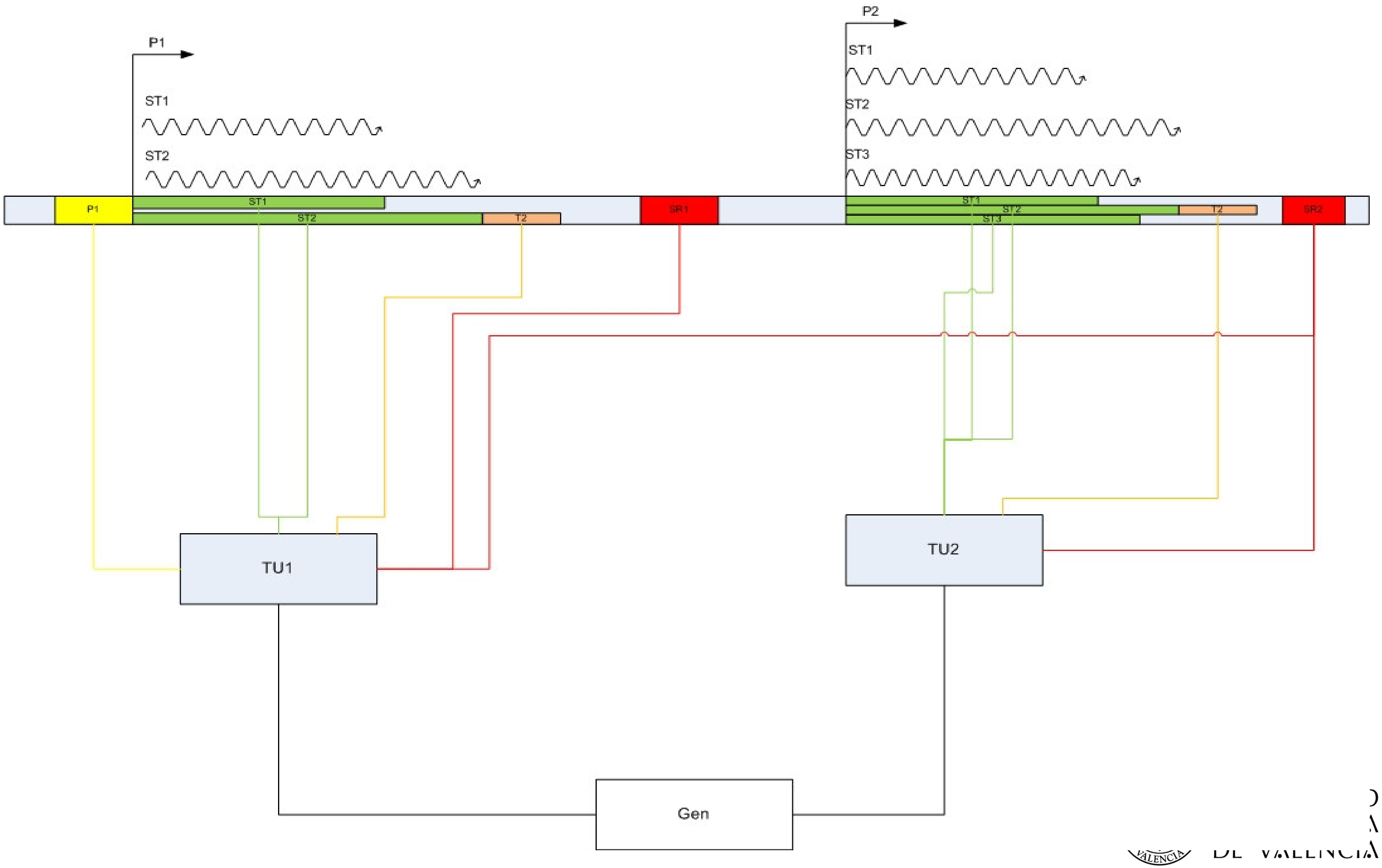
Abbreviation: N, any nucleotide.

Table 1.3 *Genomes 3* (© Garland Science 2007)

Building an ER Model

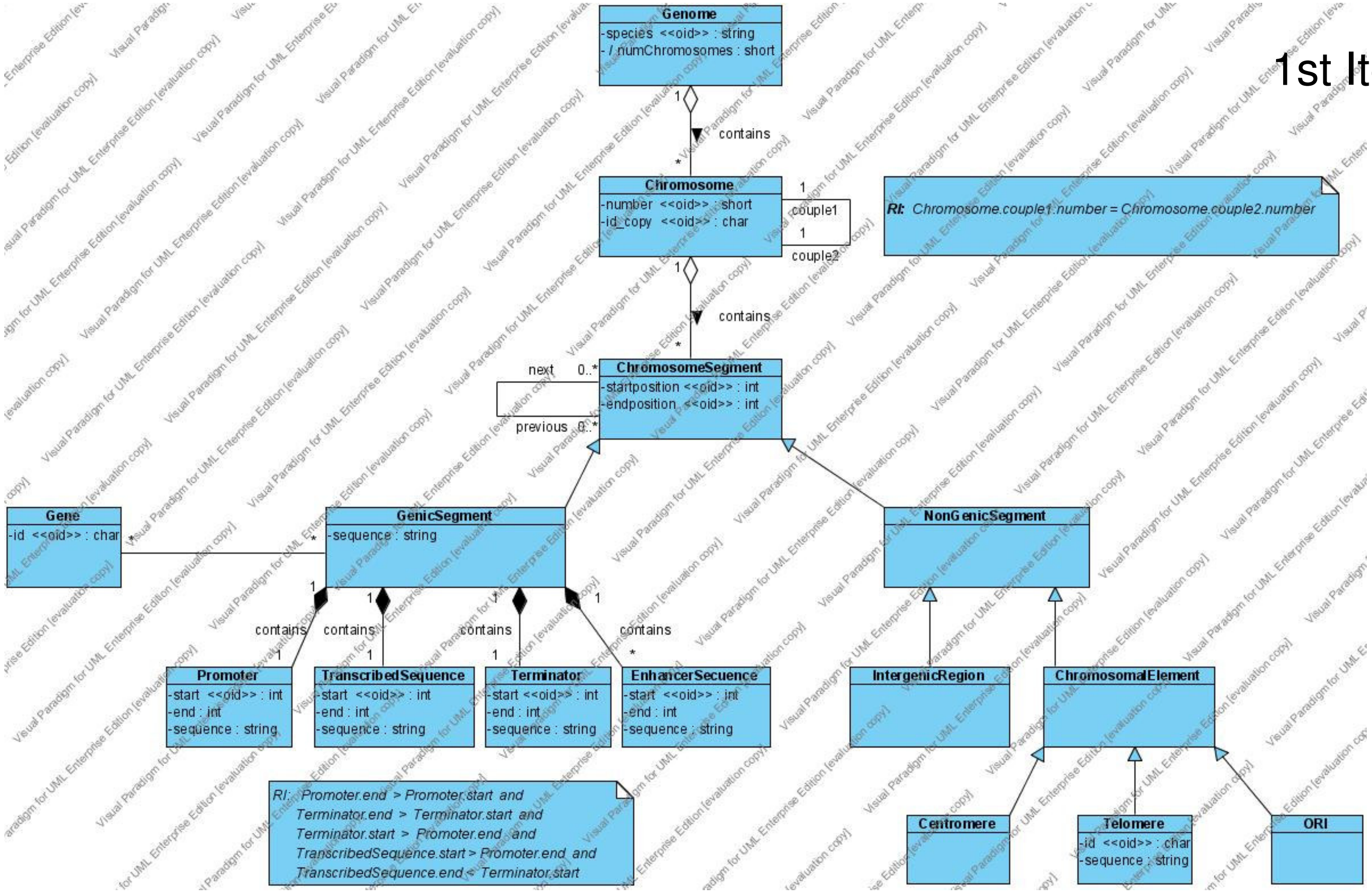
- Gene: A DNA segment containing biological information and hence coding for a RNA and/or polypedtide mollecule.
- Allele : One or two or more alternatives forms of a gene.

