IV107 Bioinformatika I

Přednáška 5

Katedra informačních technologií
Masarykova Univerzita Brno

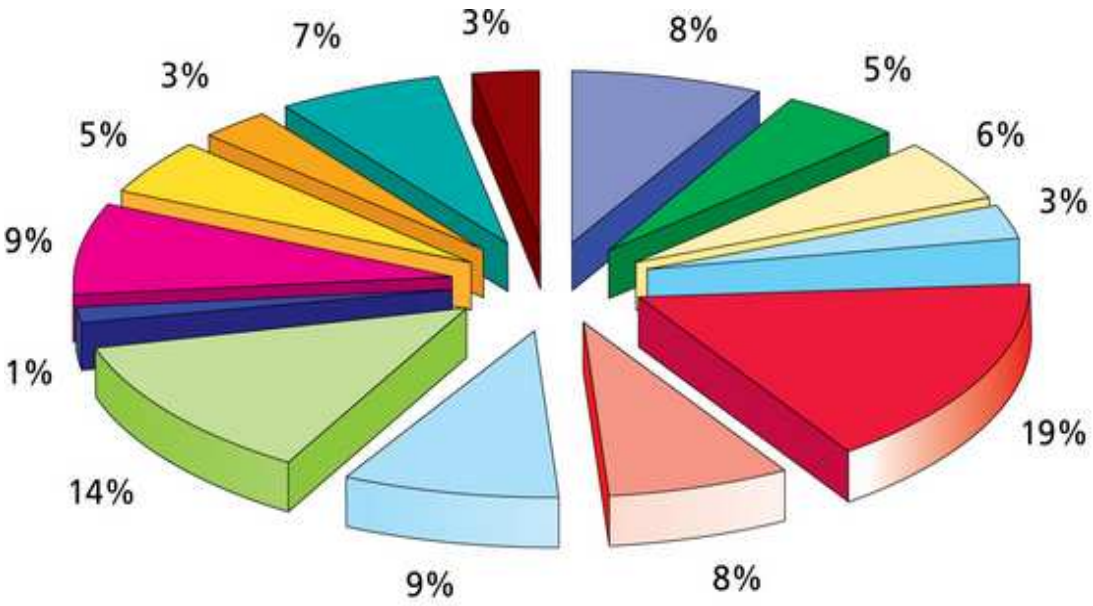
Jaro 2008

Předchozí týden

- Struktura genu
  - prokaryotického
  - eukaryotického
- Porovnání sekvencí
  - globální (Needleman–Wunsch)
  - semi-globální
  - lokální (Smith–Waterman)

