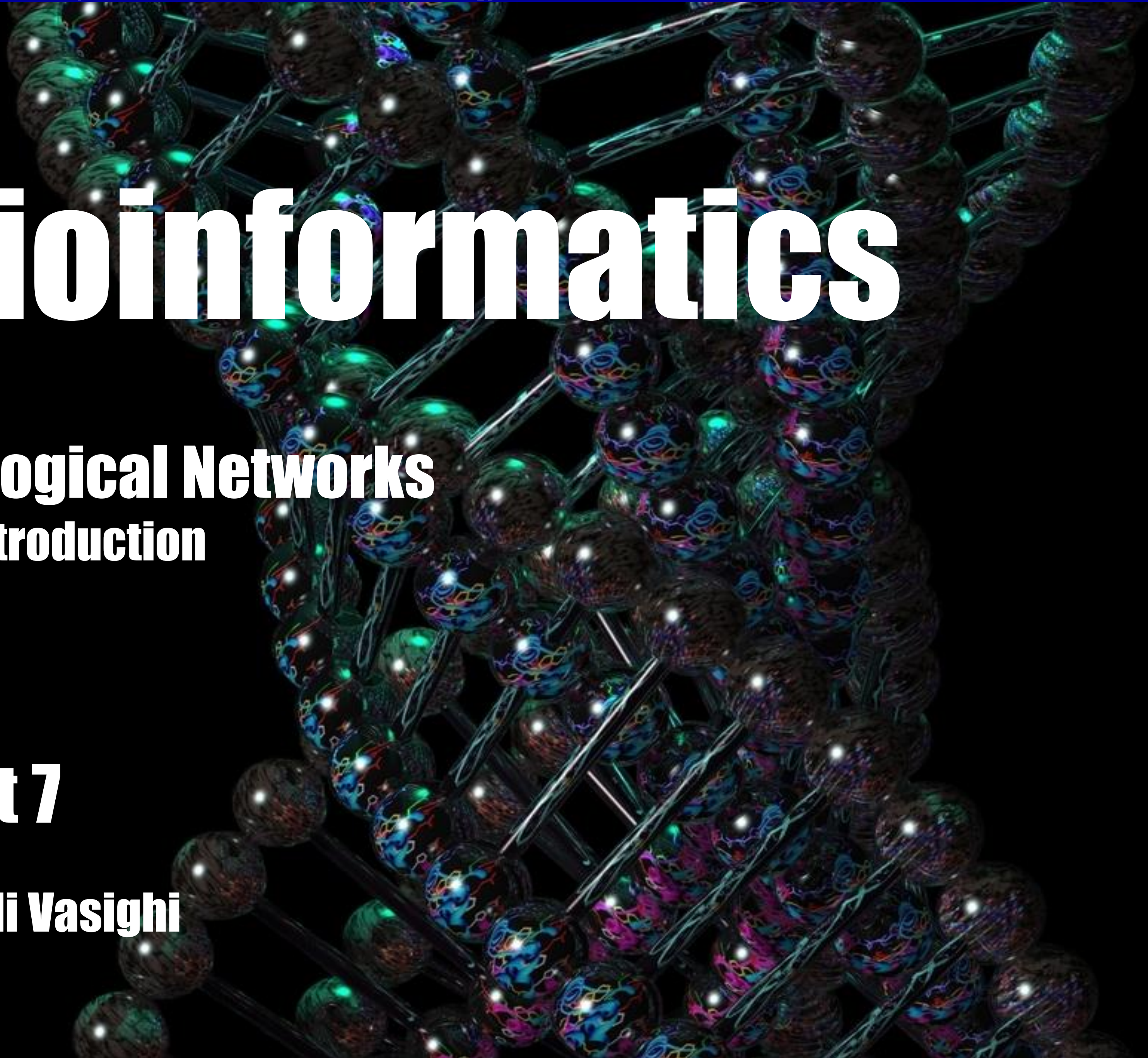
Institute for Advanced Studies in Basic Sciences – Zanjan
Department of Computer Science and Information Technology

By: Mahdi Vasighi



Bioinformatics

Biological Networks
An introduction

Part 7

Mahdi Vasighi

Gene Regulation

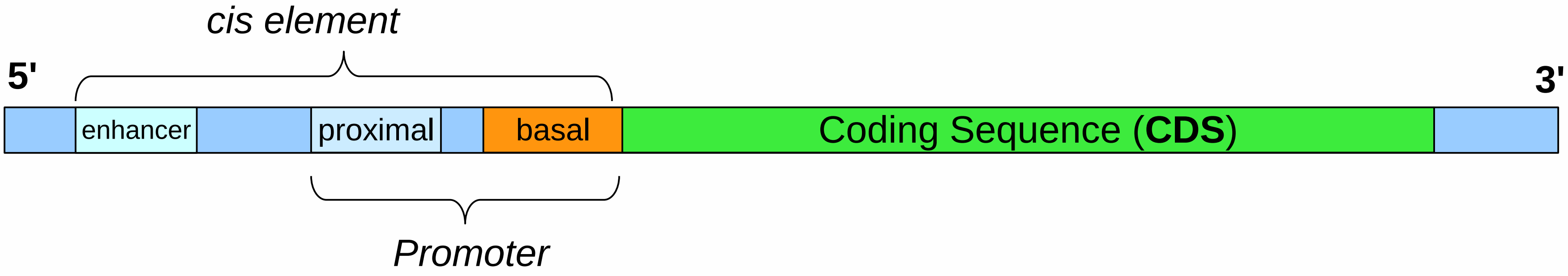
If transcription and translation were only dependent on random collisions thousands, if not millions, of times within the nucleus (transcription) and at ribosomes (translation), there would be no cell differentiation and organisms would be shapeless and lack structure.

The latest estimates are that a human cell, a eukaryotic cell, contains approximately 35,000 genes.

- **housekeeping genes** are expressed in all cells all the time.
- Some are expressed all the time in only those cells that have **differentiated** in a particular way.
- Some are expressed only as conditions around and in the cell change. For example, the **arrival of a hormone** may turn on (or off) certain genes in that cell.

- The gene must have Exon; start signals; stop signals; regulatory control elements

Gene Regulation


cis element

5'

3'

enhancer

proximal

basal

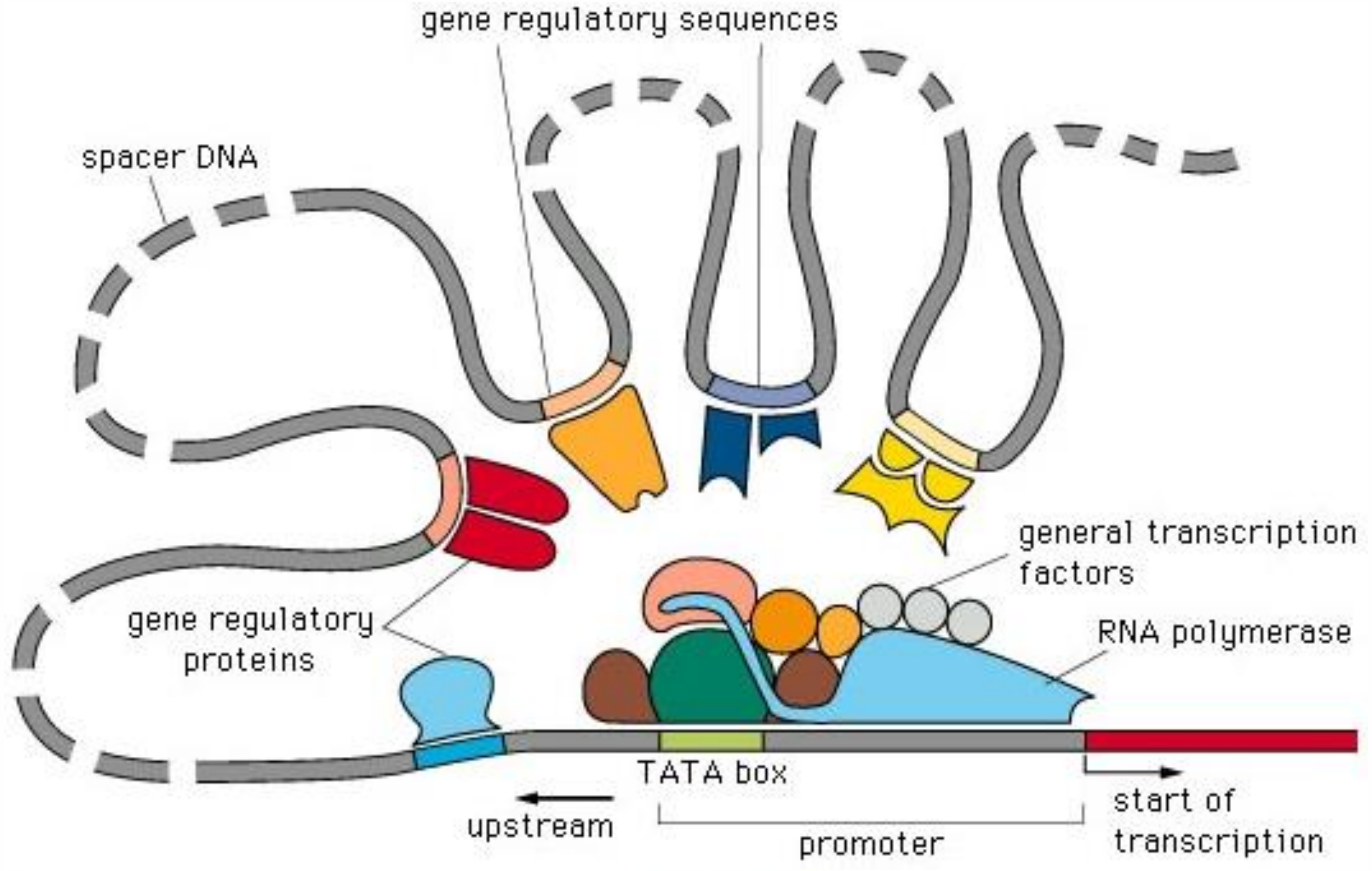
Coding Sequence (**CDS**)

Promoter

- **The basal (core) promoter**: provides recognition and binding sites for the RNAP II and is located about 40 base pairs (bps) from the start of the gene. It is typically contains a sequence of seven bases (TATAAAA, the '**TATA box**').
- **Proximal promoter** tends to contain primary regulatory elements
- **Enhancer** (and/or **Silencer**) serve to activate or repress transcription.
- **promoters** and **enhancers** form a DNA sequence, called a ***cis element***, which is recognized by regulatory proteins, called a ***trans element***.

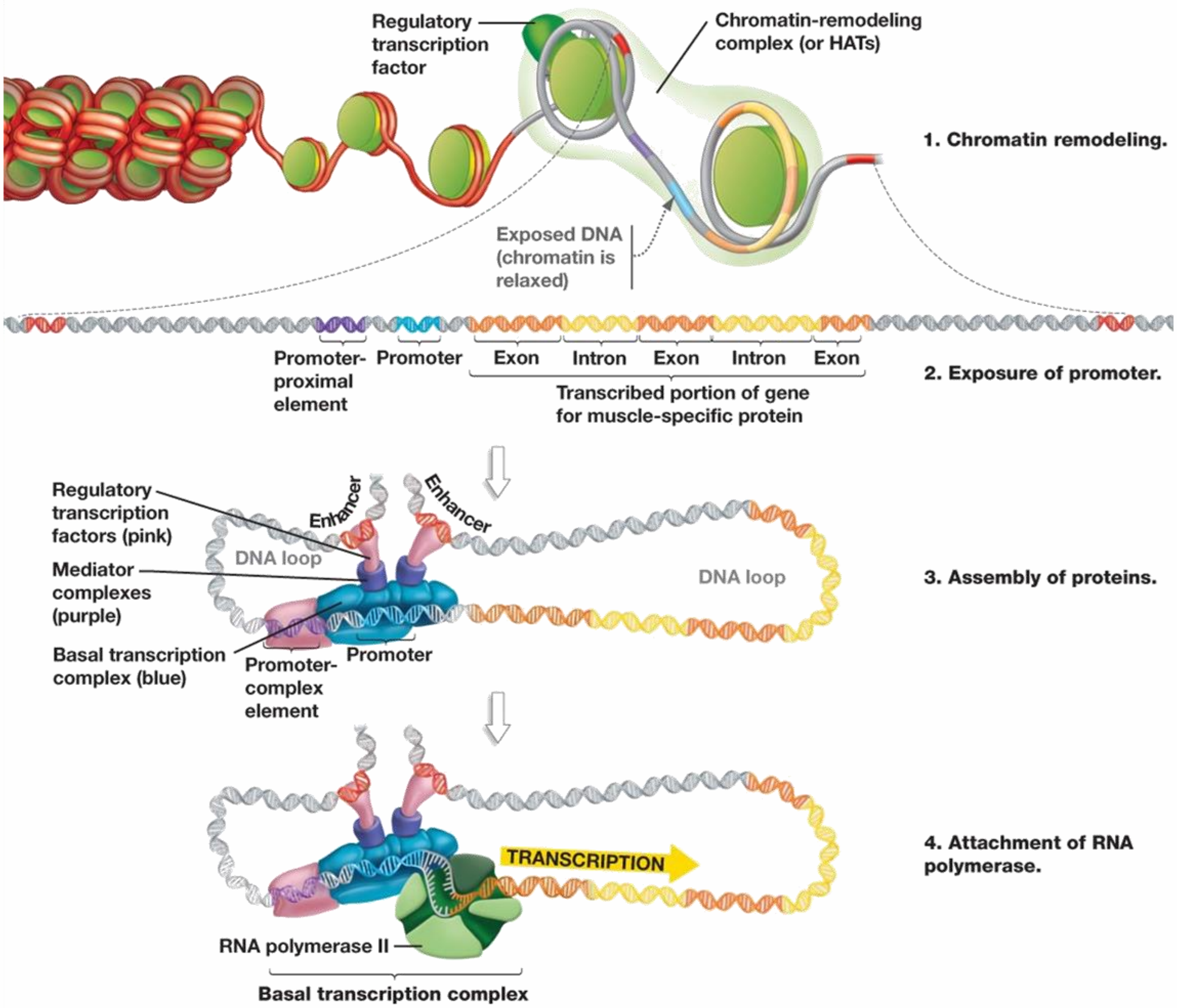
- This still leaves the question of where the first trans element comes from.