Building an ER Model



Genomic ER Model: Evolution

1st Iteration

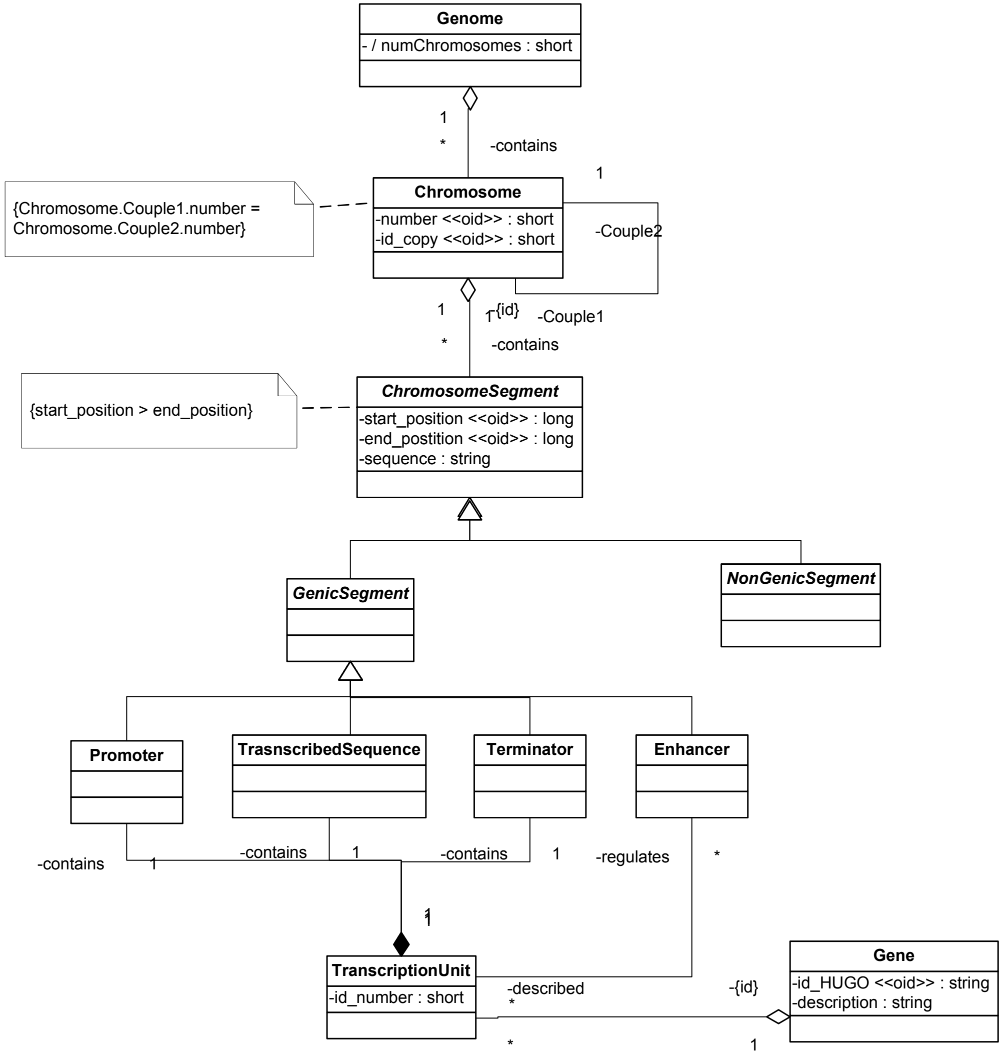


RI: Chromosome.couple1.number = Chromosome.couple2.number

RI: Promoter.end > Promoter.start and
Terminator.end > Terminator.start and
Terminator.start > Promoter.end and
TranscribedSequence.start > Promoter.end and
TranscribedSequence.end < Terminator.start