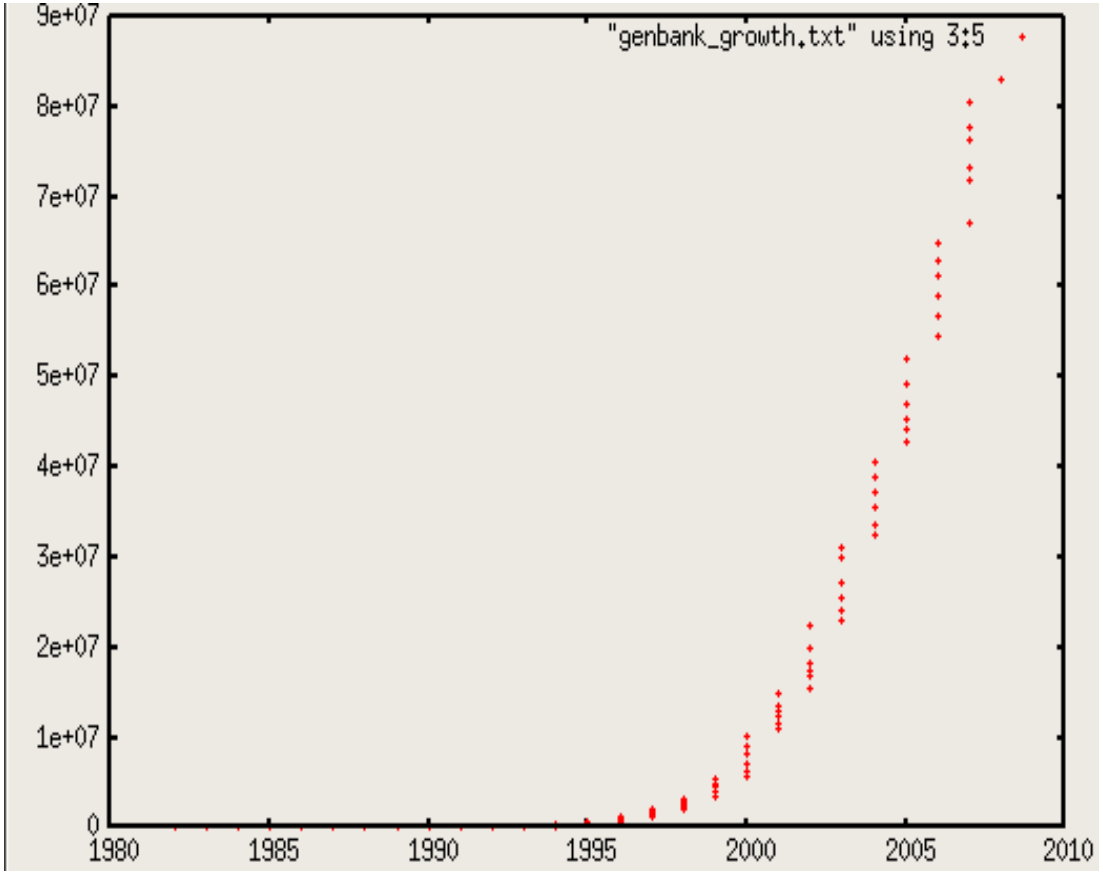
Outline

Typy dat v databázích



http://www.agr.kuleuven.ac.be/vakken/i287/bioinformatica.htm

Nárůst databáze GenBank



GenBank

Genetic Sequence Data Bank
February 15 2008
NCBI-GenBank Flat File Release 164.0
National Center for Biotechnology Information

- 85759586764 bases
- 82853685 sequences

ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

Součásti databáze GenBank

- INV, VRT, MAM, PLN, PRI, ROD, BCT, VRL
- PAT (Patents)
- HTGS (High Throughput Genomic Sequences)
- GSS (Genome Survey Sequences)
- ETS (Expressed Sequence Tags)
- STS (Sequence Tagged Sites)

Příklad záznamu v databázi GenBank

```
LOCUS       SCU49845     5028 bp    DNA
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene,
            Axl2p
            (AXL2) and Rev7p (REV7) genes, complete
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomy
           Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; S
```

Vyhledávání v sekvenčních databázích

- textové (klíčová slova)
- sekvenční (BLAST)

Příklad záznamu v databázi UniProt

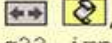
Entry information

Entry name	**LMO7_HUMAN**
Primary accession number	**Q8WWI1**
Secondary accession numbers	O15462 O95346 Q9UKC1 Q9UQM5 Q9Y6A7
Integrated into Swiss-Prot on	March 15, 2004
Sequence was last modified on	March 15, 2004 (Sequence version 2)
Annotations were last modified on	July 25, 2006 (Entry version 39)

Name and origin of the protein

Protein name	**LIM domain only protein 7**
Synonyms	**LOMP** **F-box only protein 20**
Gene name	**Name:** LMO7 Synonyms: FBX20, FBXO20, KIAA0858
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