Gene Regulation



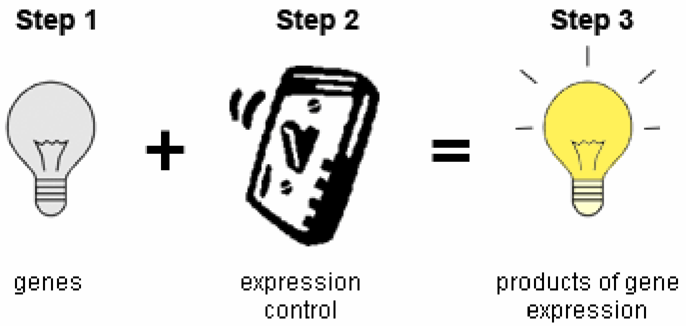
http://pbil.univ-lyon1.fr/members/sagot/htdocs/team/projects/chromo_net/images/reg.jpg

Gene Regulation



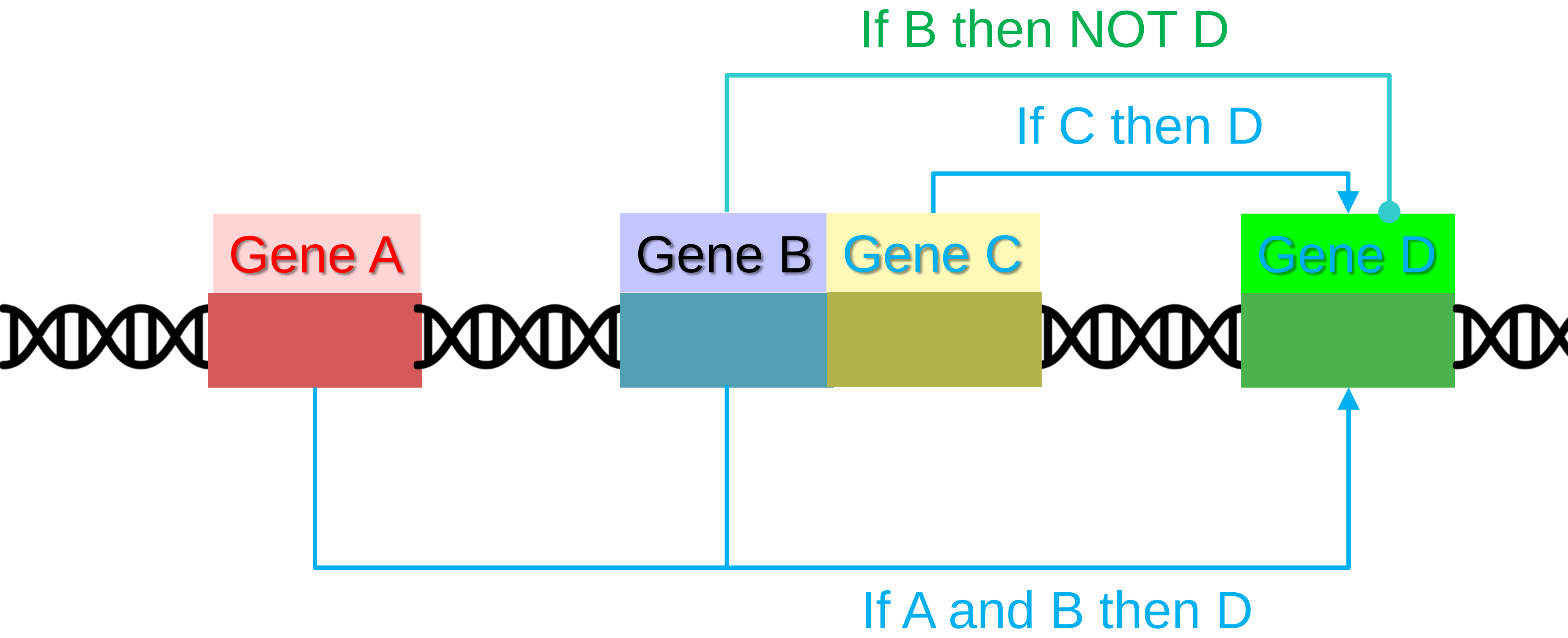
Gene Regulatory Network

- Gene regulatory networks (GRNs) are the on-off switches of a cell operating at the gene level.



- Two genes are connected if the expression of one gene modulates expression of another one by either activation or inhibition

Gene Regulatory Network



If B then NOT D

If C then D

Gene A

Gene B

Gene C

Gene D

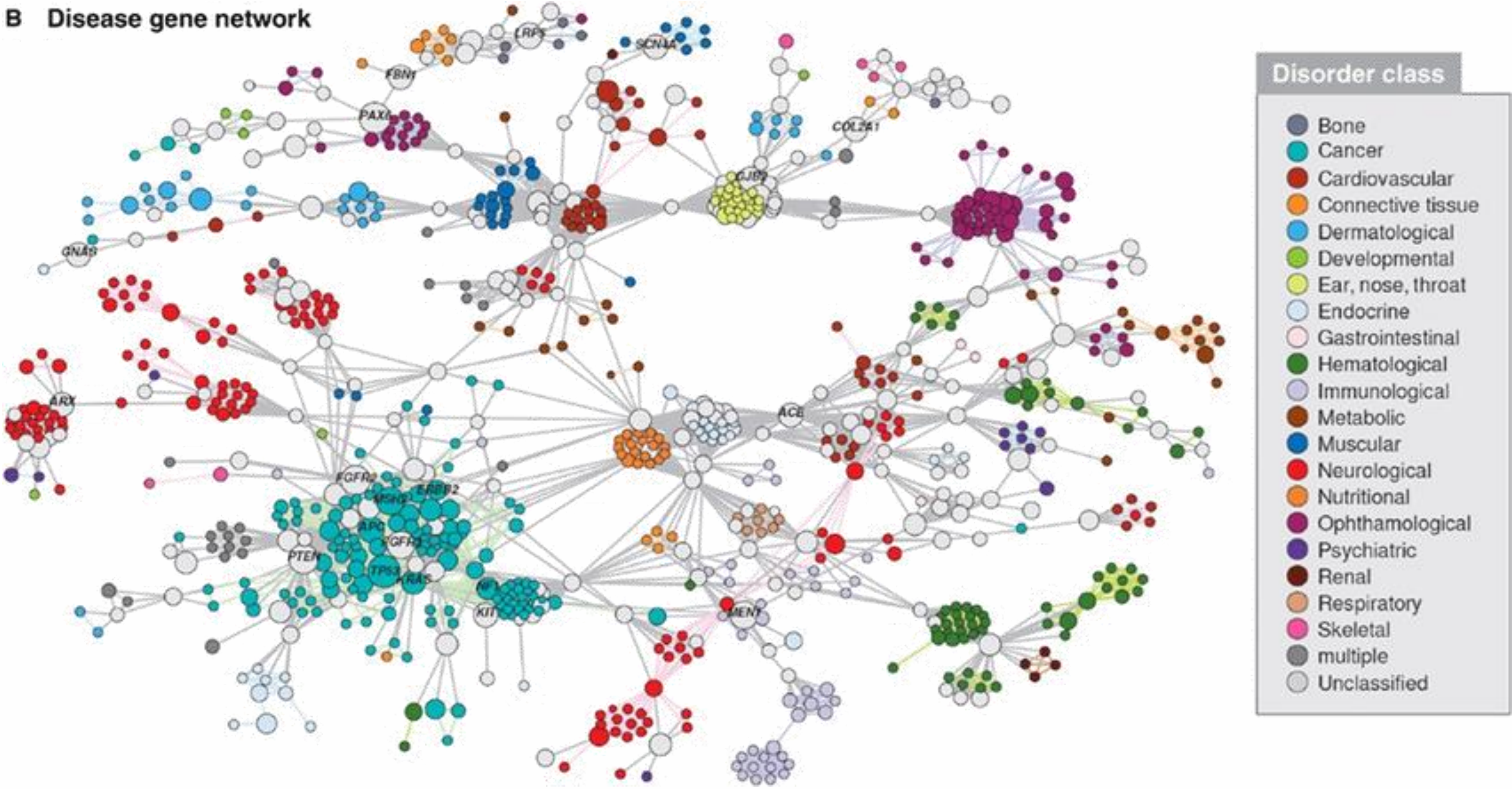
If A and B then D

Gene Regulatory Network

Why Study GRNs?

- Genes are **not independent**
  - They regulate each other and act collectively;
  - This collective behavior can be observed using microarray
- Some genes **control the response** of the cell to changes in the environment by regulating other genes
- Potential discovery of **triggering mechanism** and **treatments for disease**