Centro de Investigación en Métodos de Producción de Software

Genomic ER Model: Evolution



2nd Iteration

UNIVERSIDAD POLITECNICA DE VALENCIA

Genomic ER Model: Evolution

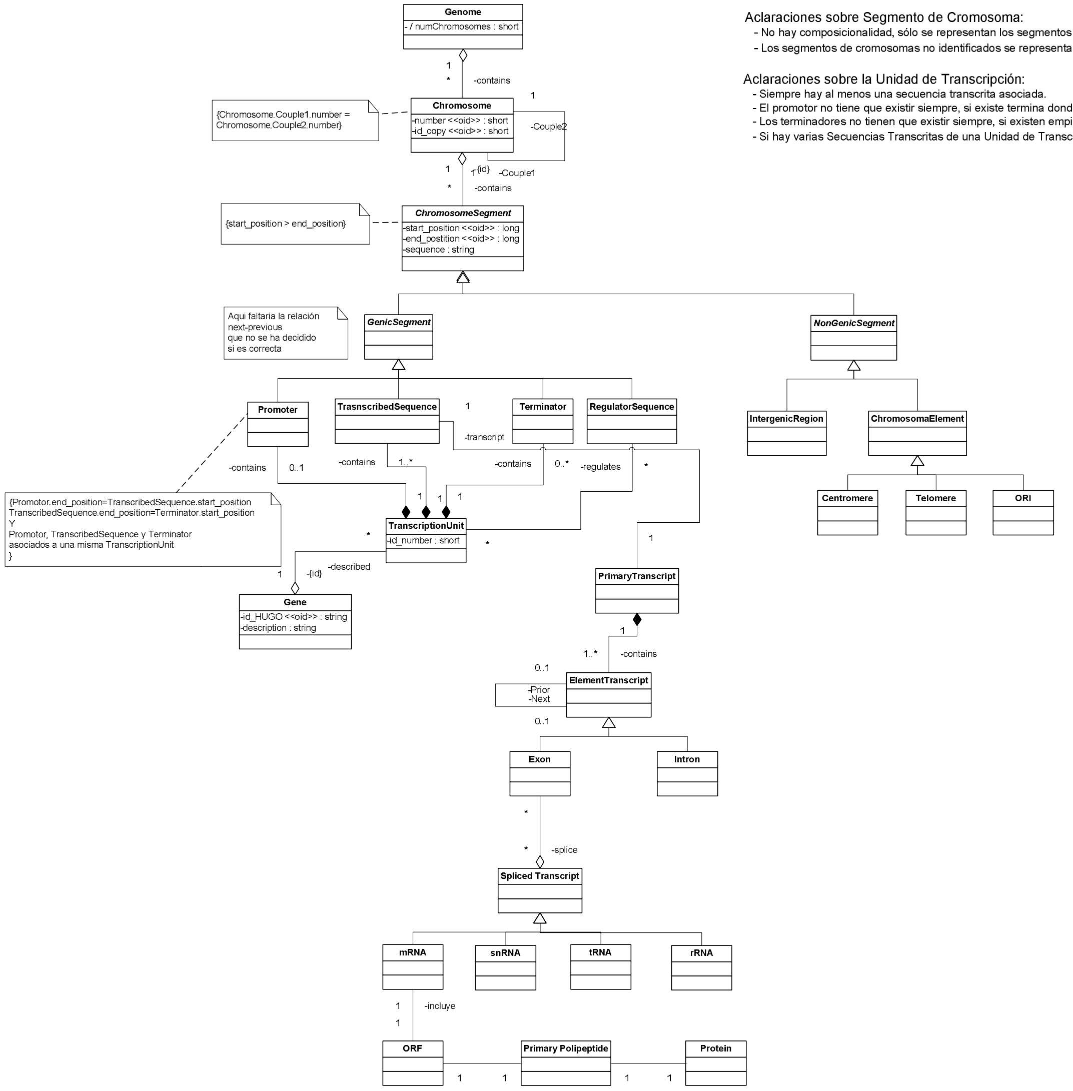
3rd Iteration

Aclaraciones sobre Segmento de Cromosoma:

- No hay composicionalidad, sólo se representan los segmentos indivisibles.
- Los segmentos de cromosomas no identificados se representan como instancias del Región InterGénica.

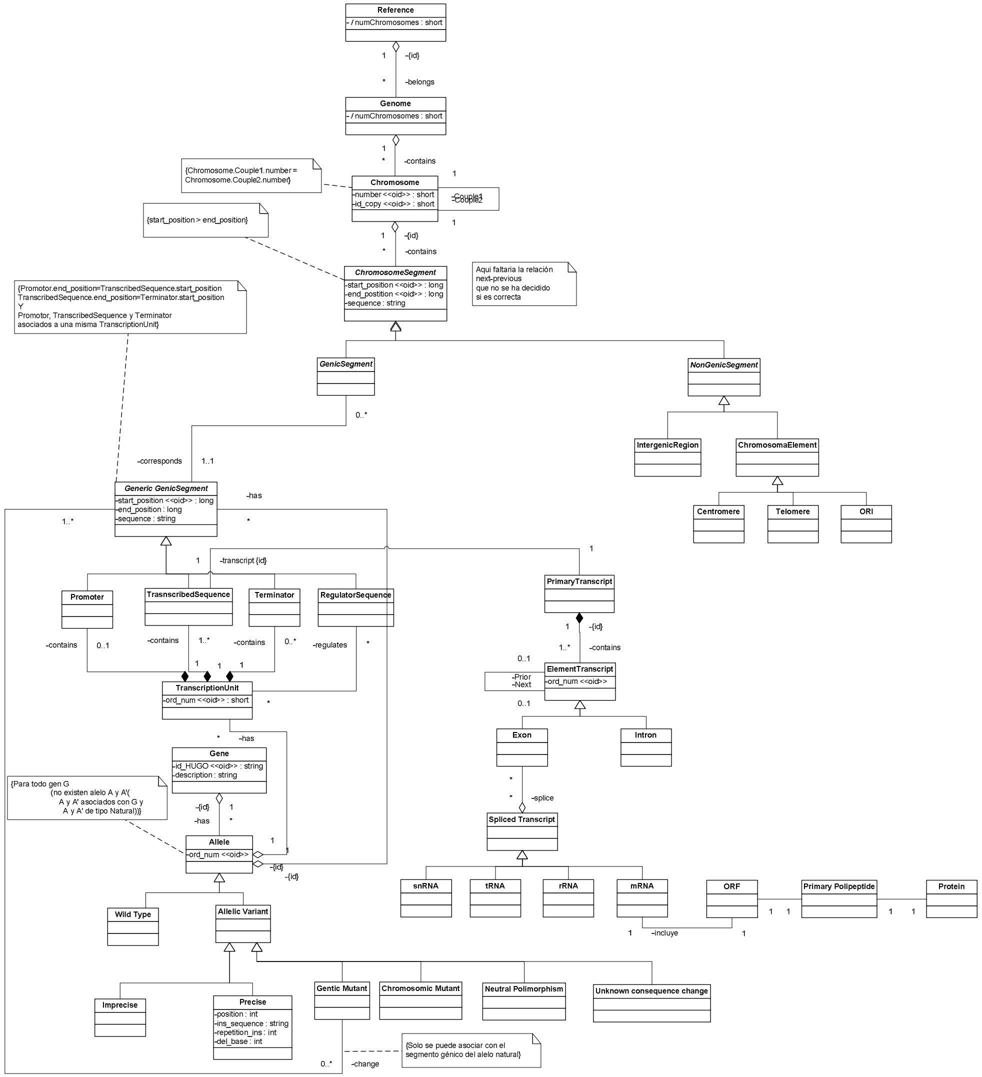
Aclaraciones sobre la Unidad de Transcripción:

- Siempre hay al menos una secuencia transcrita asociada.
- El promotor no tiene que existir siempre, si existe termina donde empiezan las secuencias transcritas.
- Los terminadores no tienen que existir siempre, si existen empiezan donde terminan las secuencias transcritas.
- Si hay varias Secuencias Transcritas de una Unidad de Transcripción, todas tienen el mismo comienzo.