References

[1] NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 3), AND TISSUE SPECIFICITY.
TISSUE=Brain, and Peripheral blood leukocyte;
DOI=10.1007/s00439-001-0646-6; PubMed=11935316 [NCBI, ExPASy, EBI, Israel, Japan]
Rozenblum E., Vahteristo P., Sandberg T., Bergthorsson J.T., Syrjakoski K., Weaver D., Haraldsson K., Johannsdottir H.K., Vehmanen P., Nigam S., Golberger N., Robbins C., Pak E., Dutra A., Gillander E., Stephan D.A., Bailey-Wilson J., Juo S.-H.H., Kainu T.,  , Kallioniemi O.-P.;
"A genomic map of a 6-Mb region at 13q21-q22 implicated in cancer development: identification and characterization of candidate genes.";
Hum. Genet. 110:111-121(2002).

http://www.uniprot.org/

Příklad záznamu v databázi UniProt

Key	From	To	Length	Description	FTId
CHAIN	1	1683	1683	LIM domain only protein 7.	PRO_0000075824
DOMAIN	54	168	115	CH.	
DOMAIN	1042	1128	87	PDZ.	
DOMAIN	1612	1678	67	LIM zinc-binding.	

```
        10         20         30         40         50         60
MKKIRICHIF TFYSWMSYDV LFQRTELGAL EIWRQLICAH VCICVGWLYL RDRVCSKKDI

        70         80         90        100        110        120
ILRTEQNSGR TILIKAVTEK NFETKDFRAS LENGVLLCDL INKLKPGVIK KINRLSTPIA

       130        140        150        160        170        180
GLDNINVFLK ACEQIGLKEA QLFHPGDLQD LSNRVTVKQE ETDRRVKNVL ITLYWLGRKA
```

PDB

Záznam v PDB

```
HEADER    HYDROLASE(O-GLYCOSYL)                    20-JAN-92   1HEW      1HEW   2
COMPND    LYSOZYME (E.C.3.2.1.17) COMPLEXED WITH THE INHIBITOR          1HEW   3
COMPND   2 TRI-N-ACETYLCHITOTRIOSE                                      1HEW   4
SOURCE    HEN (GALLUS GALLUS) EGG WHITE                                 1HEW   5
AUTHOR    J.C.CHEETHAM,P.J.ARTYMIUK,D.C.PHILLIPS                        1HEW   6
REVDAT   1   31-JAN-94 1HEW    0                                        1HEW   7
JRNL        AUTH   J.C.CHEETHAM,P.J.ARTYMIUK,D.C.PHILLIPS              1HEW   8
JRNL        TITL   REFINEMENT OF AN ENZYME COMPLEX WITH INHIBITOR       1HEW   9
JRNL        TITL 2 BOUND AT PARTIAL OCCUPANCY. HEN EGG-WHITE            1HEW  10
JRNL        TITL 3 LYSOZYME AND TRI-N-ACETYLCHITOTRIOSE AT 1.75         1HEW  11
JRNL        TITL 4 ANGSTROMS RESOLUTION                                 1HEW  12
JRNL        REF    J.MOL.BIOL.                   V. 224   613 1992      1HEW  13
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836                  070  1HEW  14
REMARK   1                                                            1HEW  15
REMARK   1 REFERENCE 1                                                1HEW  16
REMARK   1  AUTH   L.N.JOHNSON,J.C.CHEETHAM,P.J.MC*LAUGHLIN,         1HEW  17
REMARK   1  AUTH 2 K.R.ACHARYA,D.BARFORD,D.C.PHILLIPS                 1HEW  18
REMARK   1  TITL   PROTEIN-OLIGOSACCHARIDE INTERACTIONS: LYSOZYME,    1HEW  19
REMARK   1  TITL 2 PHOSPHORYLASE, AMYLASES                            1HEW  20
REMARK   1  REF    CURR.TOP.MICROBIOL.IMMUNOL.   V. 139    81 1988      1HEW  21
REMARK   1  REFN   ASTM CTMIA3  GW ISSN 0070-217X                761  1HEW  22
```