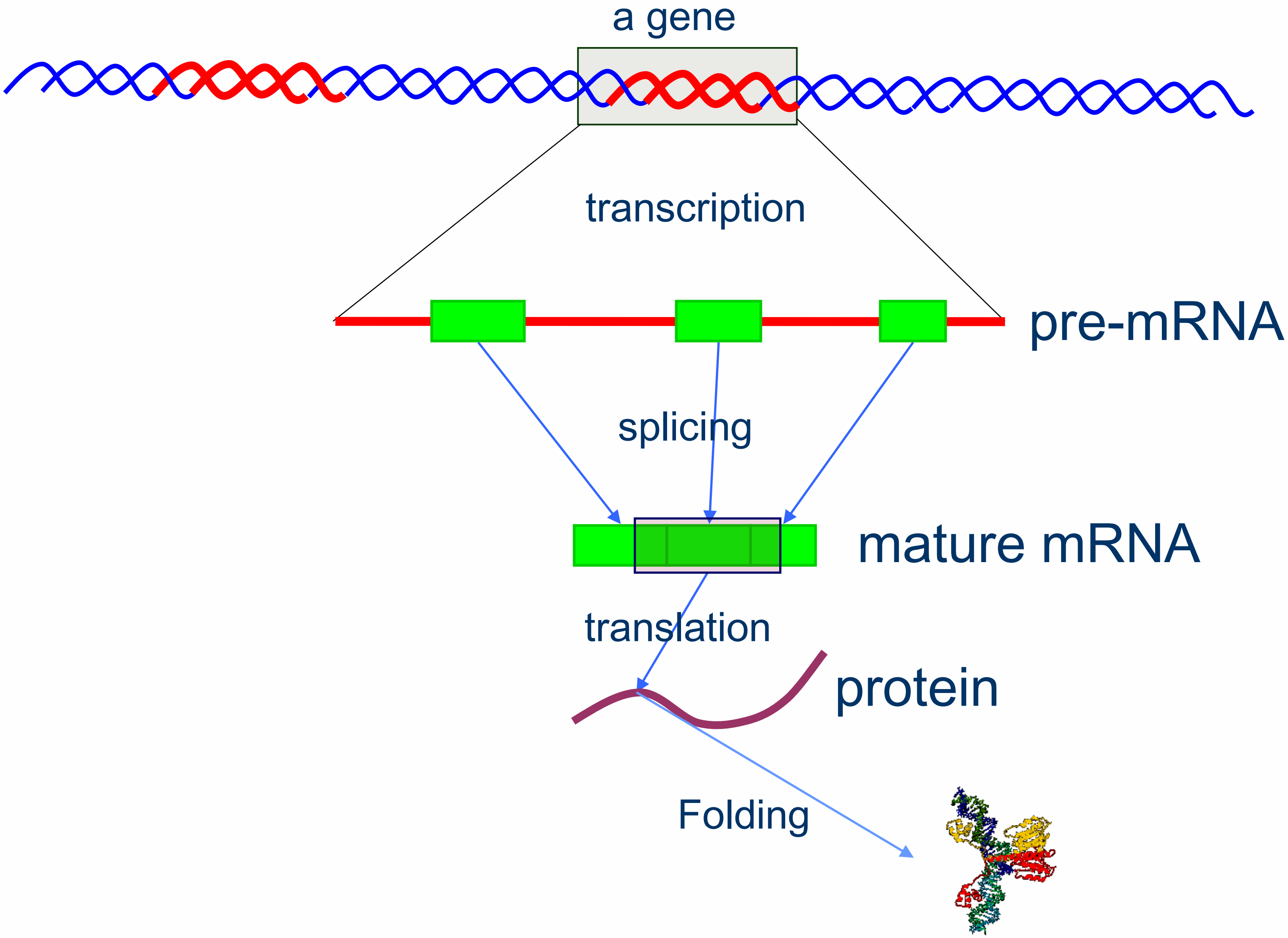
Gene Regulatory Network



B Disease gene network

Molecular Systems Biology (2007) 3, 124

Sequence story



a gene

transcription

pre-mRNA

splicing

mature mRNA

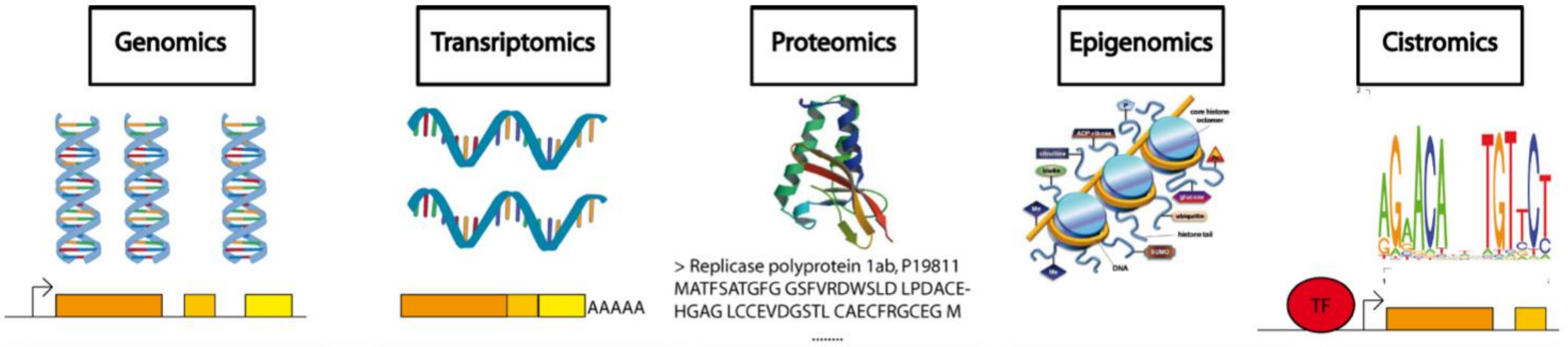
translation

protein

Folding

Sequence story

Schematic representation of the five omics technologies



> Replicase polyprotein 1ab, P19811
MATFSATGFG GSFVRDWSLD LPDACE-
HGAG LCCEVDGSTL CAECFRGCEG M
....

Genes sequences

mRNA sequences
mRNA expression level

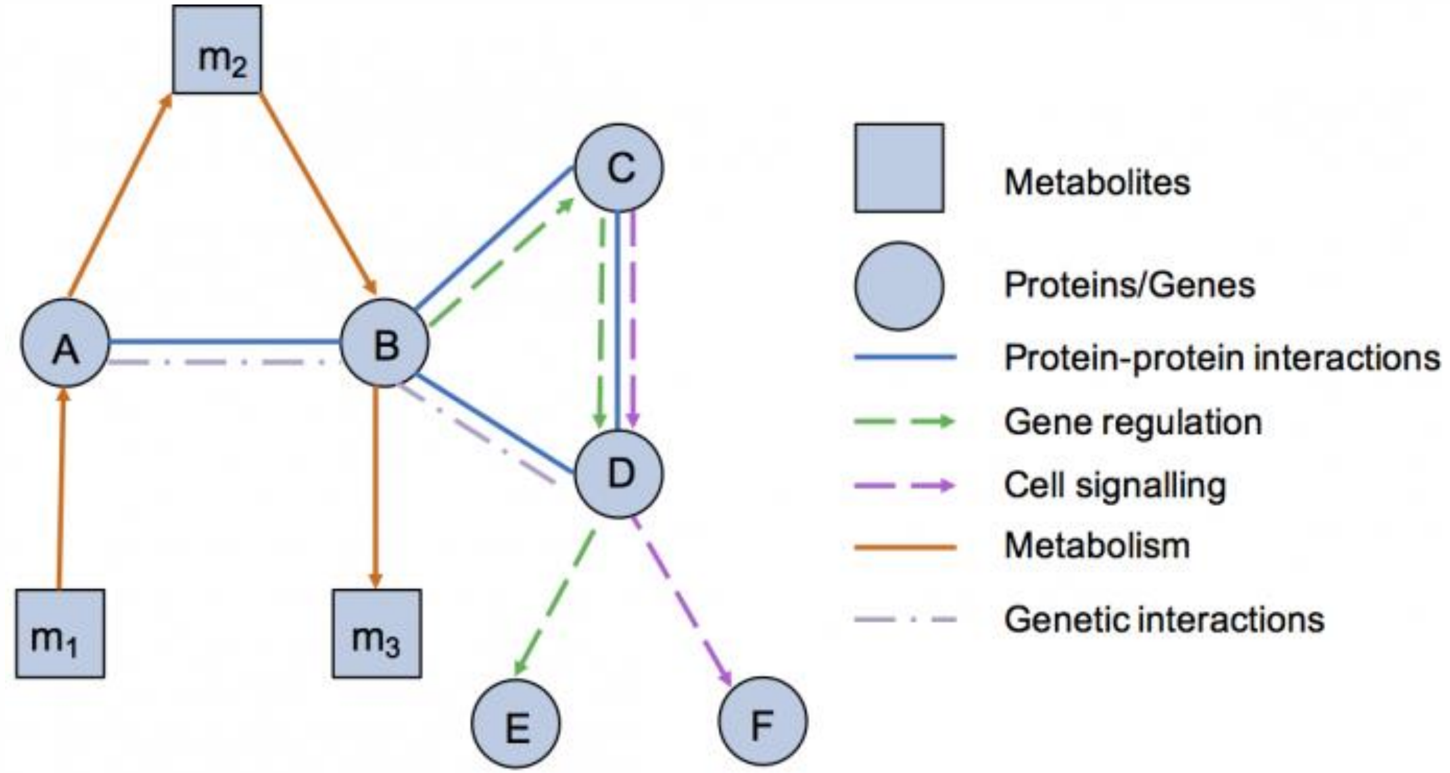
Protein sequences

Histones modifications
DNA methylation

Genomic repertoire of the binding sequences of a transcriptome factor

Biological networks

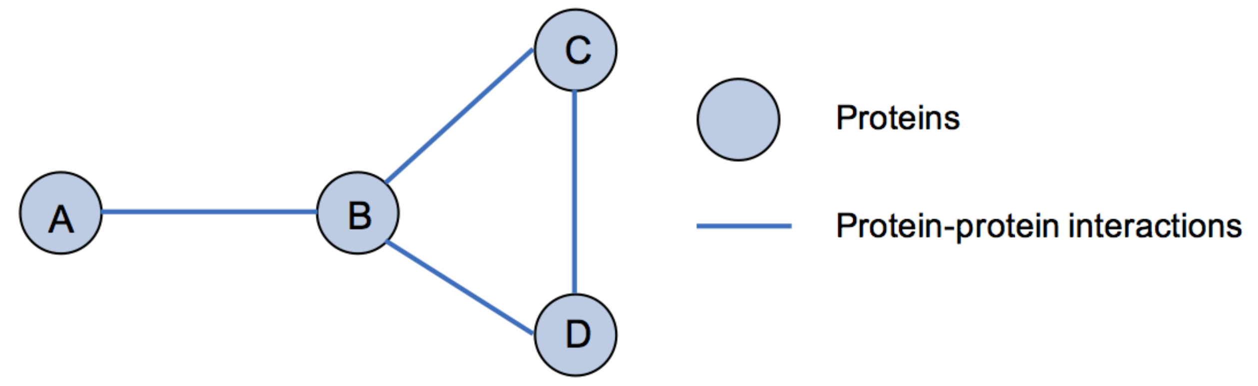
Different types of information can be represented in the shape of networks in order to model the cell

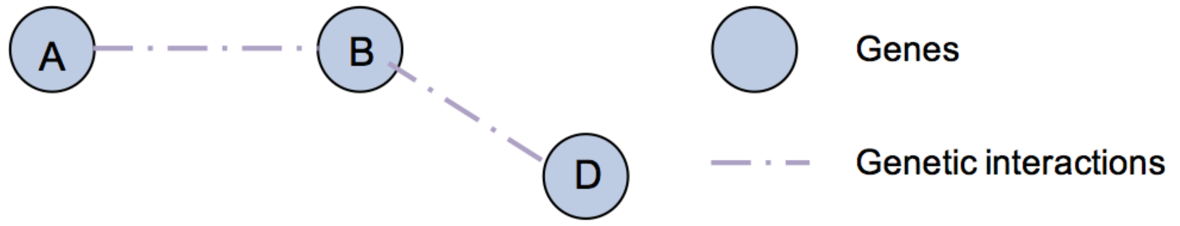


https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/networks-cell-biology-summary-0

Biological networks

Different types of information can be represented in the shape of networks in order to model the cell

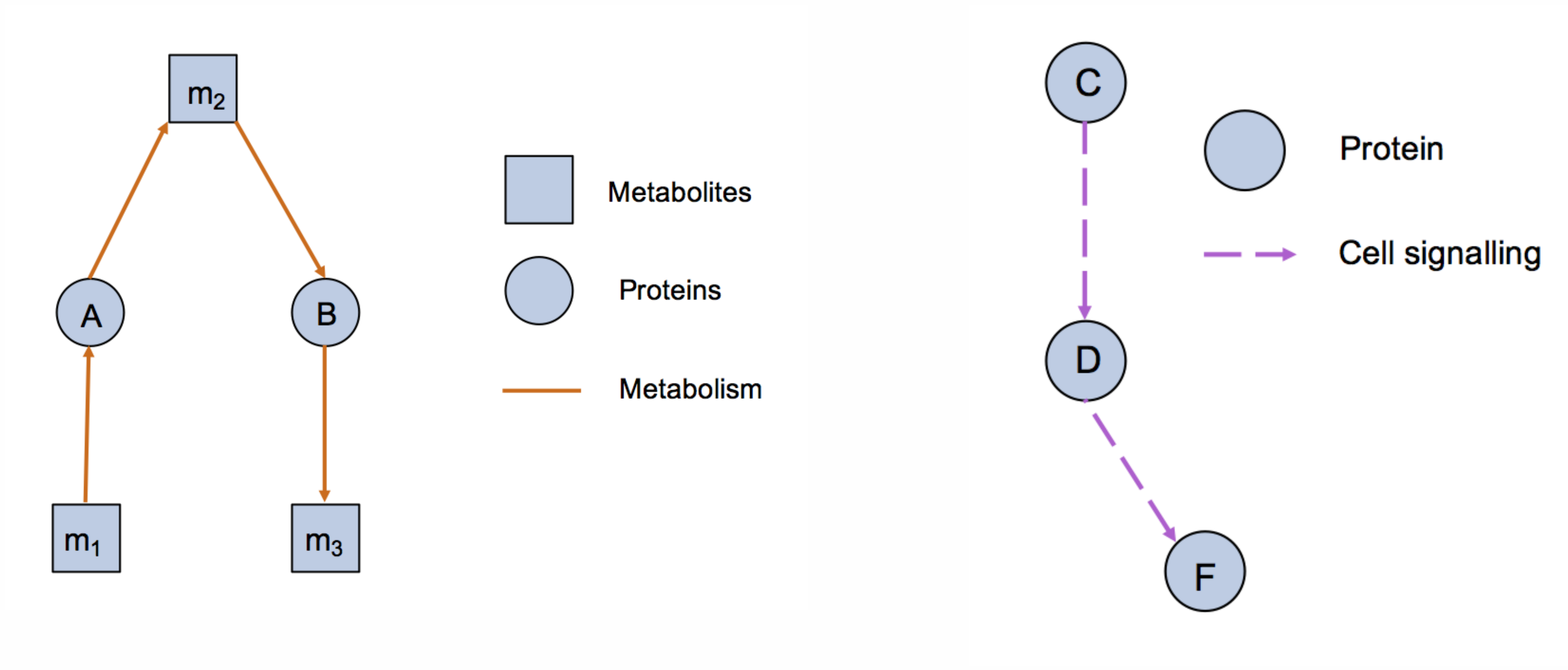




https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/networks-cell-biology-summary-0