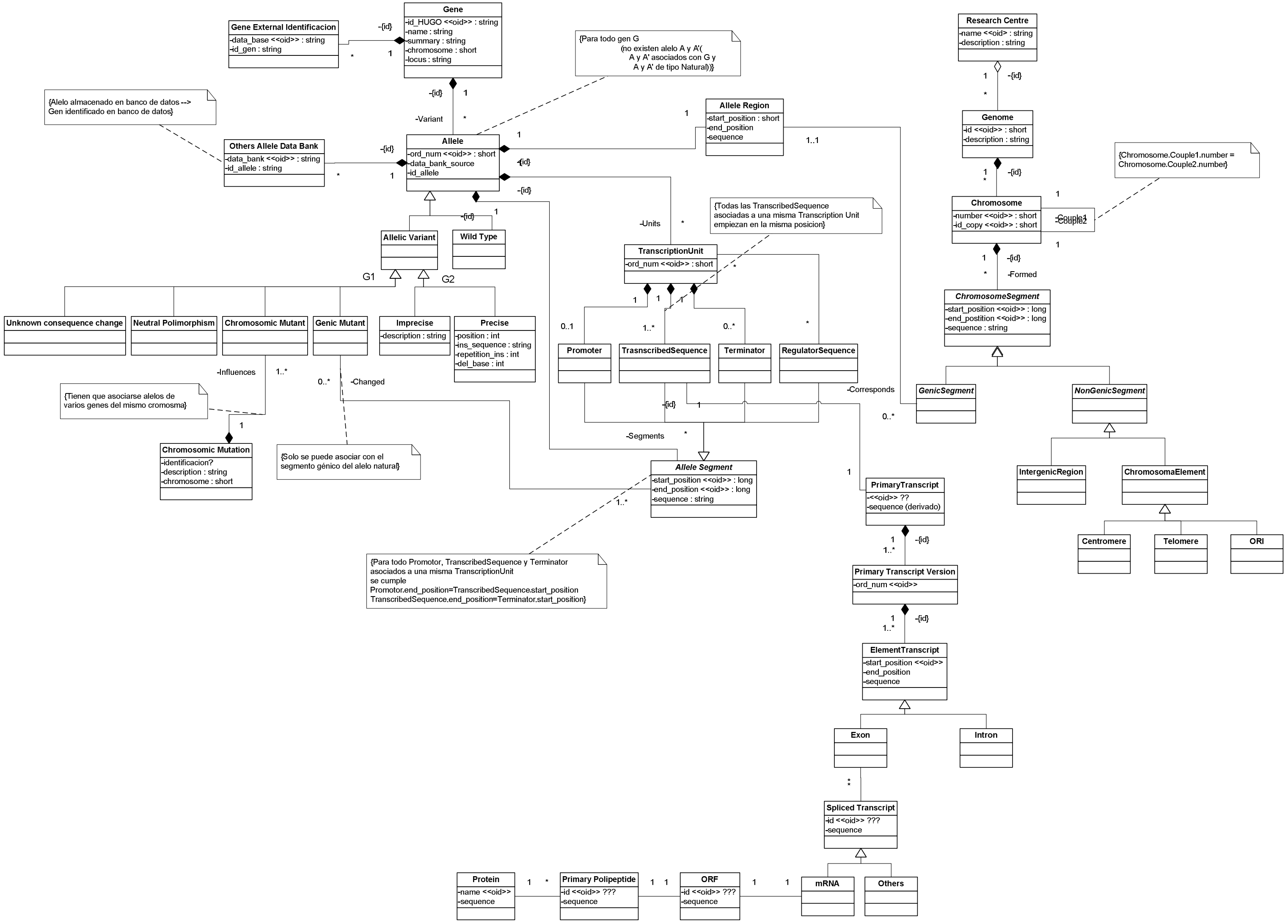
Genomic ER Model: Evolution

4th Iteration



Genomic ER Model: Evolution

Current
Iteration



{Alelo almacenado en banco de datos -->
Gen identificado en banco de datos}

{Para todo gen G
(no existen alelo A y A'(
A y A' asociados con G y
A y A' de tipo Natural))}

{Tienen que asociarse alelos de
varios genes del mismo cromosma}

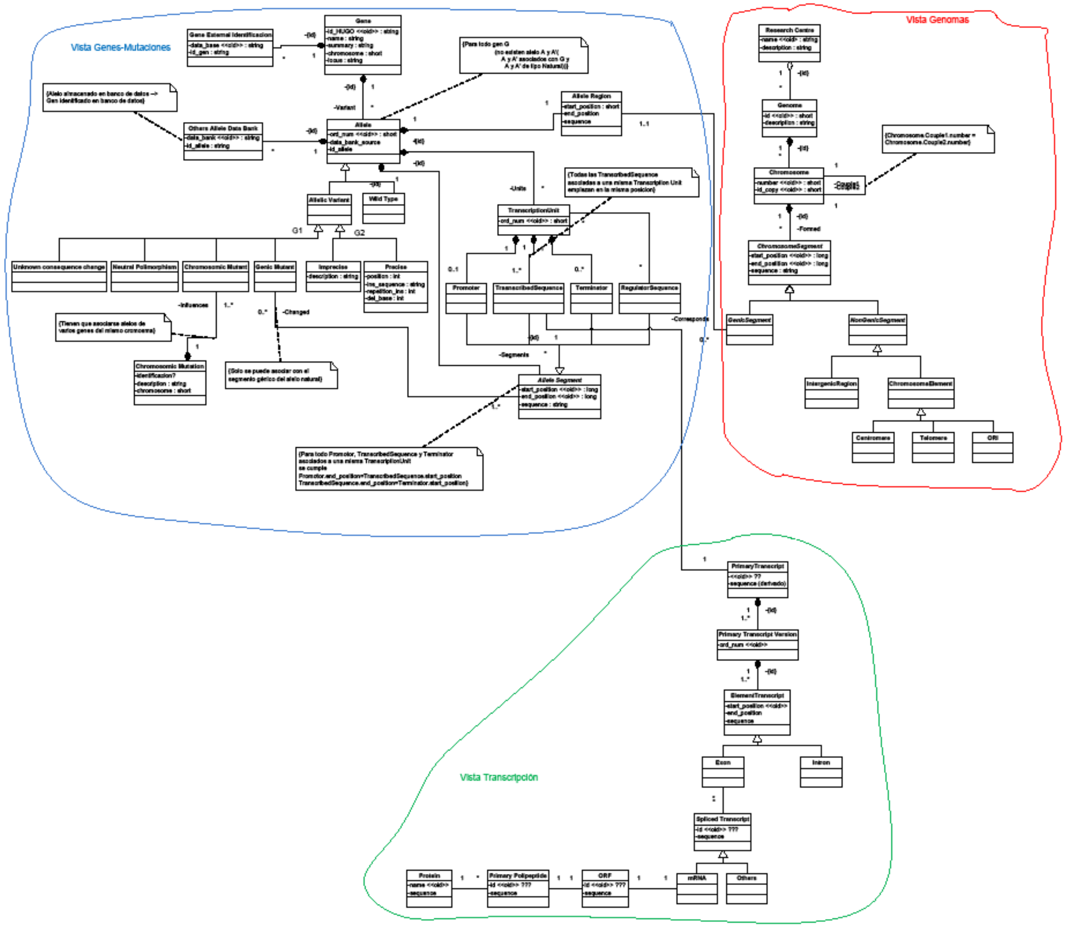
{Solo se puede asociar con el
segmento génico del alelo natural}

{Todas las TranscribedSequence
asociadas a una misma Transcription Unit
empiezan en la misma posicion}

{Para todo Promotor, TranscribedSequence y Terminator
asociados a una misma TranscriptionUnit
se cumple
Promotor.end_position=TranscribedSequence.start_position
TranscribedSequence.end_position=Terminator.start_position}