Záznam v PDB

```
REMARK   5 THE THREE SUGAR UNITS OF THE INHIBITOR MOLECULE ARE BOUND    1HEW  56
REMARK   5 IN THE UPPER THREE SITES (A TO C) OF THE LYSOZYME ACTIVE     1HEW  57
REMARK   5 SITE CLEFT.  NAG MOLECULES, NUMBERED 203, 202, AND 201, ARE  1HEW  58
REMARK   5 BOUND IN SITES A, B, AND C, RESPECTIVELY.                    1HEW  59
SEQRES   1    129  LYS VAL PHE GLY ARG CYS GLU LEU ALA ALA ALA MET LYS  1HEW  60
SEQRES   2    129  ARG HIS GLY LEU ASP ASN TYR ARG GLY TYR SER LEU GLY  1HEW  61
SEQRES   3    129  ASN TRP VAL CYS ALA ALA LYS PHE GLU SER ASN PHE ASN  1HEW  62
SEQRES   4    129  THR GLN ALA THR ASN ARG ASN THR ASP GLY SER THR ASP  1HEW  63
SEQRES   5    129  TYR GLY ILE LEU GLN ILE ASN SER ARG TRP TRP CYS ASN  1HEW  64
SEQRES   6    129  ASP GLY ARG THR PRO GLY SER ARG ASN LEU CYS ASN ILE  1HEW  65
SEQRES   7    129  PRO CYS SER ALA LEU LEU SER SER ASP ILE THR ALA SER  1HEW  66
SEQRES   8    129  VAL ASN CYS ALA LYS LYS ILE VAL SER ASP GLY ASN GLY  1HEW  67
SEQRES   9    129  MET ASN ALA TRP VAL ALA TRP ARG ASN ARG CYS LYS GLY  1HEW  68
SEQRES  10    129  THR ASP VAL GLN ALA TRP ILE ARG GLY CYS ARG LEU      1HEW  69
HET    NAG    201      15     N-ACETYL-D-GLUCOSAMINE                    1HEW  70
HET    NAG    202      14     N-ACETYL-D-GLUCOSAMINE                    1HEW  71
HET    NAG    203      14     N-ACETYL-D-GLUCOSAMINE                    1HEW  72
FORMUL   2  NAG    3(C8 H15 N1 O6)                                      1HEW  73
```

Záznam v PDB

```
HELIX    1   A ARG      5  HIS     15  1                                  1HEW  75
HELIX    2   B LEU     25  GLU     35  1                                  1HEW  76
HELIX    3   C CYS     80  LEU     84  5                                  1HEW  77
HELIX    4   D THR     89  ILE     98  1                                  1HEW  78
HELIX    5   E VAL    109  ASN    113  1                                  1HEW  79
SHEET    1  S1 2 LYS     1  PHE     3  0                                  1HEW  80
SHEET    2  S1 2 PHE    38  THR    40 -1  N  THR    40   O  LYS     1    1HEW  81
SHEET    1  S2 3 ALA    42  ASN    46  0                                  1HEW  82
SHEET    2  S2 3 SER    50  GLY    54 -1  O  SER    50   N  ASN    46    1HEW  83
SHEET    3  S2 3 GLN    57  SER    60 -1  O  ILE    58   N  TYR    53    1HEW  84
TURN     1  T1 MET    12  HIS    15     TYPE III                         1HEW  85
TURN     2  T2 LYS    13  GLY    16     TYPE I                           1HEW  86
TURN     3  T3 LEU    17  TYR    20     TYPE II                          1HEW  87
TURN     4  T4 ASN    19  GLY    22     DISTORTED TYPE II                1HEW  88
TURN     5  T5 TYR    20  TYR    23     TYPE I'                          1HEW  89
TURN     6  T6 SER    24  ASN    27     TYPE III                         1HEW  90
TURN     7  T7 LEU    25  TRP    28     TYPE III                         1HEW  91
TURN     8  T8 SER    36  ASN    39     TYPE III'                        1HEW  92
```