Biological networks

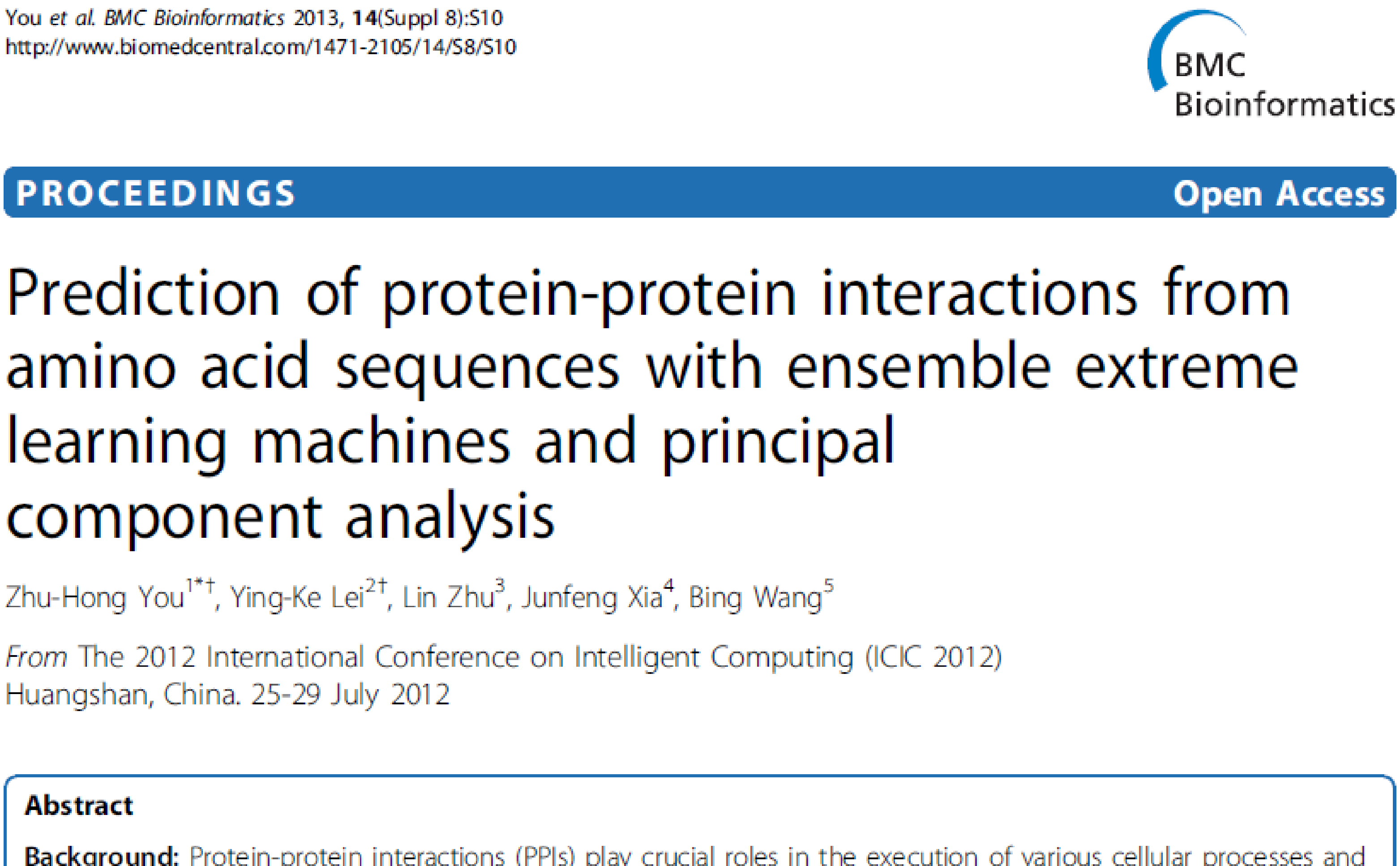
Different types of information can be represented in the shape of networks in order to model the cell



https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-introduction/networks-cell-biology-summary-0

Biological networks

http://www.biomedcentral.com/1471-2105/14/S8/S10



You *et al.* *BMC Bioinformatics* 2013, **14**(Suppl 8):S10
http://www.biomedcentral.com/1471-2105/14/S8/S10

BMC Bioinformatics

PROCEEDINGS **Open Access**

Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis

Zhu-Hong You[1*†], Ying-Ke Lei[2†], Lin Zhu[3], Junfeng Xia[4], Bing Wang[5]

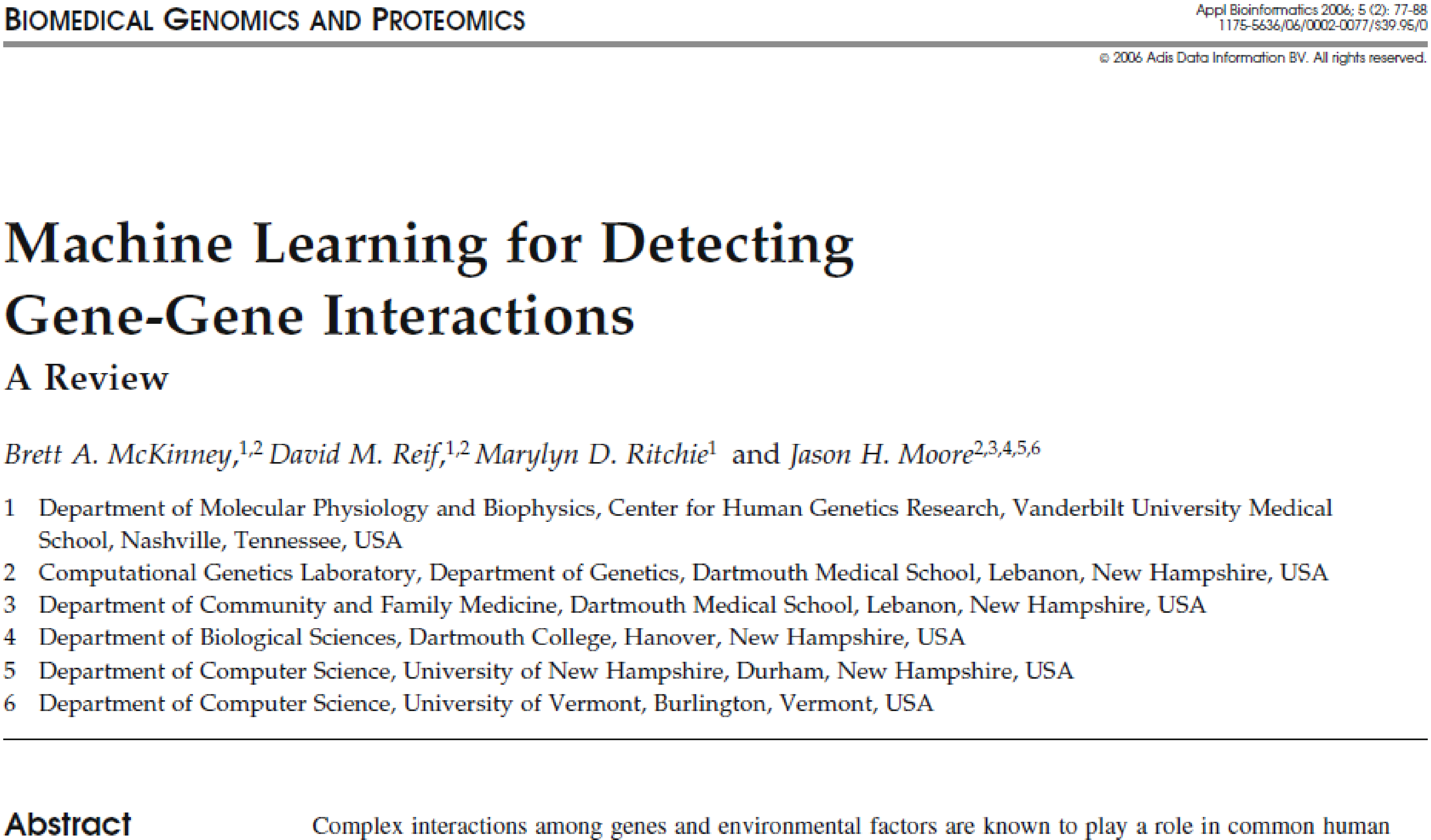
From The 2012 International Conference on Intelligent Computing (ICIC 2012)
Huangshan, China. 25-29 July 2012

Abstract

Background: Protein-protein interactions (PPIs) play crucial roles in the execution of various cellular processes and

Biological networks

Appl Bioinformatics 2006; 5 (2): 77-88



BIOMEDICAL GENOMICS AND PROTEOMICS

Appl Bioinformatics 2006; 5 (2): 77-88
1175-5636/06/0002-0077/$39.95/0

Machine Learning for Detecting Gene-Gene Interactions

A Review

Brett A. McKinney,[1,2] *David M. Reif,*[1,2] *Marylyn D. Ritchie*[1] and *Jason H. Moore*[2,3,4,5,6]

1 Department of Molecular Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University Medical School, Nashville, Tennessee, USA
2 Computational Genetics Laboratory, Department of Genetics, Dartmouth Medical School, Lebanon, New Hampshire, USA
3 Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, New Hampshire, USA
4 Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire, USA
5 Department of Computer Science, University of New Hampshire, Durham, New Hampshire, USA
6 Department of Computer Science, University of Vermont, Burlington, Vermont, USA

Abstract

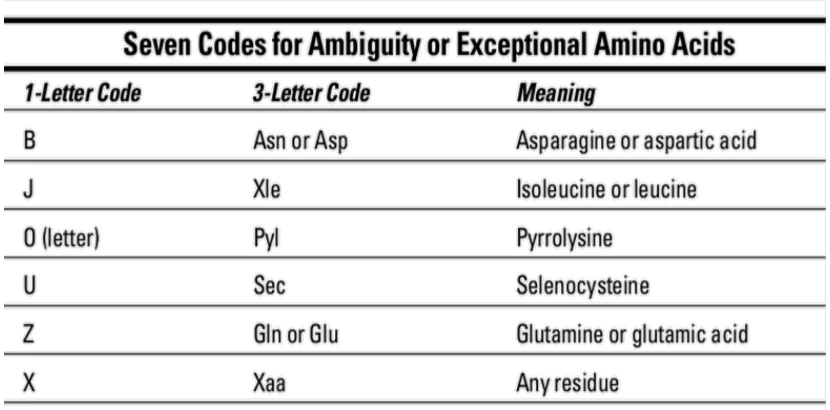
Complex interactions among genes and environmental factors are known to play a role in common human

Protein Sequences

Nomenclature and Symbolism for Amino Acids

#	1-Letter Code	3-Letter Code	Name
1	A	Ala	Alanine
2	R	Arg	Arginine
3	N	Asn	Asparagine
4	D	Asp	Aspartic acid
5	C	Cys	Cysteine
6	Q	Gln	Glutamine
7	E	Glu	Glutamic acid
8	G	Gly	Glycine
9	H	His	Histidine
10	I	Ile	Isoleucine
11	L	Leu	Leucine
12	K	Lys	Lysine
13	M	Met	Methionine
14	F	Phe	Phenylalanine
15	P	Pro	Proline
16	S	Ser	Serine
17	T	Thr	Threonine
18	W	Trp	Tryptophan
19	Y	Tyr	Tyrosine
20	V	Val	Valine

by IUPAC (International Union of Pure and Applied Chemistry)