{Chromosome.Couple1.number =
Chromosome.Couple2.number}

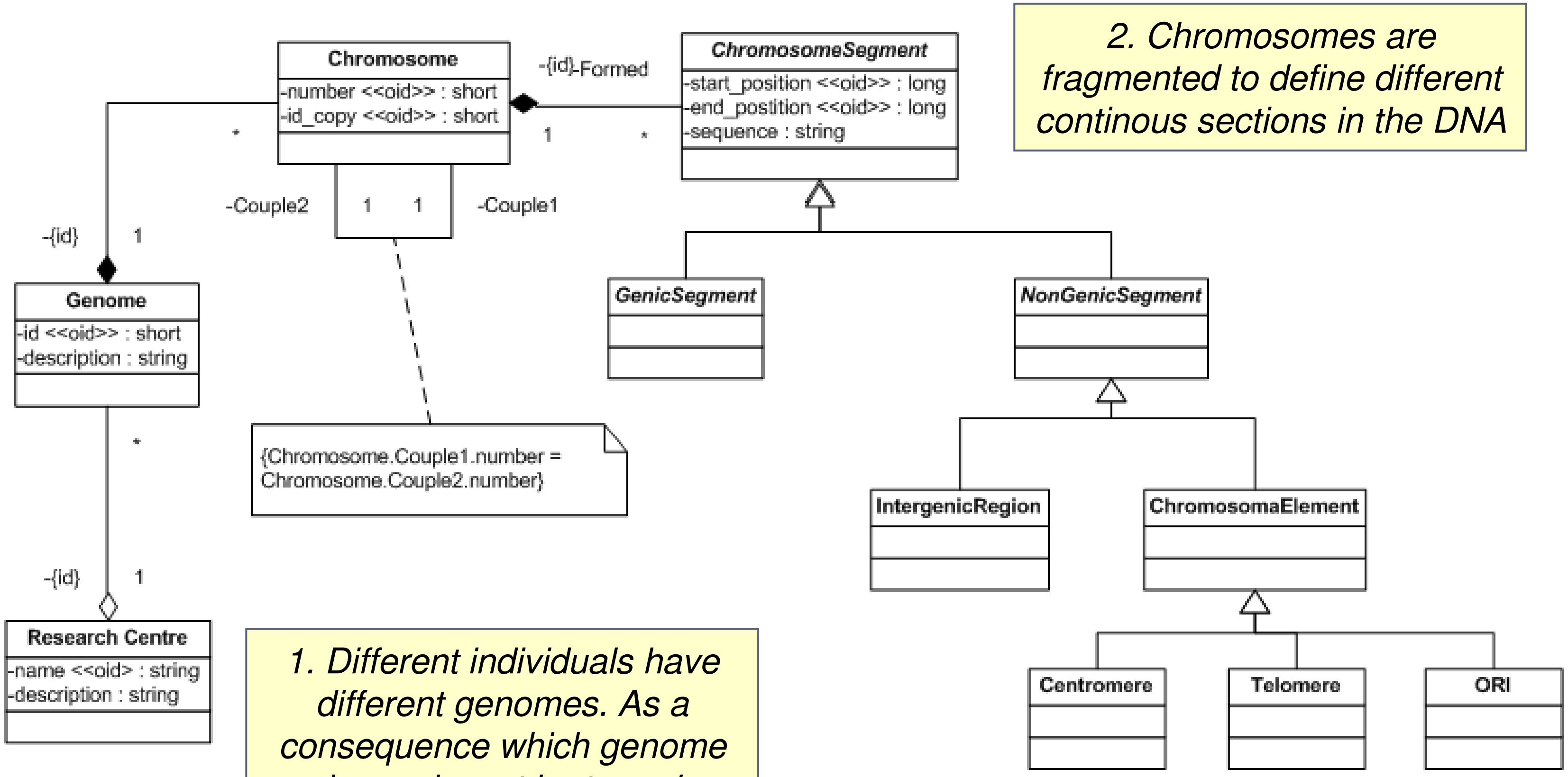
Genomic ER Model: Evolution



Genomic ER Model: Evolution

- Conceptual Genome - ER Model

Genomic ER Model : Genomic View



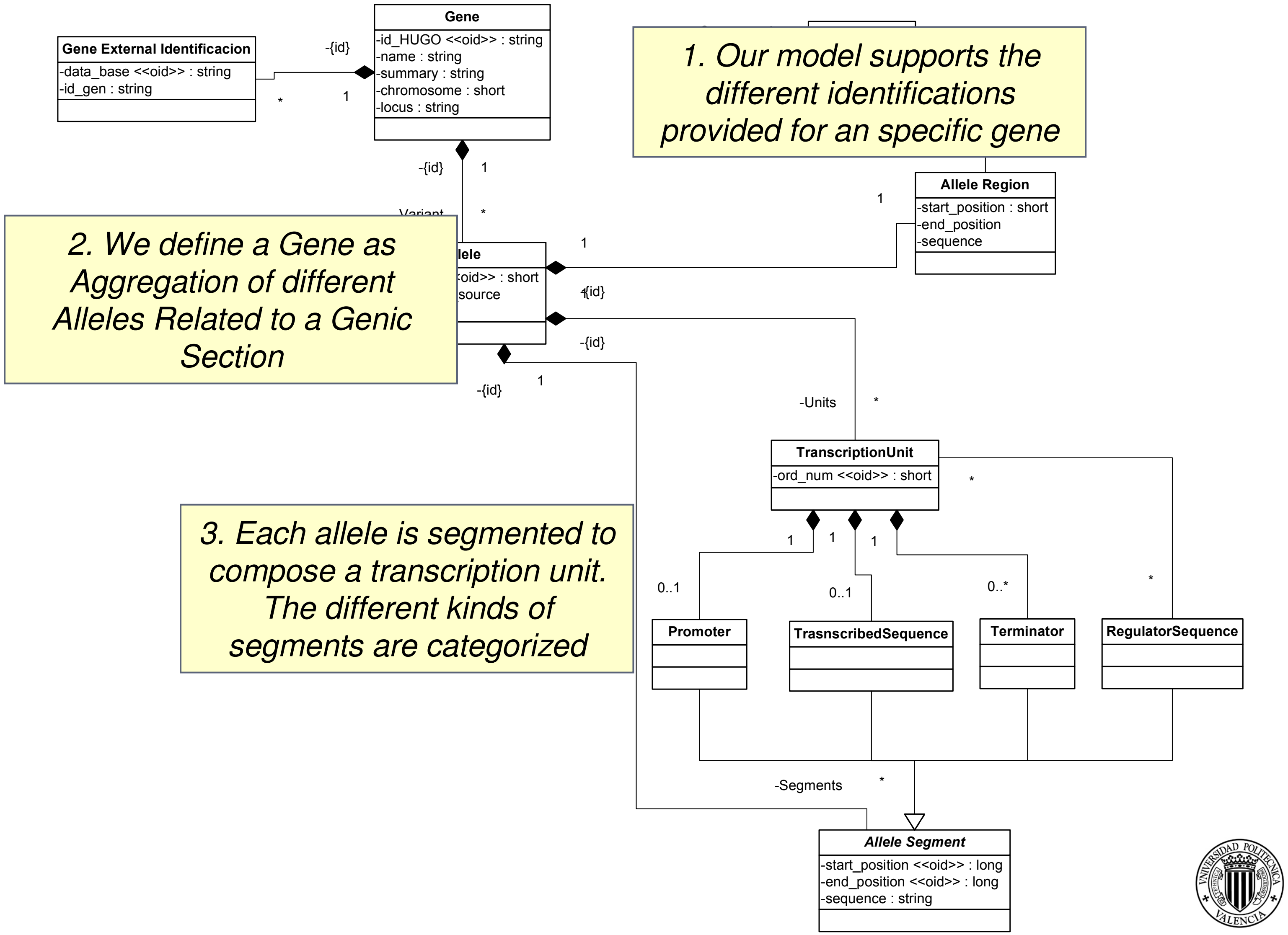
2. Chromosomes are fragmented to define different continous sections in the DNA

