

GENE ONTOLOGY

Classificare la materia vivente e comprenderla




Francesca Cordero, Ph.D. Student

















fcordero@di.unito.it




Search across databases

GO CLEAR Help

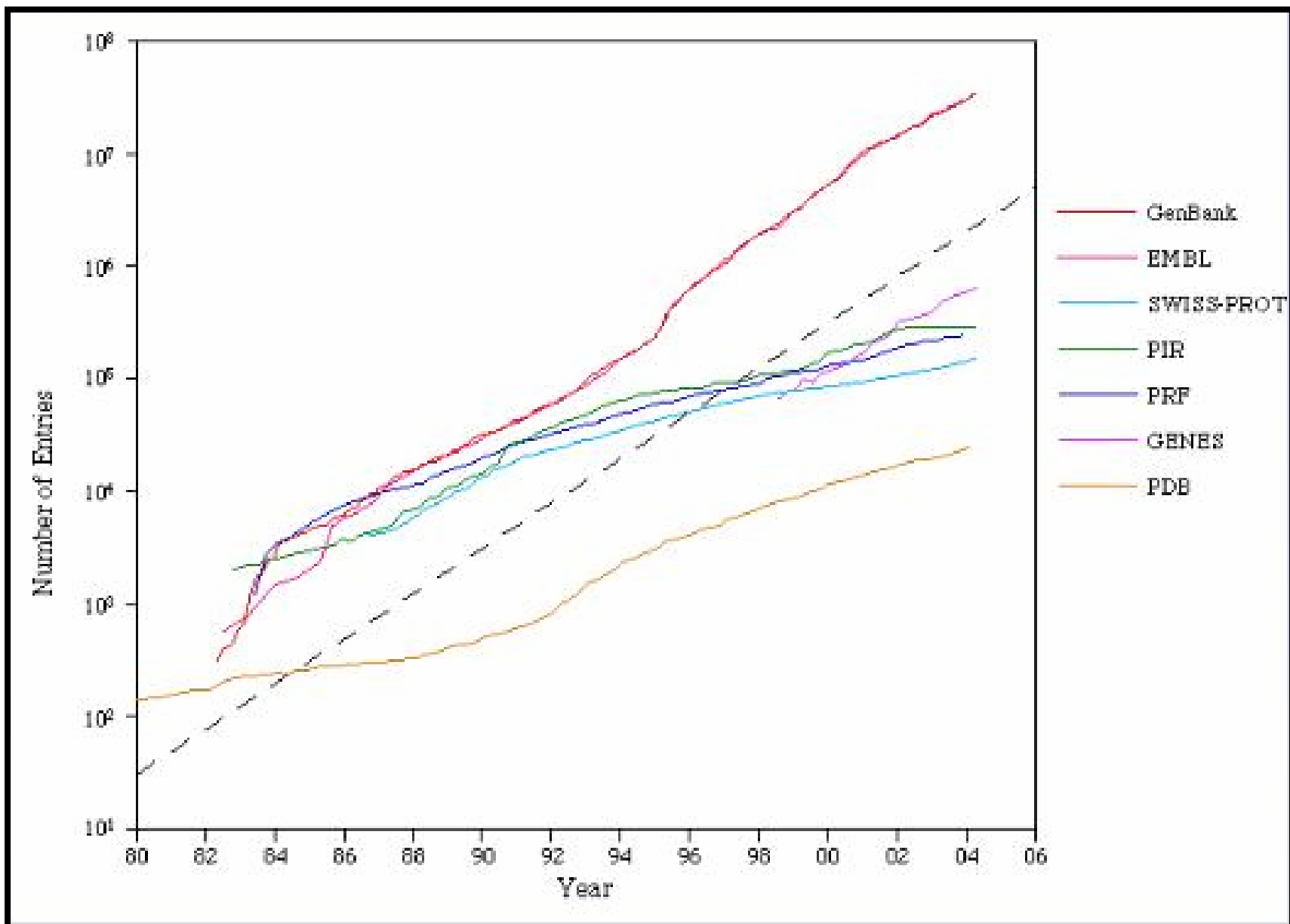
Welcome to the new Entrez cross-database search page

 PubMed: biomedical literature citations and abstracts ?	 Books: online books ?
 PubMed Central: free, full text journal articles ?	 OMIM: online Mendelian Inheritance in Man ?
	 Site Search: NCBI web and FTP sites ?