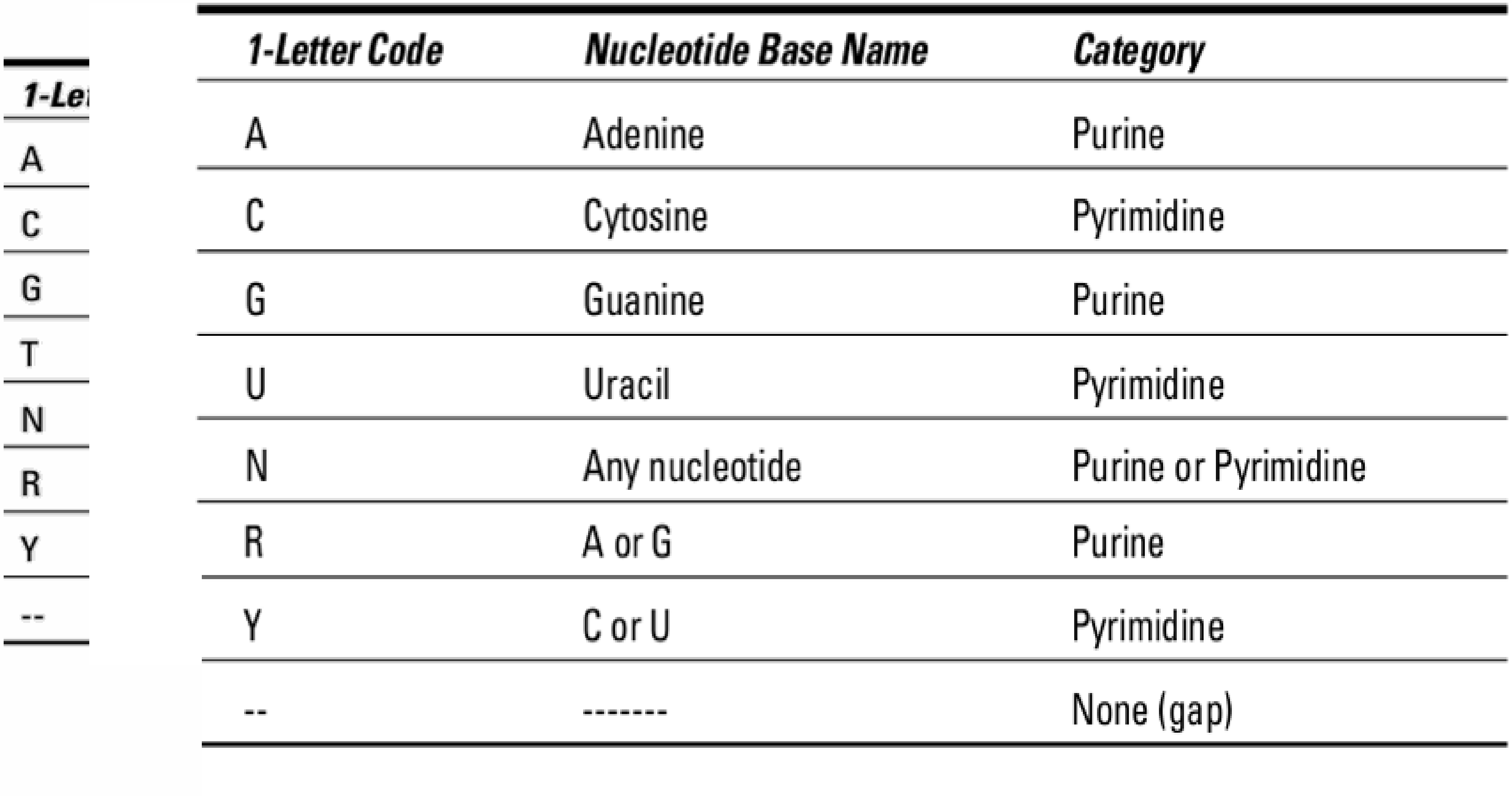


Seven Codes for Ambiguity or Exceptional Amino Acids

1-Letter Code	3-Letter Code	Meaning
B	Asn or Asp	Asparagine or aspartic acid
J	Xle	Isoleucine or leucine
O (letter)	Pyl	Pyrrolysine
U	Sec	Selenocysteine
Z	Gln or Glu	Glutamine or glutamic acid
X	Xaa	Any residue

A protein fragment defines as the succession of its constituent amino acids, listed in order from the **N-terminus** to the **C-terminus**.

DNA Sequences

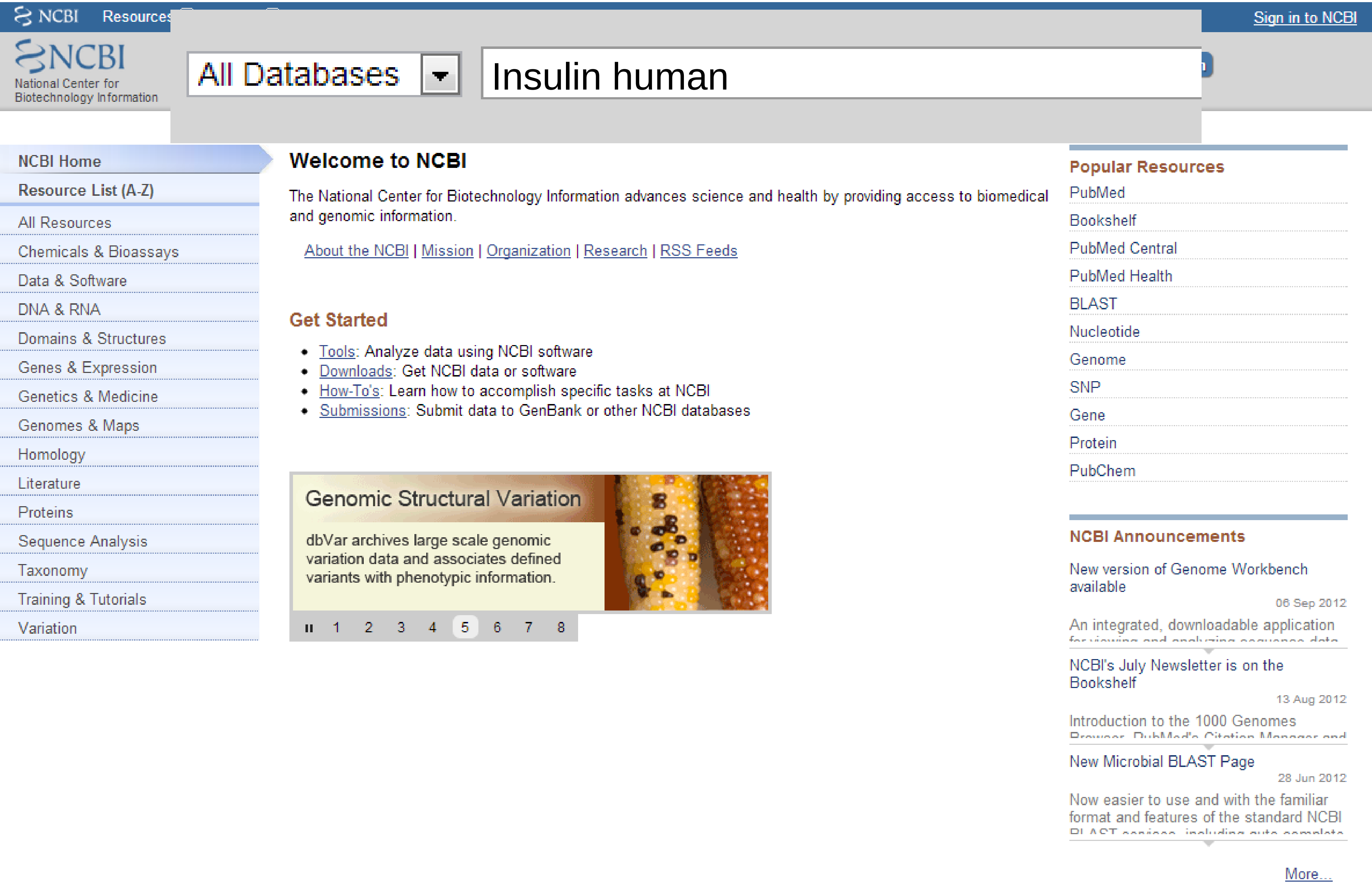


1-Letter Code	Nucleotide Base Name	Category
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
U	Uracil	Pyrimidine
N	Any nucleotide	Purine or Pyrimidine
R	A or G	Purine
Y	C or U	Pyrimidine
--	-------	None (gap)

1-Le
A
C
G
T
N
R
Y
--

By convention, the primary structure of a DNA or RNA molecule is reported from the 5' end to the 3' end.

http://www.ncbi.nlm.nih.gov



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genomic Structural Variation

dbVar archives large scale genomic variation data and associates defined variants with phenotypic information.

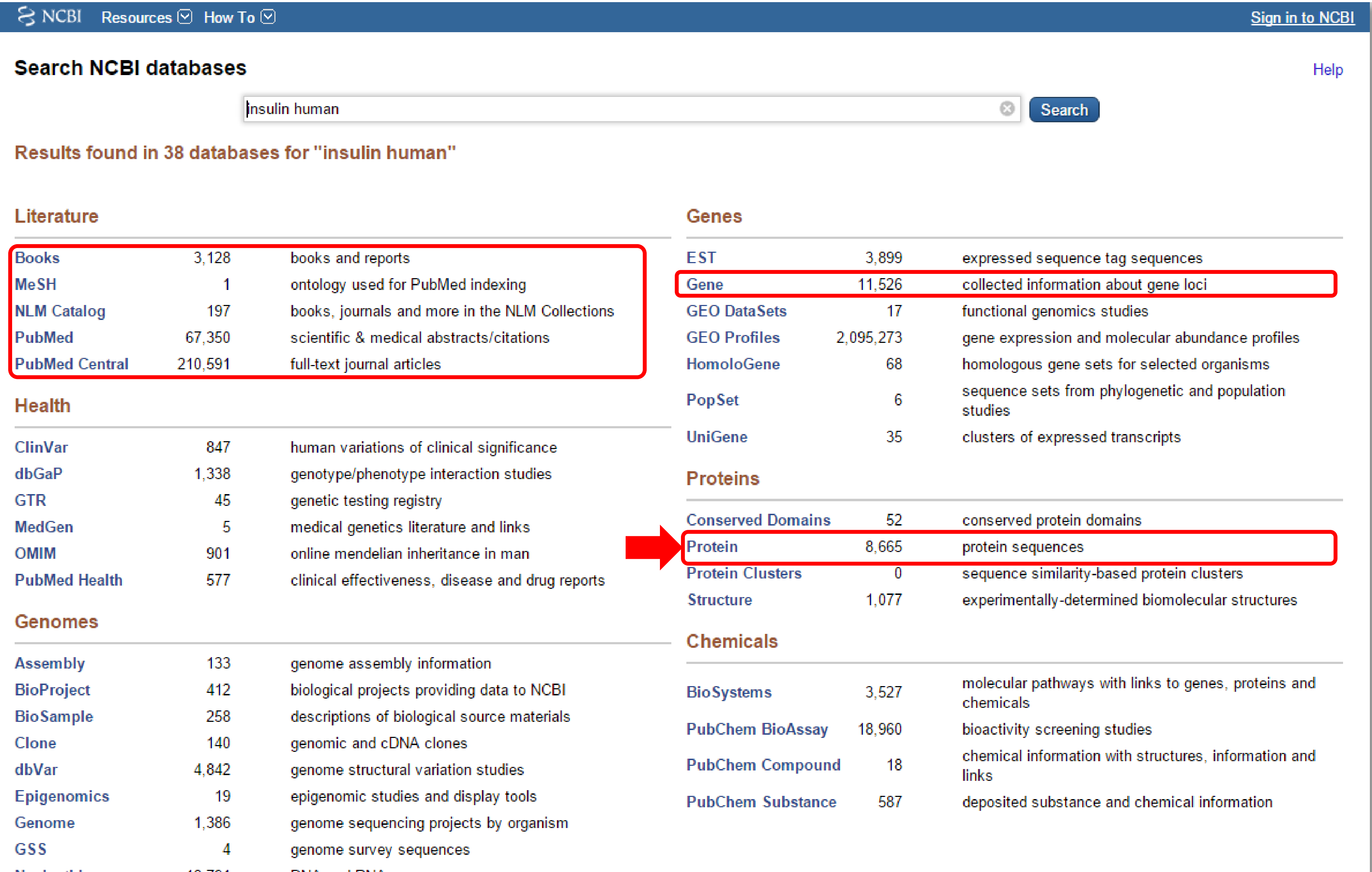
NCBI Announcements

New version of Genome Workbench available
06 Sep 2012
An integrated, downloadable application

NCBI's July Newsletter is on the Bookshelf
13 Aug 2012
Introduction to the 1000 Genomes

New Microbial BLAST Page
28 Jun 2012
Now easier to use and with the familiar format and features of the standard NCBI

http://www.ncbi.nlm.nih.gov



Search NCBI databases

Results found in 38 databases for "insulin human"

Literature

Books	3,128	books and reports
MeSH	1	ontology used for PubMed indexing
NLM Catalog	197	books, journals and more in the NLM Collections
PubMed	67,350	scientific & medical abstracts/citations
PubMed Central	210,591	full-text journal articles

Health

ClinVar	847	human variations of clinical significance
dbGaP	1,338	genotype/phenotype interaction studies
GTR	45	genetic testing registry
MedGen	5	medical genetics literature and links
OMIM	901	online mendelian inheritance in man
PubMed Health	577	clinical effectiveness, disease and drug reports

Genomes

Assembly	133	genome assembly information
BioProject	412	biological projects providing data to NCBI
BioSample	258	descriptions of biological source materials
Clone	140	genomic and cDNA clones
dbVar	4,842	genome structural variation studies
Epigenomics	19	epigenomic studies and display tools
Genome	1,386	genome sequencing projects by organism
GSS	4	genome survey sequences

Genes

EST	3,899	expressed sequence tag sequences
Gene	11,526	collected information about gene loci
GEO DataSets	17	functional genomics studies
GEO Profiles	2,095,273	gene expression and molecular abundance profiles
HomoloGene	68	homologous gene sets for selected organisms
PopSet	6	sequence sets from phylogenetic and population studies
UniGene	35	clusters of expressed transcripts