1. Different individuals have different genomes. As a consequence which genome is used must be traced

3. A taxonomy of the different fragments provide a clear structure to determine the relevant information

Centro de Investigación en Métodos de Producción de Software

Genomic ER Model: Gene View



1. Our model supports the different identifications provided for an specific gene

2. We define a Gene as Aggregation of different Alleles Related to a Genic Section

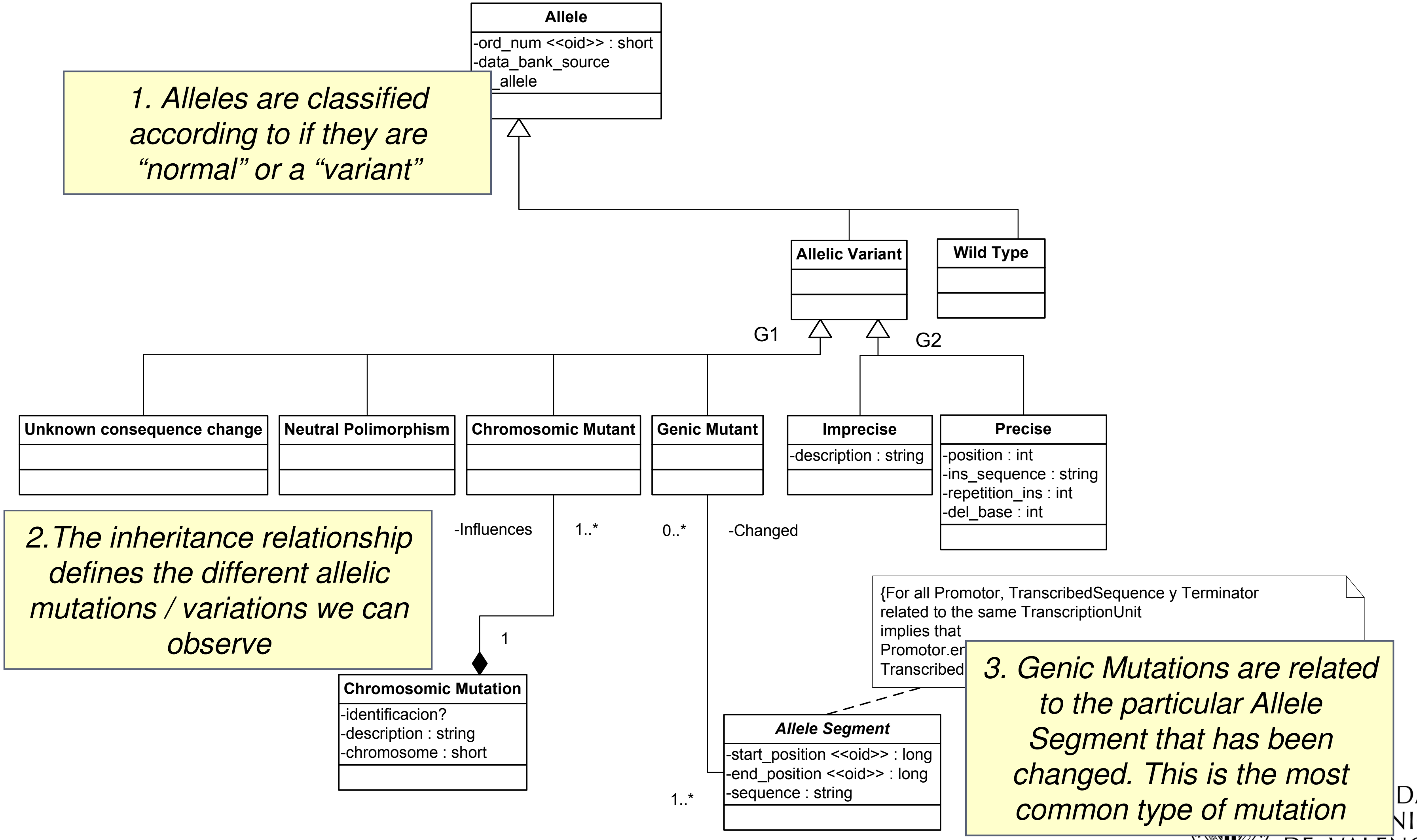
3. Each allele is segmented to compose a transcription unit. The different kinds of segments are categorized

UNIVERSIDAD POLITECNICA DE VALENCIA

- ***New objects***
  - *Allele Segment*: renaming the *Segment* object to emphasize that we mean segments of alleles.
  - *Gene External Identification*: to store gene identifiers used in different data repositories.
  - *Allele External Identification*: to know data repositories where the allele abd its identiier are stored.
  - *Allele Region*: to keep the chromosomic region and the subsequent sequence where a given gene allele is.

- ***New associations***
  - *Corresponds (Allele Region – Genic Segment)*: This association links two similar concepts; one at the gene information reference level (*Allele Region*) and another at the particular genome level (*Genic Segment*).

Genomic ER Model : Mutation View



Allele

-ord_num <<oid>> : short
-data_bank_source
_allele

1. Alleles are classified according to if they are “normal” or a “variant”

Allelic Variant

Wild Type

G1

G2

Unknown consequence change

Neutral Polimorphism

Chromosomic Mutant

Genic Mutant

Imprecise

-description : string

Precise

-position : int
-ins_sequence : string
-repetition_ins : int
-del_base : int

2.The inheritance relationship defines the different allelic mutations / variations we can observe

-Influences

1..*

0..*

-Changed

1

Chromosomic Mutation

-identificacion?
-description : string
-chromosome : short

{For all Promotor, TranscribedSequence y Terminator
related to the same TranscriptionUnit
implies that
Promotor.en
Transcribed