 Nucleotide: sequence database (GenBank) ?	 UniGene: gene-oriented clusters of transcript sequences ?
 Protein: sequence database ?	 CDD: conserved protein domain database ?
 Genome: whole genome sequences ?	 3D Domains: domains from Entrez Structure ?
 Structure: three-dimensional macromolecular structures ?	 UniSTS: markers and mapping data ?
 Taxonomy: organisms in GenBank ?	 PopSet: population study data sets ?
 SNP: single nucleotide polymorphism ?	 GEO Profiles: expression and molecular abundance profiles ?
 Gene: gene-centered information ?	 GEO DataSets: experimental sets of GEO data ?
 HomoloGene: eukaryotic homology groups ?	 Cancer Chromosomes: cytogenetic databases ?
 PubChem Compound: small molecule chemical structures ?	 PubChem BioAssay: bioactivity screens of chemical substances ?
 PubChem Substance: chemical substances screened for bioactivity ?	 GENSAT: gene expression atlas of mouse central nervous system ?
 Genome Project: genome project information ?	

 Journals: detailed information <i>about</i> the journals indexed in PubMed and other Entrez databases ?	 MeSH: detailed information about NLM's controlled vocabulary ?
 NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections ?	

Numero di record nei diversi database / Anno



Cosa si può chiedere ad un database?

Dove è espresso il gene?

Qual è la sua localizzazione (intra)cellulare del suo prodotto?

Quando è espresso?

Qual'è la funzione del suo prodotto?

Qual'è la struttura del suo prodotto?

Di quale processo più generale è parte?

Da chi o da cosa è controllato?

Di quale complesso la sua funzione è parte?

Search PubMed for angiogenesis [Save Search](#)

- About Entrez
- Text Version
- Entrez PubMed
 - Overview
 - Help | FAQ
 - Tutorial
 - New/Noteworthy
 - E-Utilities
- PubMed Services
 - Journals Database
 - MeSH Database
 - Single Citation Matcher
 - Batch Citation Matcher
 - Clinical Queries
 - LinkOut
 - My NCBI (Cubby)

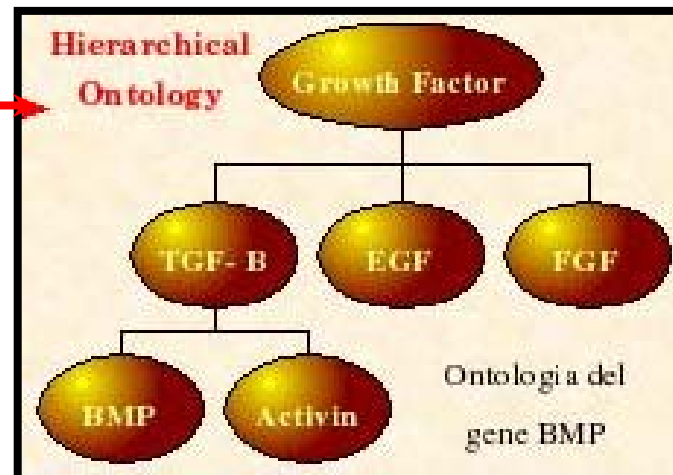
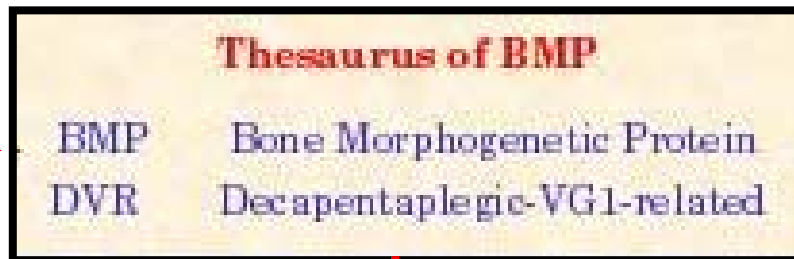
Display Show Sort by Send to