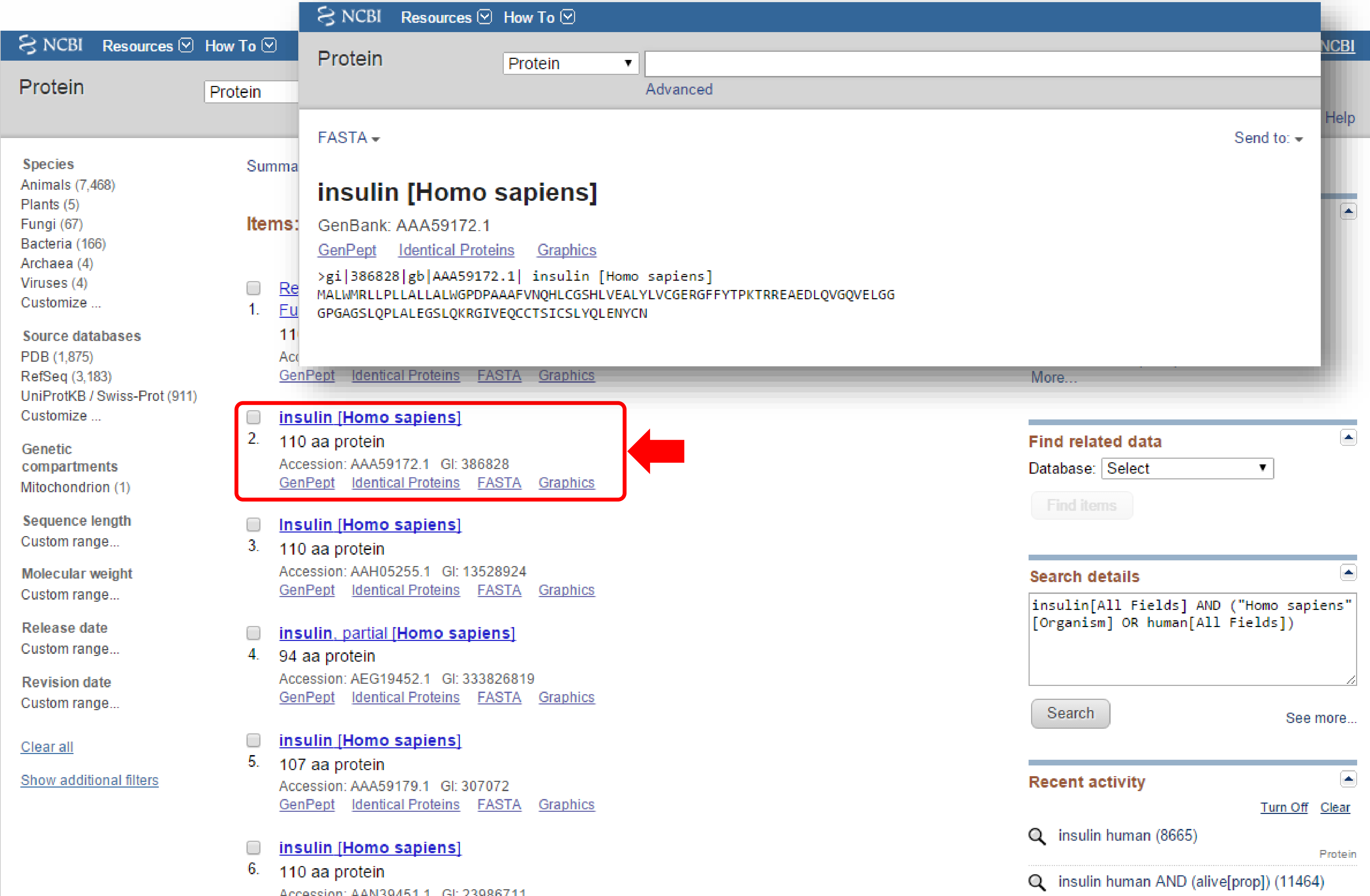
Proteins

Conserved Domains	52	conserved protein domains
Protein	8,665	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	1,077	experimentally-determined biomolecular structures

Chemicals

BioSystems	3,527	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	18,960	bioactivity screening studies
PubChem Compound	18	chemical information with structures, information and links
PubChem Substance	587	deposited substance and chemical information

http://www.ncbi.nlm.nih.gov



insulin [Homo sapiens]