Allele Segment

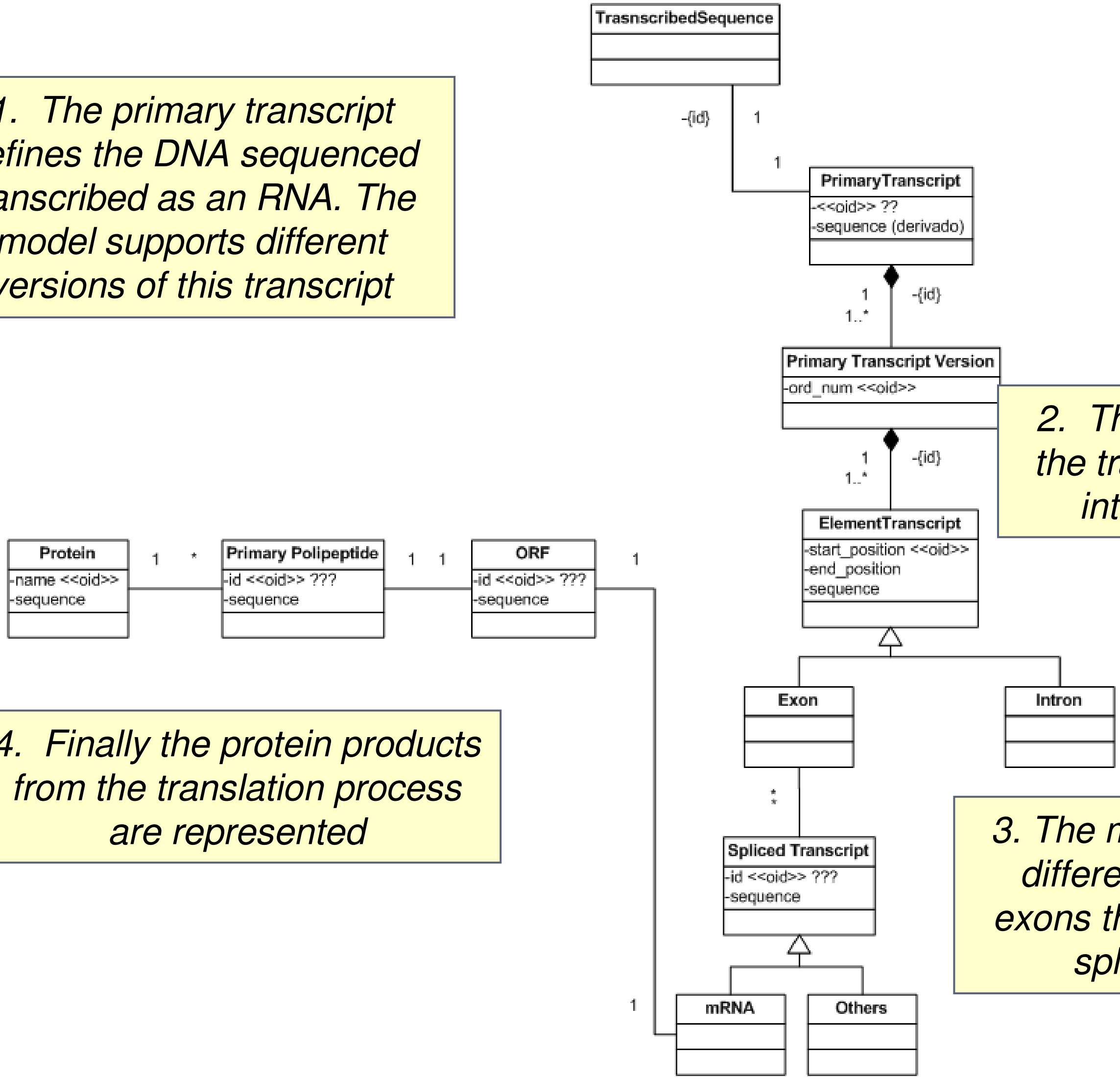
-start_position <<oid>> : long
-end_position <<oid>> : long
-sequence : string

1..*

3. Genic Mutations are related to the particular Allele Segment that has been changed. This is the most common type of mutation

Genomic ER Model : Transcription View

1. The primary transcript defines the DNA sequenced transcribed as an RNA. The model supports different versions of this transcript



2. The different elements of the transcript are subdivided into exons and introns.

3. The model represents the different combinations of exons that produce different spliced transcripts

4. Finally the protein products from the translation process are represented

Genomic ER Model: Advantages

- Can be associated to different genomic databases and allows to use several gene identifications
- It has been described using terminology commonly used by biologists
- The definition of gene take into account that is not (always) a continuous sequence of bases
- The model does not include implementation details to a particular physical database schema

Genomic ER Model: Advantages

- The Model is still to be refined and conceptually fixed…
- …but it provides a solid basis to incorporate contents in a precise and structured way
- … and the subsequent database can make possible an efficient use, content-oriented, where any human behaviour characteristic could be traced from fenotype to the involved gene(s)

So many opportunities for the future!

- **Repairing Genetic Mutations With Lasers?**
  - *Physical base: DNA strands differ in their light sensitivity depending on their base sequences.*
  - *Conceptual base: need of understanding semantics behind given sequences of nucleotides*

- **Nature versus nurture**

- **Pre-implant Genetic Diagnosis**: a technique that allows to check if an embryo is/isn’t healthy from a genetic perspective, before transfered to the maternal uterus.
  - Physical base: “assisted reproduction” technologies
  - Conceptual base: need to understand semantics of specific gene mutations

- Discovered a gene –**EYS** (for “Eyes Shut”) that **causes *inherited blindness***.
    - Physical base: mutation that gives rise to the problem
    - Conceptual base: why the mutation occurs? How to prevent it?

- Identified **295 potential therapeutics targets against AIDS**
  - Physical base: 295 human proteins that “probably” helps the AIDS to establish in the human cells
  - Conceptual base: “probably”? Under which conditions / interactions?

More and more related news...

- Tuberculosis uses a protein of the human immune system to proliferate
- Genes against the Malaria
- Genetic influences on female infidelity and number of sexual partners in humans (2004)
- Allele 334: an ‘infidelity' gene for men? (2008)
- 'Fat' gene makes millions of Britons more greedy

Conclusions

- **Understanding the Human Genome** can become an extremely hard task if research is more and more oriented to the solution space
- Discovering “human” patterns in the genomic code is really like looking for a needle in a haystack.
- **Conceptual Modeling-based** approaches and techniques applied to this challenging domain should guide the efforts to succeed

And more and more challenges to be explored…

- Linking diseases with genes with therapeutical purposes as a main application
- Gene mutations that enforce expression of some other genes while delaying or reducing the expression of others
- Gene regulators

ProS
Centro de Investigación
en Métodos de
Producción de Software

Conclusions

Una polla xica, pica, pellarica, camatorta i becarica...