Items 1 - 20 of 22502 Page of 112

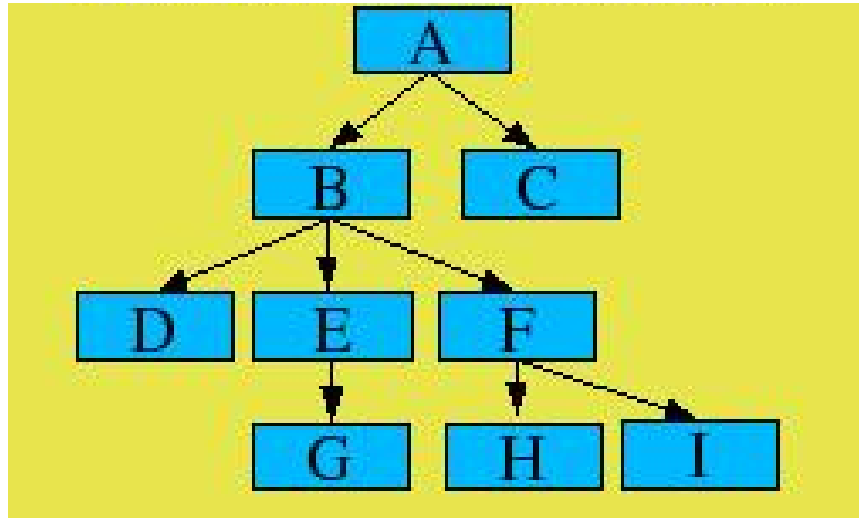
- 1: [Mayer H, Bertram H, Lindenmaier W, Korff T, Weber H, Weich H.](#)
 Vascular endothelial growth factor (VEGF-A) expression in human mesenchymal stem cells: Autocrine and paracrine role on osteoblastic and endothelial differentiation.
J Cell Biochem. 2005 Apr 18; [Epub ahead of print]
 PMID: 15838884 [PubMed - as supplied by publisher]
- 2: [Wadgaonkar R, Dudek SM, Zaiman AL, Linz-McGillen L, Verin AD, Nurmukhambetova S, Romer LH, Garcia JG.](#)
 Intracellular interaction of myosin light chain kinase with macrophage migration inhibition factor (MIF) in endothelium.
J Cell Biochem. 2005 Apr 18; [Epub ahead of print]
 PMID: 15838879 [PubMed - as supplied by publisher]
- 3: [Hu B, Wei Y, Tian L, Zhao X, Lu Y, Wu Y, Yao B, Liu J, Niu T, Wen Y, He Q, Su J, Huang M, Lou Y, Luo Y, Kan B.](#)
 Active Antitumor Immunity Elicited by Vaccine Based on Recombinant Form of Epidermal Growth Factor Receptor.
J Immunother. 2005 May/June;28(3):236-244.
 PMID: 15838380 [PubMed - as supplied by publisher]



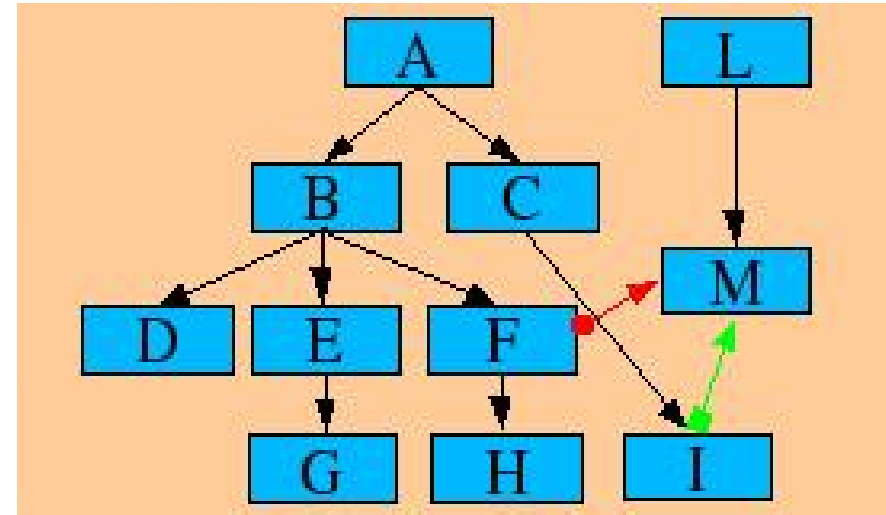
Le **Ontologie** differiscono dalle terminologie controllate (Lexicon) poiché si tratta di una struttura dati gerarchica che contiene tutte le entità rilevanti, le relazioni esistenti fra di esse, le regole, gli assiomi, ed i vincoli specifici del dominio; mentre le terminologie controllate semplicemente restringono l'insieme di parole usate per descrivere il dominio.