GenBank: AAA59172.1

GenPept Identical Proteins Graphics

```
>gi|386828|gb|AAA59172.1| insulin [Homo sapiens]
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGG
GPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

☐ 1. Re
Fu
11
Ac
GenPept Identical Proteins FASTA Graphics

☐ 2. **insulin [Homo sapiens]**
110 aa protein
Accession: AAA59172.1 GI: 386828
GenPept Identical Proteins FASTA Graphics

☐ 3. **Insulin [Homo sapiens]**
110 aa protein
Accession: AAH05255.1 GI: 13528924
GenPept Identical Proteins FASTA Graphics

☐ 4. **insulin, partial [Homo sapiens]**
94 aa protein
Accession: AEG19452.1 GI: 333826819
GenPept Identical Proteins FASTA Graphics

☐ 5. **insulin [Homo sapiens]**
107 aa protein
Accession: AAA59179.1 GI: 307072
GenPept Identical Proteins FASTA Graphics

☐ 6. **insulin [Homo sapiens]**
110 aa protein
Accession: AAN39451.1 GI: 23986711

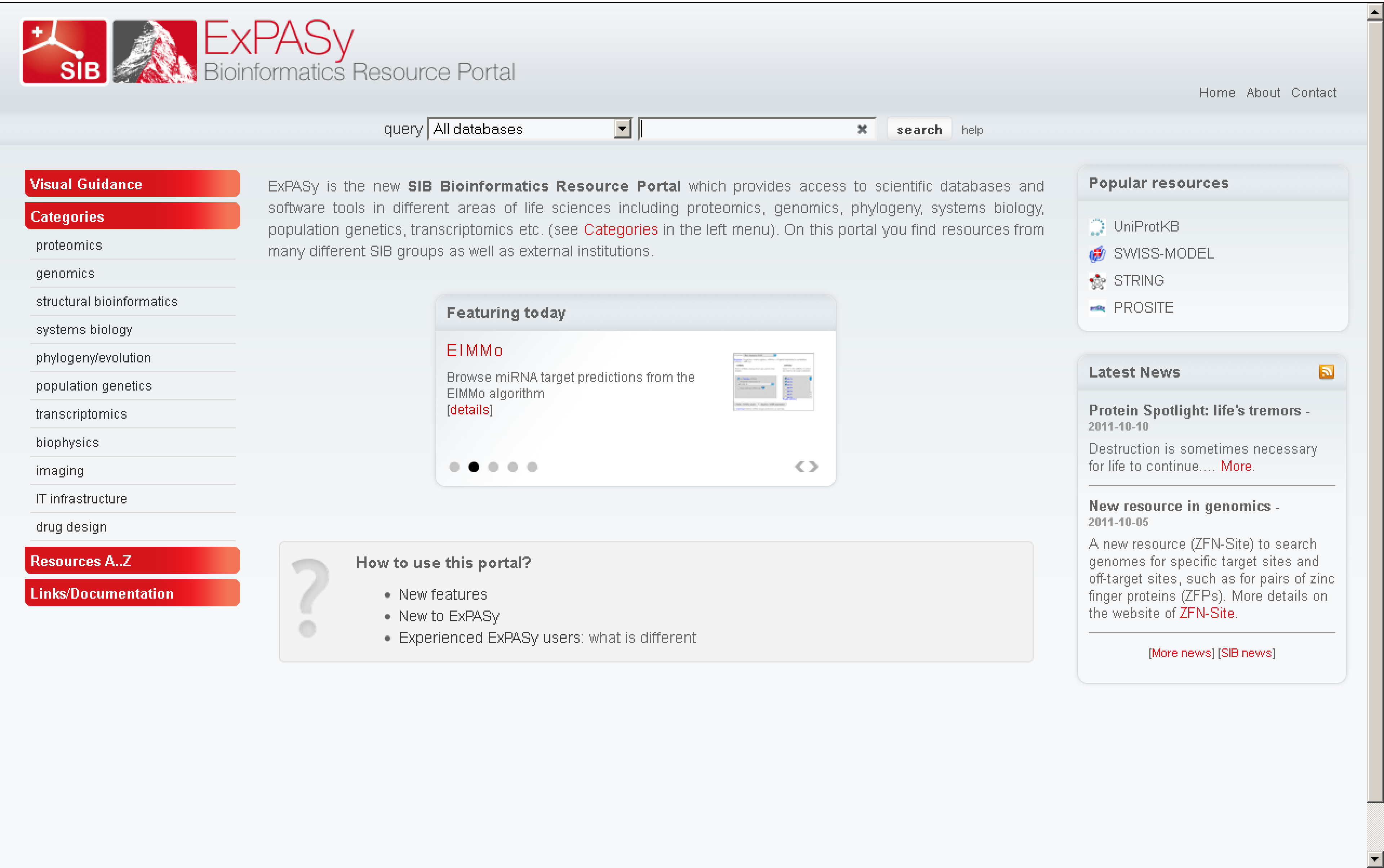
```
insulin[All Fields] AND ("Homo sapiens"
[Organism] OR human[All Fields])
```

insulin human (8665)

insulin human AND (alive[prop]) (11464)

Retrieving Protein Sequences

http://www.expasy.org/



ExPASy
Bioinformatics Resource Portal

Home About Contact

query All databases search help

ExPASy is the new **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

Featuring today

EIMMo

Browse miRNA target predictions from the ElMMo algorithm
[details]

How to use this portal?

- New features
- New to ExPASy
- Experienced ExPASy users: what is different

Popular resources

- UniProtKB
- SWISS-MODEL
- STRING
- PROSITE

Latest News

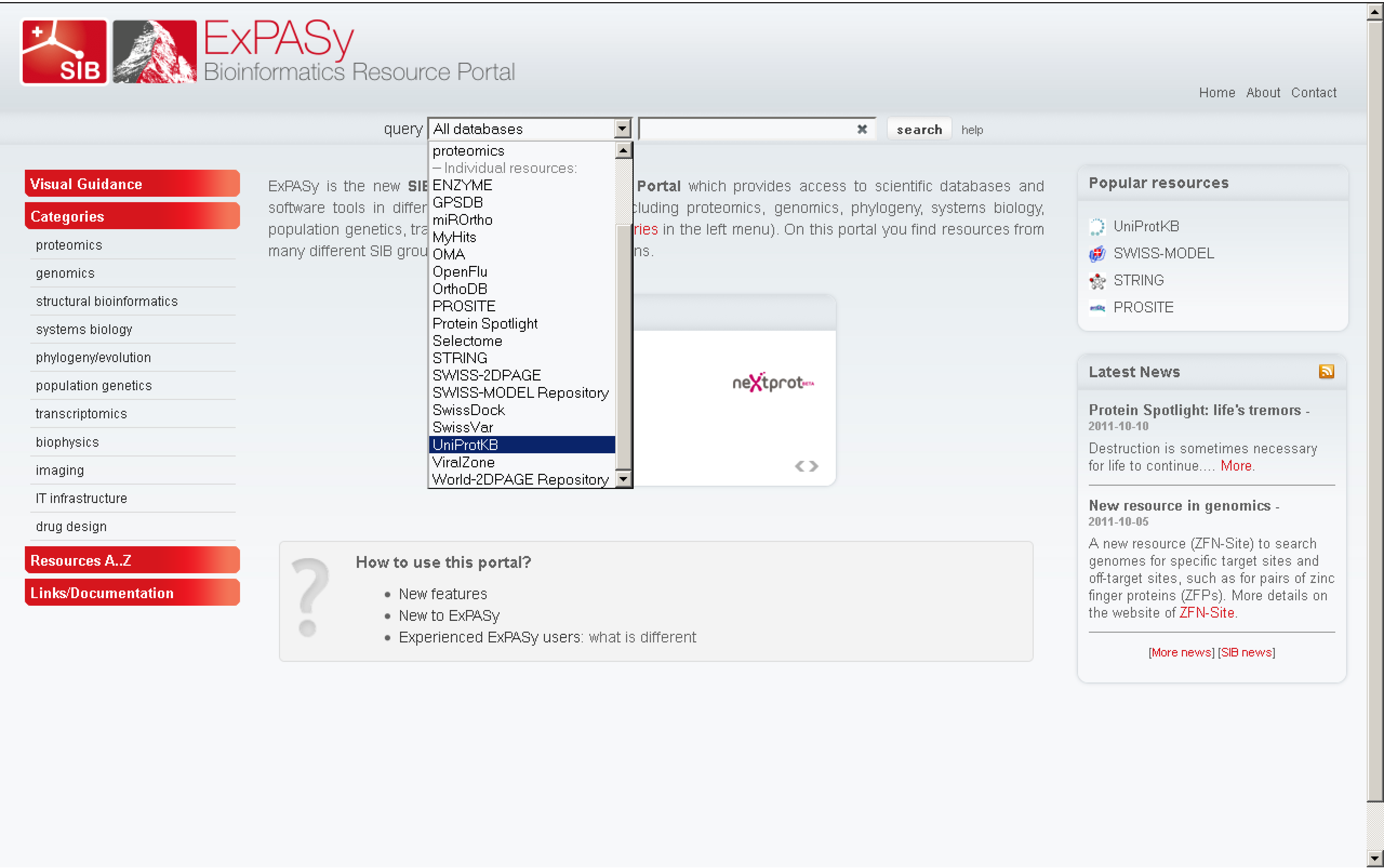
Protein Spotlight: life's tremors - 2011-10-10

Destruction is sometimes necessary for life to continue.... More.

New resource in genomics - 2011-10-05

A new resource (ZFN-Site) to search genomes for specific target sites and off-target sites, such as for pairs of zinc finger proteins (ZFPs). More details on the website of ZFN-Site.

[More news] [SIB news]



ExPASy is the new **SIB** … **Portal** which provides access to scientific databases and software tools in differ… …cluding proteomics, genomics, phylogeny, systems biology, population genetics, tra… …ries in the left menu). On this portal you find resources from many different SIB grou… …ns.

How to use this portal?

- New features
- New to ExPASy
- Experienced ExPASy users: what is different

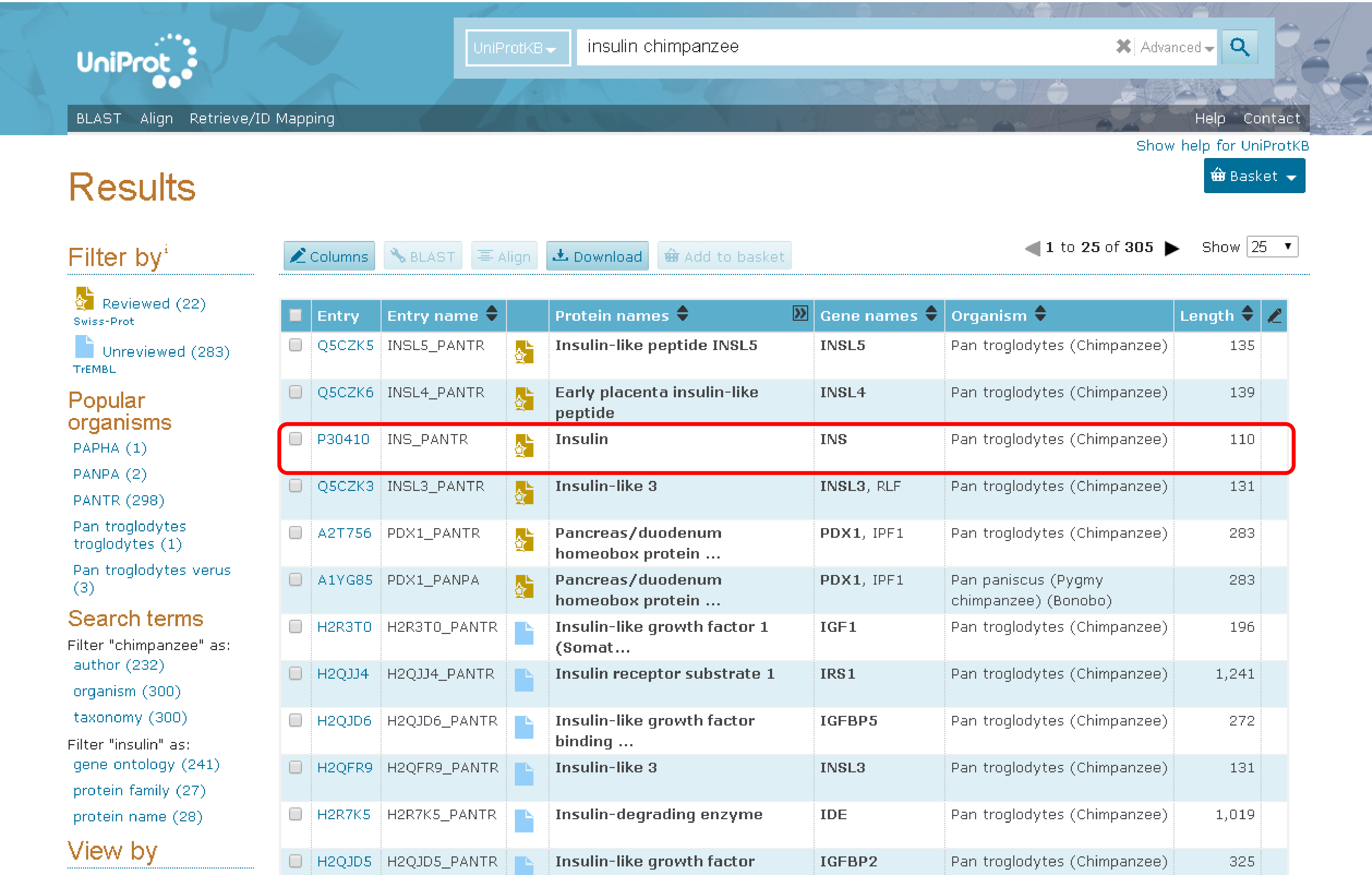
Protein Spotlight: life's tremors - 2011-10-10

Destruction is sometimes necessary for life to continue.... More.

New resource in genomics - 2011-10-05

A new resource (ZFN-Site) to search genomes for specific target sites and off-target sites, such as for pairs of zinc finger proteins (ZFPs). More details on the website of ZFN-Site.

[More news] [SIB news]



insulin chimpanzee

Results

1 to 25 of 305 Show 25

Filter by

Reviewed (22) Swiss-Prot

Unreviewed (283) TrEMBL

Popular organisms

PAPHA (1)

PANPA (2)

PANTR (298)

Pan troglodytes troglodytes (1)

Pan troglodytes verus (3)

Search terms

Filter "chimpanzee" as:

author (232)

organism (300)

taxonomy (300)

Filter "insulin" as:

gene ontology (241)

protein family (27)

protein name (28)

View by

☐	Entry	Entry name		Protein names	Gene names	Organism	Length
☐	Q5CZK5	INSL5_PANTR	★	**Insulin-like peptide INSL5**	**INSL5**	Pan troglodytes (Chimpanzee)	135
☐	Q5CZK6	INSL4_PANTR	★	**Early placenta insulin-like peptide**	**INSL4**	Pan troglodytes (Chimpanzee)	139
☐	P30410	INS_PANTR	★	**Insulin**	**INS**	Pan troglodytes (Chimpanzee)	110
☐	Q5CZK3	INSL3_PANTR	★	**Insulin-like 3**	**INSL3**, RLF	Pan troglodytes (Chimpanzee)	131
☐	A2T756	PDX1_PANTR	★	**Pancreas/duodenum homeobox protein ...**	**PDX1**, IPF1	Pan troglodytes (Chimpanzee)	283
☐	A1YG85	PDX1_PANPA	★	**Pancreas/duodenum homeobox protein ...**	**PDX1**, IPF1	Pan paniscus (Pygmy chimpanzee) (Bonobo)	283
☐	H2R3T0	H2R3T0_PANTR	🗋	**Insulin-like growth factor 1 (Somat...**	**IGF1**	Pan troglodytes (Chimpanzee)	196
☐	H2QJJ4	H2QJJ4_PANTR	🗋	**Insulin receptor substrate 1**	**IRS1**	Pan troglodytes (Chimpanzee)	1,241
☐	H2QJD6	H2QJD6_PANTR	🗋	**Insulin-like growth factor binding ...**	**IGFBP5**	Pan troglodytes (Chimpanzee)	272
☐	H2QFR9	H2QFR9_PANTR	🗋	**Insulin-like 3**	**INSL3**	Pan troglodytes (Chimpanzee)	131
☐	H2R7K5	H2R7K5_PANTR	🗋	**Insulin-degrading enzyme**	**IDE**	Pan troglodytes (Chimpanzee)	1,019
☐	H2QJD5	H2QJD5_PANTR	🗋	**Insulin-like growth factor**	**IGFBP2**	Pan troglodytes (Chimpanzee)	325

P30410 - INS_PANTR

Protein	**Insulin**
Gene	**INS**
Organism	*Pan troglodytes (Chimpanzee)*
Status	Reviewed - ●●●●● - Experimental evidence at protein level

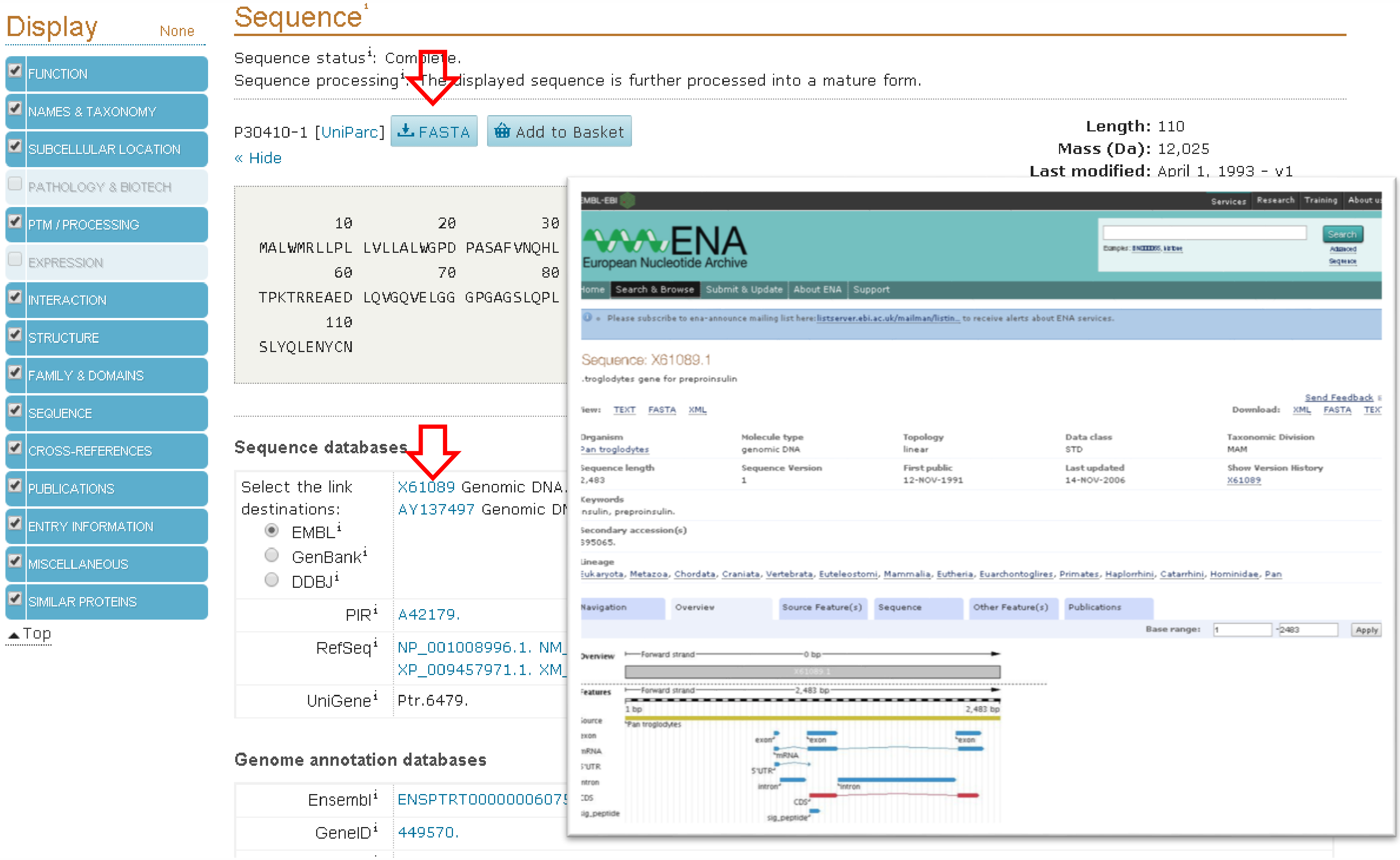
Function

Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver.

GO - Biological process

- activation of protein kinase B activity Source: Ensembl
- alpha-beta T cell activation Source: Ensembl
- glucose homeostasis Source: Ensembl
- glucose transport Source: Ensembl
- MAPK cascade Source: Ensembl
- negative regulation of acute inflammatory response Source: Ensembl
- negative regulation of fatty acid metabolic process Source: Ensembl
- negative regulation of feeding behavior Source: Ensembl
- negative regulation of NAD(P)H oxidase activity Source: Ensembl
- negative regulation of protein oligomerization Source: Ensembl
- negative regulation of proteolysis Source: Ensembl
- negative regulation of respiratory burst involved in inflammatory response Source: Ensembl
- positive regulation of cell proliferation Source: Ensembl
- positive regulation of DNA replication Source: Ensembl
- positive regulation of glycogen biosynthetic process Source: Ensembl
- positive regulation of glycolytic process Source: Ensembl
- sulin receptor signaling pathway Source: Ensembl
- acute-phase response Source: Ensembl
- fatty acid homeostasis Source: Ensembl
- glucose metabolic process Source: UniProtKB-KW
- G-protein coupled receptor signaling pathway Source: Ensembl
- negative regulation of glycogen catabolic process Source: Ensembl
- negative regulation of protein catabolic process Source: Ensembl
- negative regulation of protein secretion Source: Ensembl
- positive regulation of cytokine secretion Source: Ensembl
- positive regulation of glucose import Source: Ensembl

```
>sp|P30410|INS_PANTR Insulin OS=Pan troglodytes GN=INS PE=1 SV=1
MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```



Sequence[i]

Sequence status[i]: Complete.
Sequence processing[i]: The displayed sequence is further processed into a mature form.