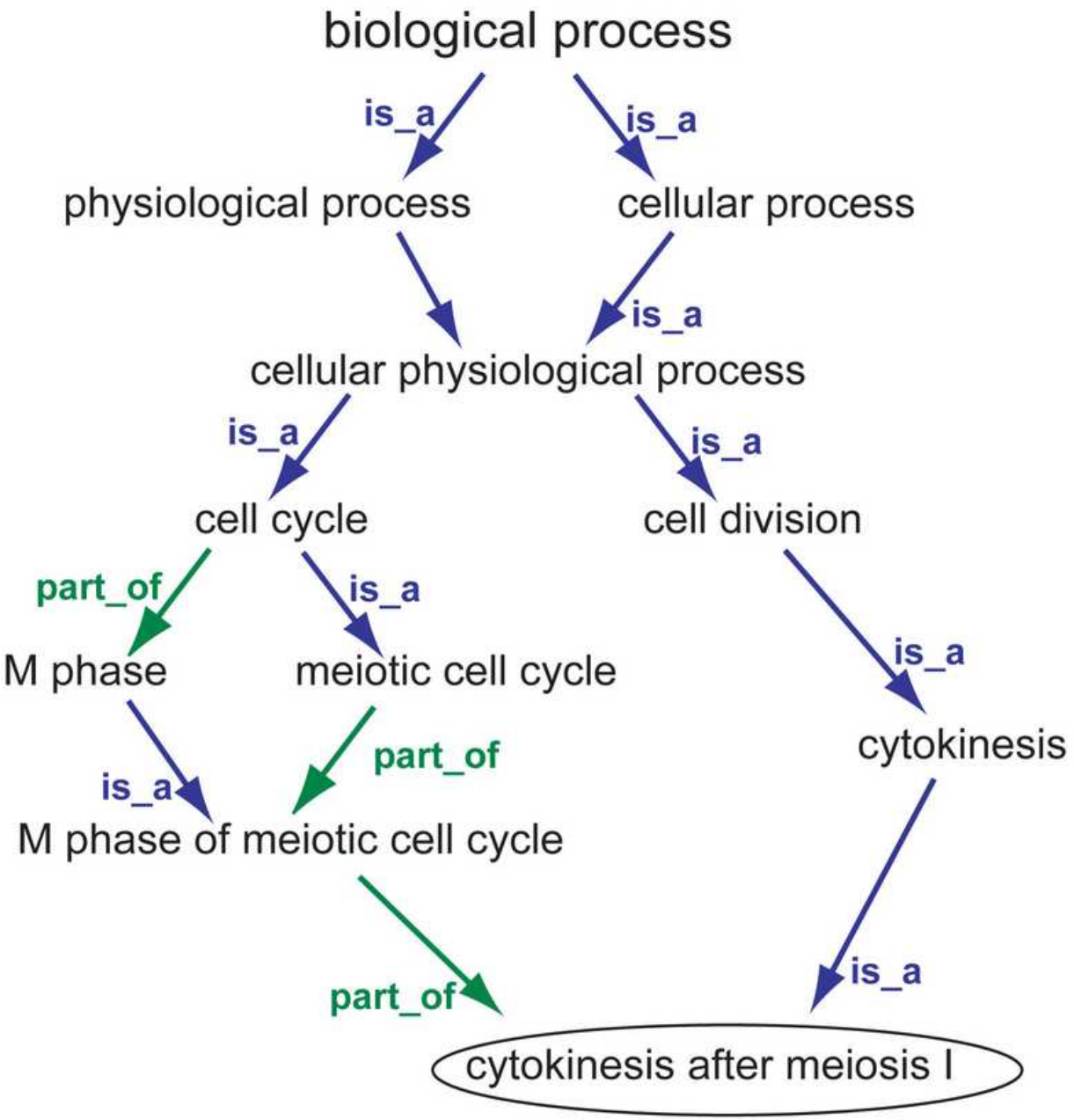
Záznam v PDB

```
CRYST1   78.860   78.860   38.250  90.00  90.00  90.00 P 43 21 2     8  1HEW 113
ORIGX1      1.000000  0.000000  0.000000        0.00000                      1HEW 114
ORIGX2      0.000000  1.000000  0.000000        0.00000                      1HEW 115
ORIGX3      0.000000  0.000000  1.000000        0.00000                      1HEW 116
SCALE1      0.012681  0.000000  0.000000        0.00000                      1HEW 117
SCALE2      0.000000  0.012681  0.000000        0.00000                      1HEW 118
SCALE3      0.000000  0.000000  0.026144        0.00000                      1HEW 119
ATOM      1  N   LYS     1       3.398   9.981  10.408  1.00 30.48      1HEW 120
ATOM      2  CA  LYS     1       2.459  10.365   9.364  1.00 28.03      1HEW 121
ATOM      3  C   LYS     1       2.458  11.880   9.149  1.00 21.93      1HEW 122
ATOM      4  O   LYS     1       2.481  12.672  10.100  1.00 14.10      1HEW 123
ATOM      5  CB  LYS     1       1.026   9.935   9.695  1.00 30.54      1HEW 124
ATOM      6  CG  LYS     1       0.028  10.169   8.558  1.00 37.93      1HEW 125
ATOM      7  CD  LYS     1      -1.415  10.089   9.048  1.00 33.23      1HEW 126
ATOM      8  CE  LYS     1      -2.357  10.822   8.082  1.00 32.17      1HEW 127
ATOM      9  NZ  LYS     1      -3.661  10.090   8.025  1.00 31.92      1HEW 128
ATOM     10  N   VAL     2       2.429  12.232   7.880  1.00 17.30      1HEW 129
ATOM     11  CA  VAL     2       2.395  13.653   7.465  1.00 14.47      1HEW 130
ATOM     12  C   VAL     2       0.977  13.868   6.903  1.00 17.58      1HEW 131
ATOM     13  O   VAL     2       0.642  13.368   5.826  1.00 32.65      1HEW 132
ATOM     14  CB  VAL     2       3.533  14.012   6.536  1.00 22.88      1HEW 133
```