Immune system **Base pair** **Protein** **Transcribed sequence** RDF

Transcription Genetic influences on female infidelity

Ontology **Exon** Human **Gene** **Conceptual Modeling-based**

Cytosine Cell **RNA polymerase** **Diagnosis** Conceptual model

Terminator **Chromosome** **Transcription unit** **Mutation** OO-Method

Genes against the malaria

Allele **Experiment** **ORF** Gene Ontology **Promoter** **Guanine**

Nature versus nurture

Centromere **Intron** **Neutral polimorphism** **Regulator sequence**

Aminoacid **DNA** GenoCAD **Widt type** **Chromosomic mutation**

Data bank **Proteone**

OWL **Primary polipeptide** BioBricks **ORI**

Inheritance **Genome** **Allelic variant** **External identification**

Hydrogene bonds **Telomere**

Spliced transcript An ‘infidelity' gene for men **Thymine**

Exon skipping **Intergenic region** HUGO **Research centre**

Pre-implant genetic diagnosis **Enhanced sequence**

Nucleotides **Mitocondrial genome** **Embryo** **Ambient** **Adenine**

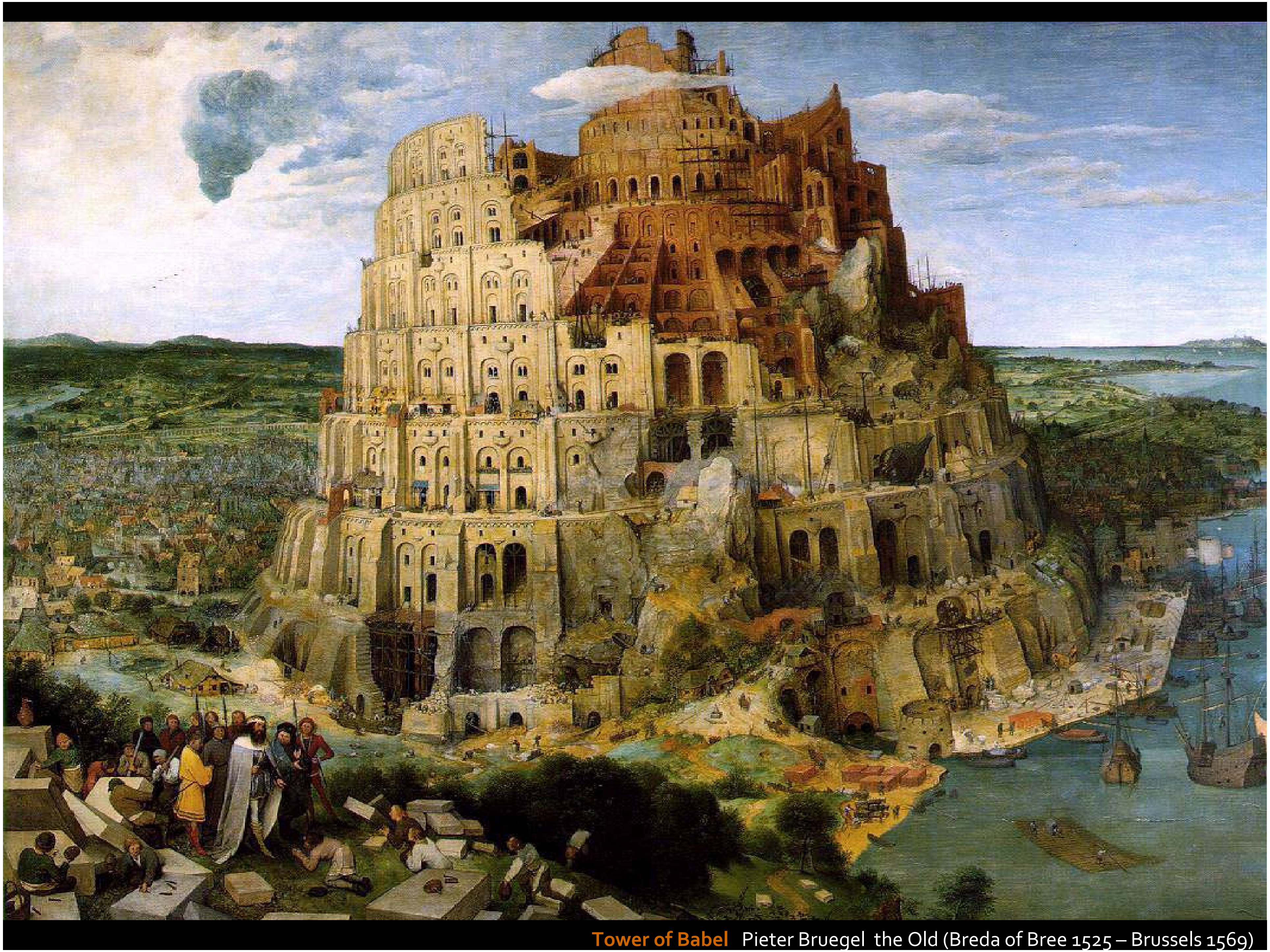
Vertebrate Genome Annotation Entrez Gene

Genic mutant Repairing genetic mutations with lasers

Codon **mRNA** 'Fat' gene makes greedy

Major groove Human Gene Mutation Database

UNIVERSIDAD
POLITECNICA
DE VALENCIA

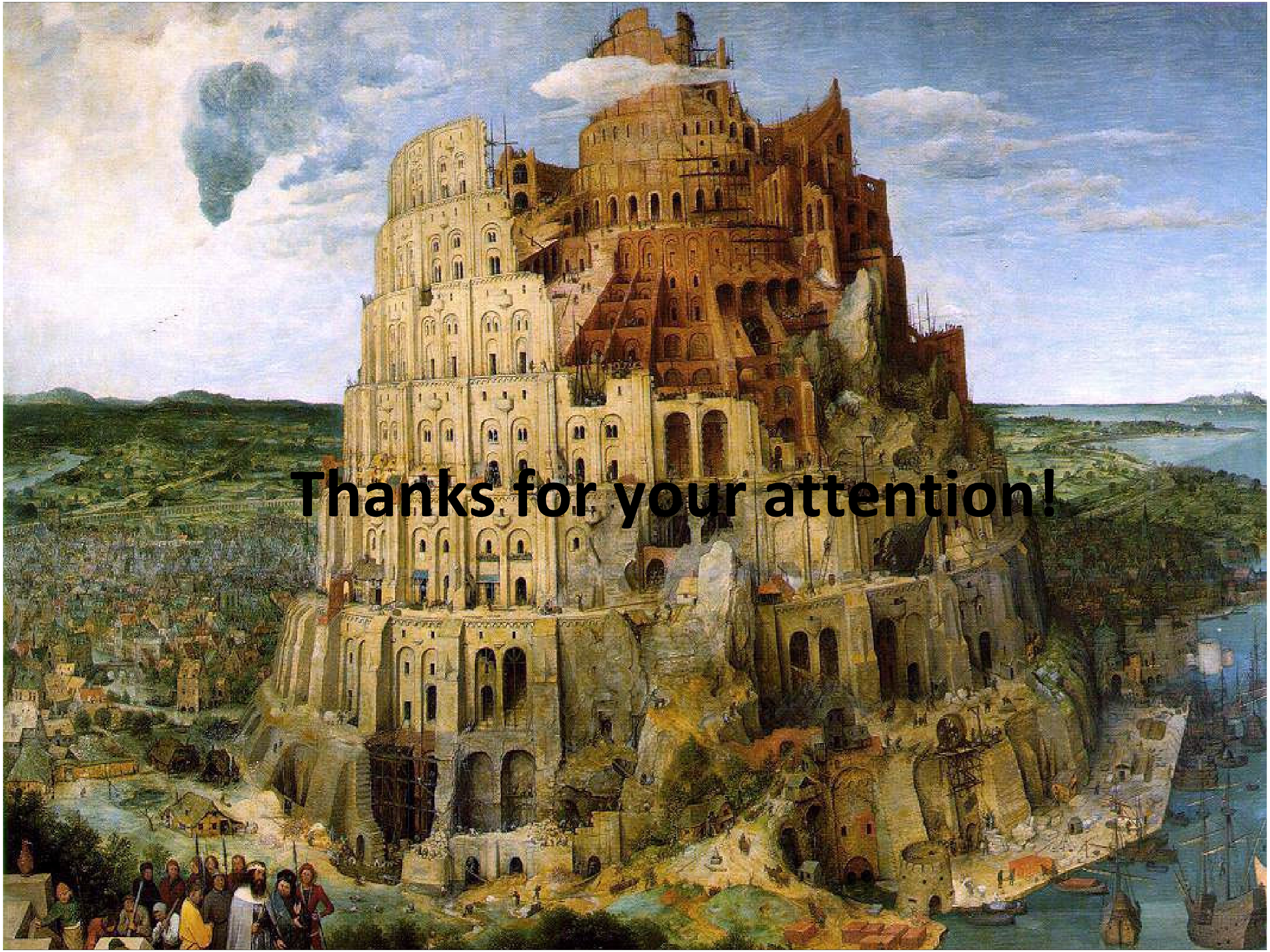


Tower of Babel Pieter Bruegel the Old (Breda of Bree 1525 – Brussels 1569)

Conclusions

- This is probably the most attractive challenge in the future of the Conceptual Modeling community:

Modeling the Real Life to understand why we are as we are, and how a human being can be seen as the "representation" of a Conceptual Model that can be specified in detail



Thanks for your attention!