P30410-1 [UniParc] FASTA Add to Basket
« Hide

Length: 110
Mass (Da): 12,025
Last modified: April 1, 1993 - v1

```
        10         20         30
MALWMRLLPL LVLLALWGPD PASAFVNQHL
        60         70         80
TPKTRREAED LQVGQVELGG GPGAGSLQPL
       110
SLYQLENYCN
```

Sequence databases

Select the link destinations: ● EMBL[i] ○ GenBank[i] ○ DDBJ[i]	X61089 Genomic DNA. AY137497 Genomic DN
PIR[i]	A42179.
RefSeq[i]	NP_001008996.1. NM_ XP_009457971.1. XM_
UniGene[i]	Ptr.6479.

Genome annotation databases

Ensembl[i]	ENSPTRT00000006075
GeneID[i]	449570.

Please subscribe to ena-announce mailing list here: listserver.ebi.ac.uk/mailman/listin... to receive alerts about ENA services.

Sequence: X61089.1

.troglodytes gene for preproinsulin

Send Feedback

View: TEXT FASTA XML
Download: XML FASTA TEX

Organism	Molecule type	Topology	Data class	Taxonomic Division
Pan troglodytes	genomic DNA	linear	STD	MAM
Sequence length	**Sequence Version**	**First public**	**Last updated**	**Show Version History**
2,483	1	12-NOV-1991	14-NOV-2006	X61089

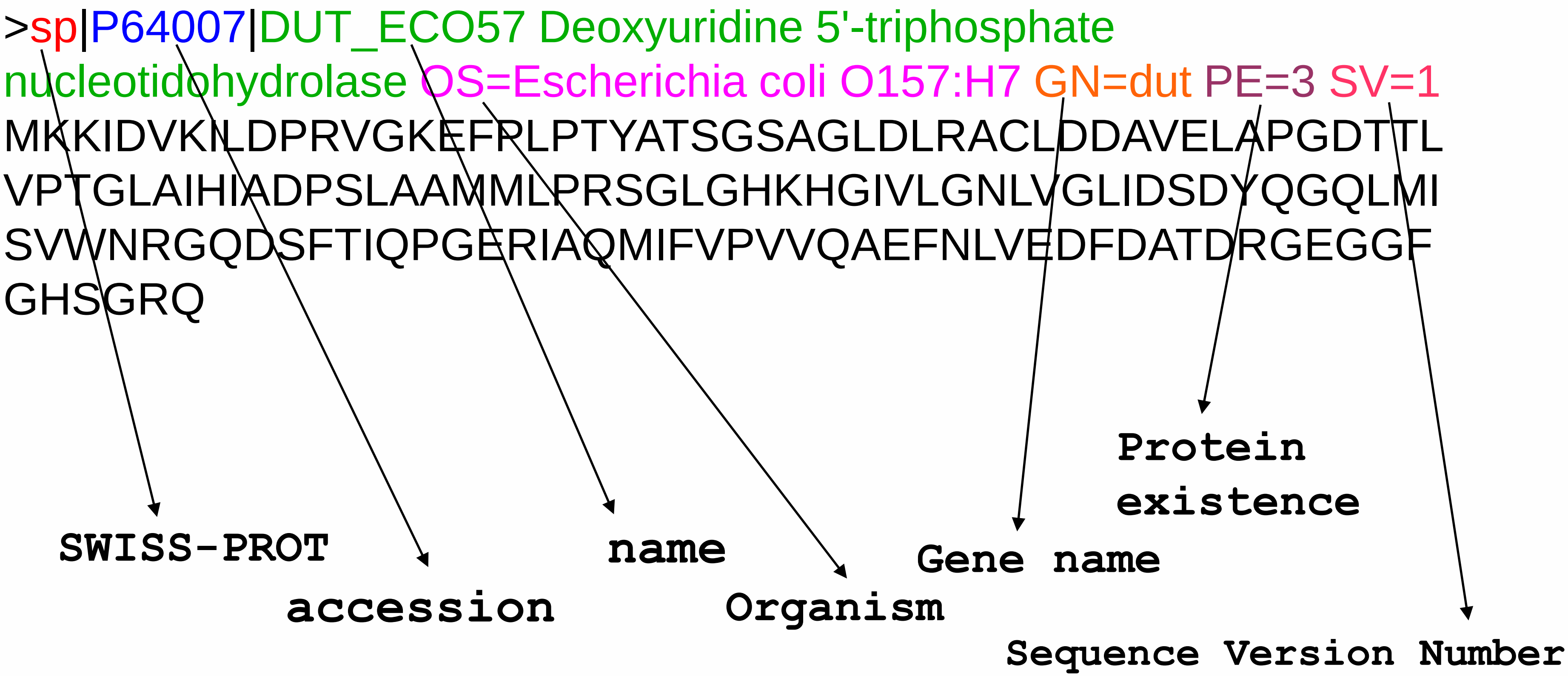
Keywords
nsulin, preproinsulin.

Secondary accession(s)
395065.

Lineage
Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini, Hominidae, Pan

A sequence in FASTA format **begins with a single-line description, followed by lines of sequence data**. The description line is distinguished from the sequence data by a greater-than ("**>**") symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length.

E. coli dUTPase protein sequence in FASTA format.



```
>sp|P64007|DUT_ECO57 Deoxyuridine 5'-triphosphate
nucleotidohydrolase OS=Escherichia coli O157:H7 GN=dut PE=3 SV=1
MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLDDAVELAPGDTTL
VPTGLAIHIADPSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMI
SVWNRGQDSFTIQPGERIAQMIFVPVVQAEFNLVEDFDATDRGEGGF
GHSGRQ
```

Protein existence

SWISS-PROT

name

Gene name

accession

Organism

Sequence Version Number

Protein Data Bank

http://www.rcsb.org

An Information Portal to Biological Macromolecular Structures

As of **Tuesday Jan 08, 2013 at 4 PM PST** there are **87279** Structures | **PDB Statistics**

Insulin human

Organism

- Homo sapiens (21609)
- Escherichia coli (4675)
- Mus musculus (3732)
- Saccharomyces cerevisiae (2209)
- Bos taurus (2117)
- Rattus norvegicus (1834)
- Escherichia coli K-12 (1486)
- Other (47060)

close

PDB-101 Hide

- Structural View of Biology
- Understanding PDB Data
- Molecule of the Month
- Educational Resources
- Author Profiles

MyPDB Hide

- Login to your Account
- Register a New Account

Home Hide

- News & Publications
- Usage/Reference Policies
- Deposition Policies
- Website FAQ
- Deposition FAQ
- Contact Us
- About Us
- Careers
- External Links
- Sitemap
- New Website Features

Deposition Hide

- All Deposit Services

Hide

e By: Title | Date | Category

Molecule of the Month

Transfer-Messenger RNA

Damaged messenger RNA poses a double danger to cells. If a messenger RNA is truncated, it will be missing its stop codon, so it will encode a faulty, truncated protein. Also, ribosomes get stalled at the end of these truncated messages and are unable to release the mRNA and move on to the next protein synthesis job. Bacteria possess an ingenious molecular method for solving both of these problems at the same time, that destroys the faulty protein and releases the ribosome all at once.

Full Article

Protein Structure Initiative Featured System

Bacterial Hemophores

Our bodies are filled with iron: iron in hemoglobin colors our blood bright red and iron is used as a chemical tool in many cellular proteins. Altogether, we have several grams of iron scattered through our cells. Surprisingly, however, lack of iron often limits the growth of pathogenic bacteria infecting our bodies. This is by design: our bodies have evolved to guard their store of iron, so the amount of free iron circulating through the body is vanishingly small. PSI biology researchers are exploring the proteins that pathogenic bacteria use to gather scarce heme groups, and the iron ions they carry.

Full Article | Archive | PSI Structural Biology Knowledgebase

Explore Archive Hide

New Structures Hide

- Latest Release
- New Structure Papers
- Search Unreleased Entries

New Features Hide

Macromolecule Names: Short Name

Latest features released:

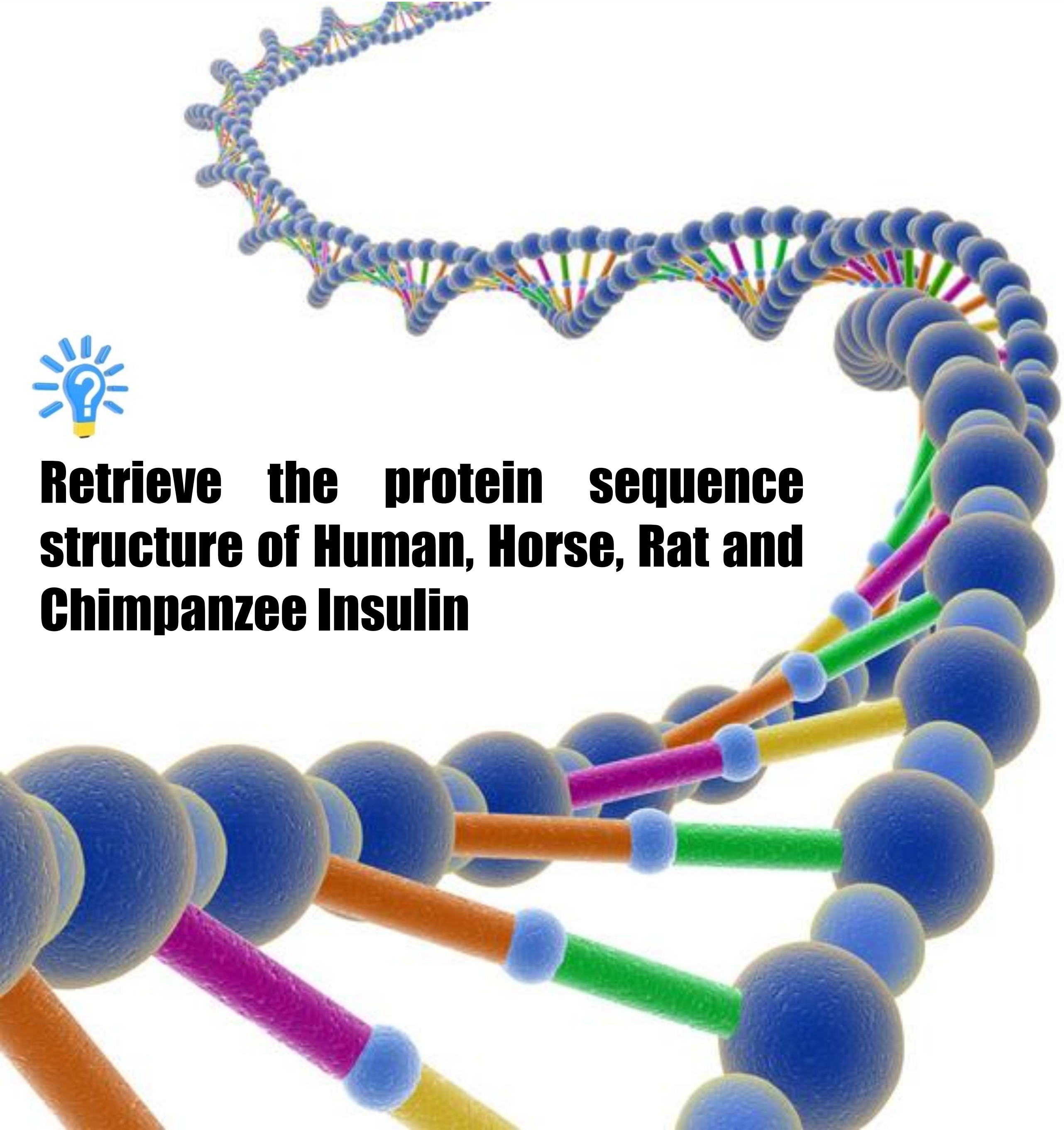
Website Release Archive:

RCSB PDB News Hide

Weekly | Quarterly | Yearly

2013-01-08

Winter Newsletter Published



Retrieve the protein sequence structure of Human, Horse, Rat and Chimpanzee Insulin