Gene Ontology

- Funkce genů a proteinů zjišťujeme experimentálně
- Slovní popis není jednoznačný
  - syntéza proteinů
  - syntéza polypeptidů
  - translace
  - aktivita ribozomů
- Ontologie je způsob jak do používaných termínů vnést systém

Gene Ontology



Gene Ontology

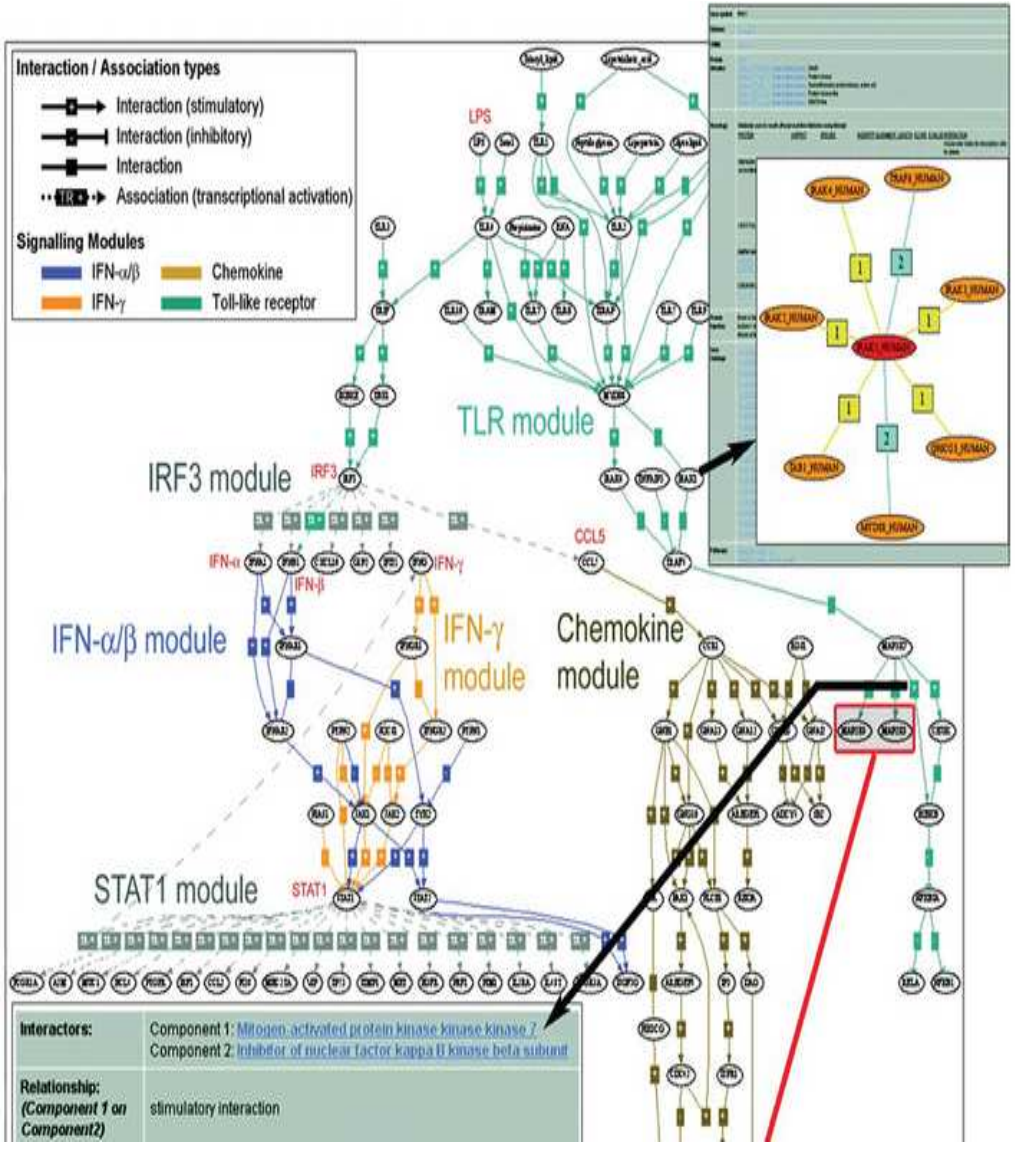
- Molekulární proces
  - katalytická aktivita
  - transport
  - intermolekulární vazba
- Biologický proces
  - přenos signálu
  - aktivace imunitního sytému
  - regulace genů
- Buněčná složka
  - buněčné jádro
  - plazmatická membrána

Gene Ontology - kódy zdroje dat

Curator-assigned Evidence Codes

- Experimental Evidence Codes
  - IDA: Inferred from Direct Assay
  - IPI: Inferred from Physical Interaction
  - IMP: Inferred from Mutant Phenotype
  - IGI: Inferred from Genetic Interaction
  - IEP: Inferred from Expression Pattern
- Computational Analysis Evidence Codes
  - ISS: Inferred from Sequence or Structural Similarity
  - IGC: Inferred from Genomic Context
  - RCA: inferred from Reviewed Computational Analysis
- Author Statement Evidence Codes
  - TAS: Traceable Author Statement
  - NAS: Non-traceable Author Statement
- Curator Statement Evidence Codes
  - IC: Inferred by Curator
  - ND: No biological Data available
- Automatically-assigned Evidence Codes
  - IEA: Inferred from Electronic Annotation
- Obsolete Evidence Codes
  - NR: Not Recorded

Metabolické dráhy



http://www.genome.jp/kegg/

Příště Analýza proteinových sekvencí, strukturních a funkčních dat

Outline

Dodatek

For Further Reading
X