Come porre in relazione due oggetti (o geni)?



Questo tipo di albero esprime solo relazioni di appartenenza per esempio “è un..”

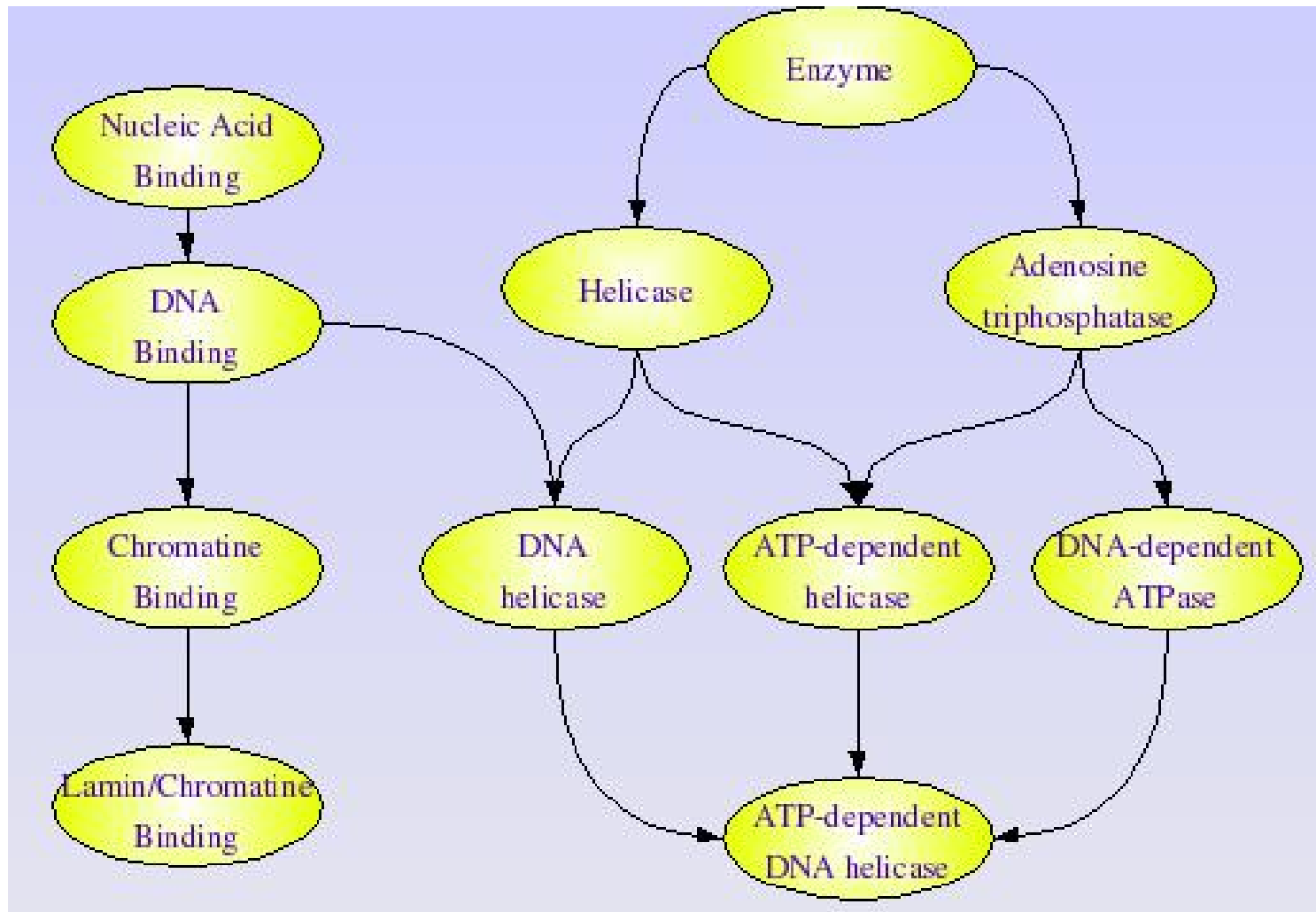


In questo grafo abbiamo diverse relazioni tra i nodi:

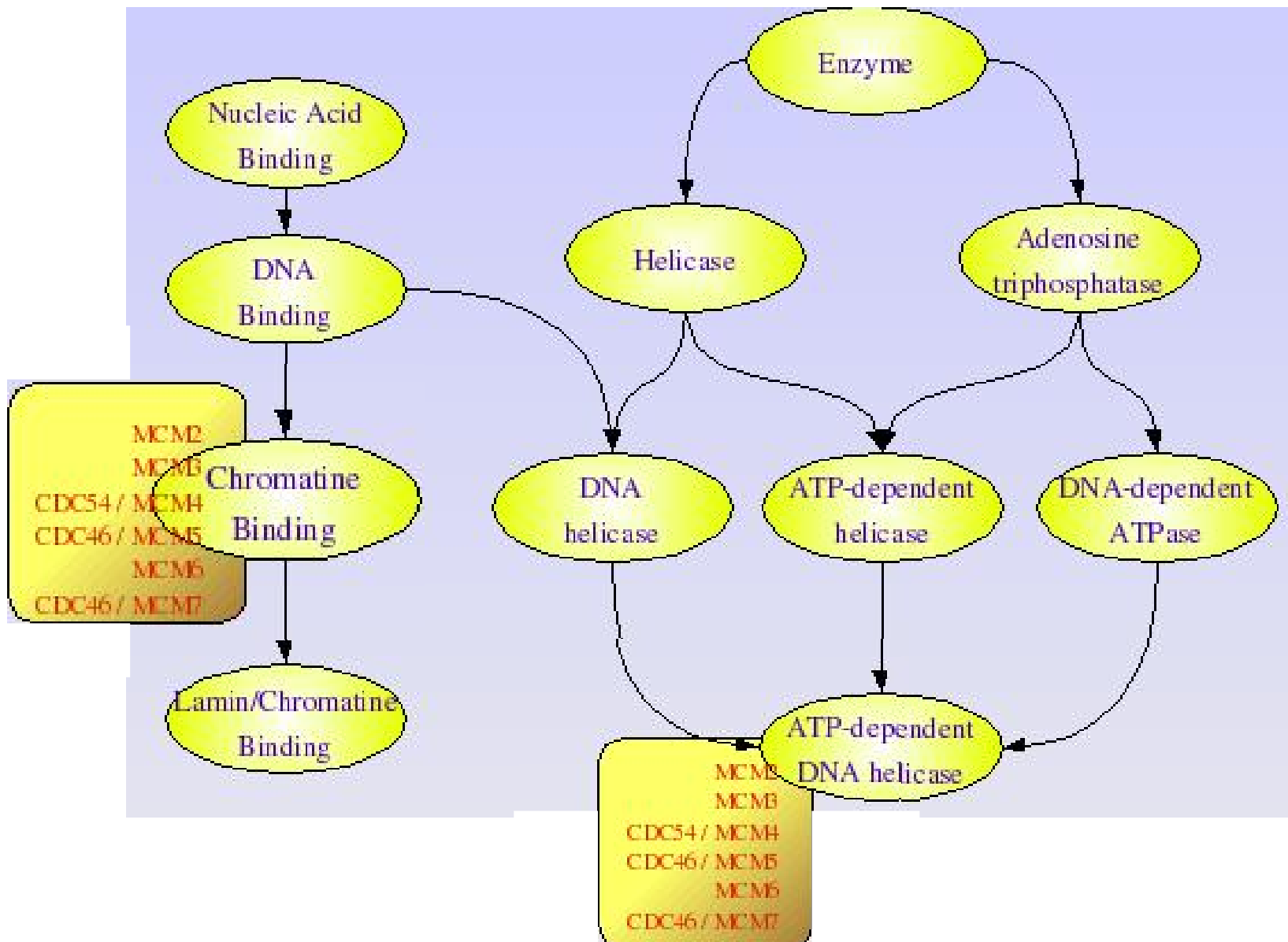
→ Istanza, è un

●→ Parte / Tutto, fa parte di

■→ Ontogenesi, origina da

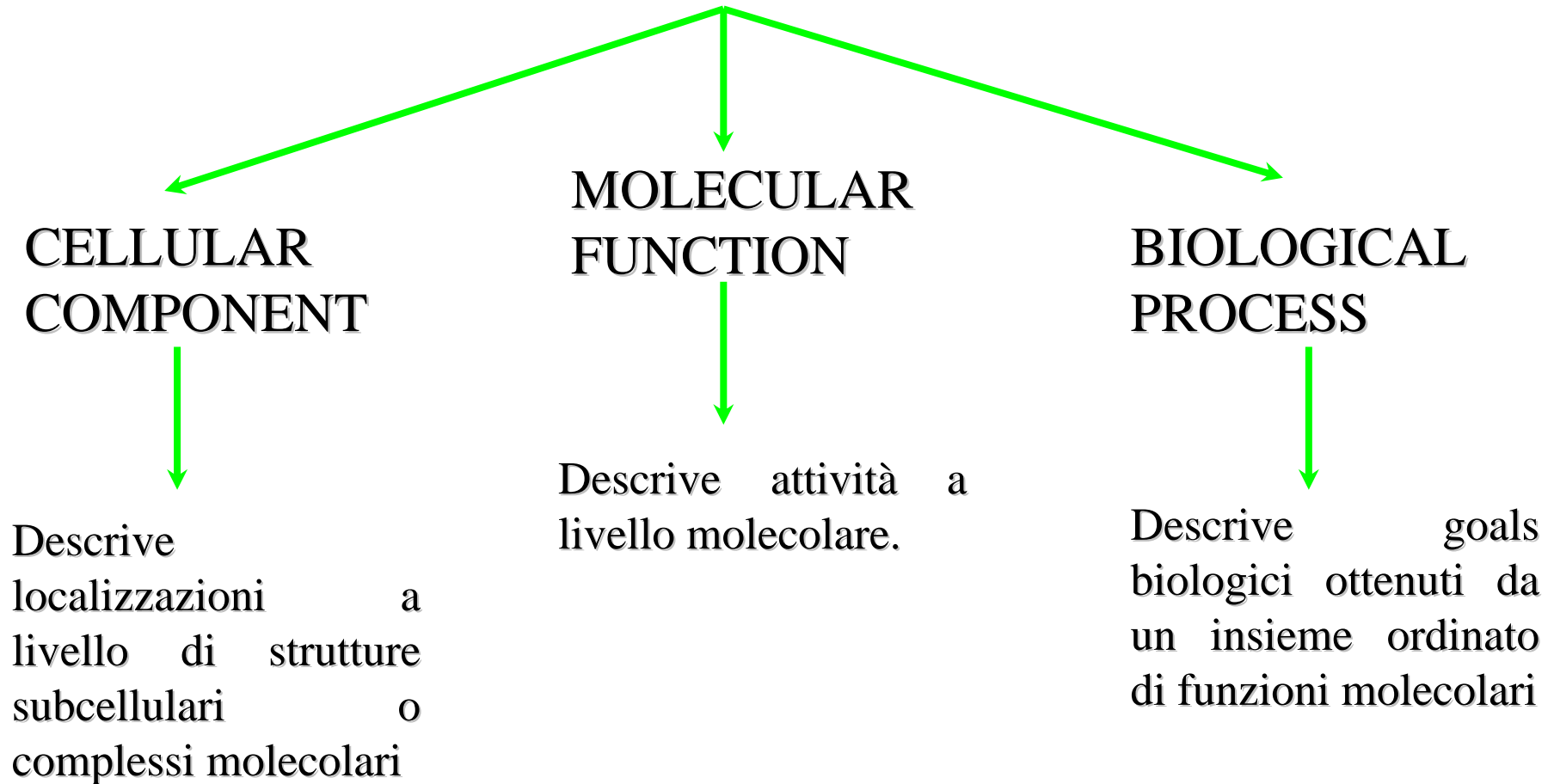


Implementando correlazioni tra oggetti diverse dal semplice istanza, si ottengono grafi più complessi in cui è possibile esprimere meglio complessi rapporti tra “oggetti” biologici memorizzati nei database.

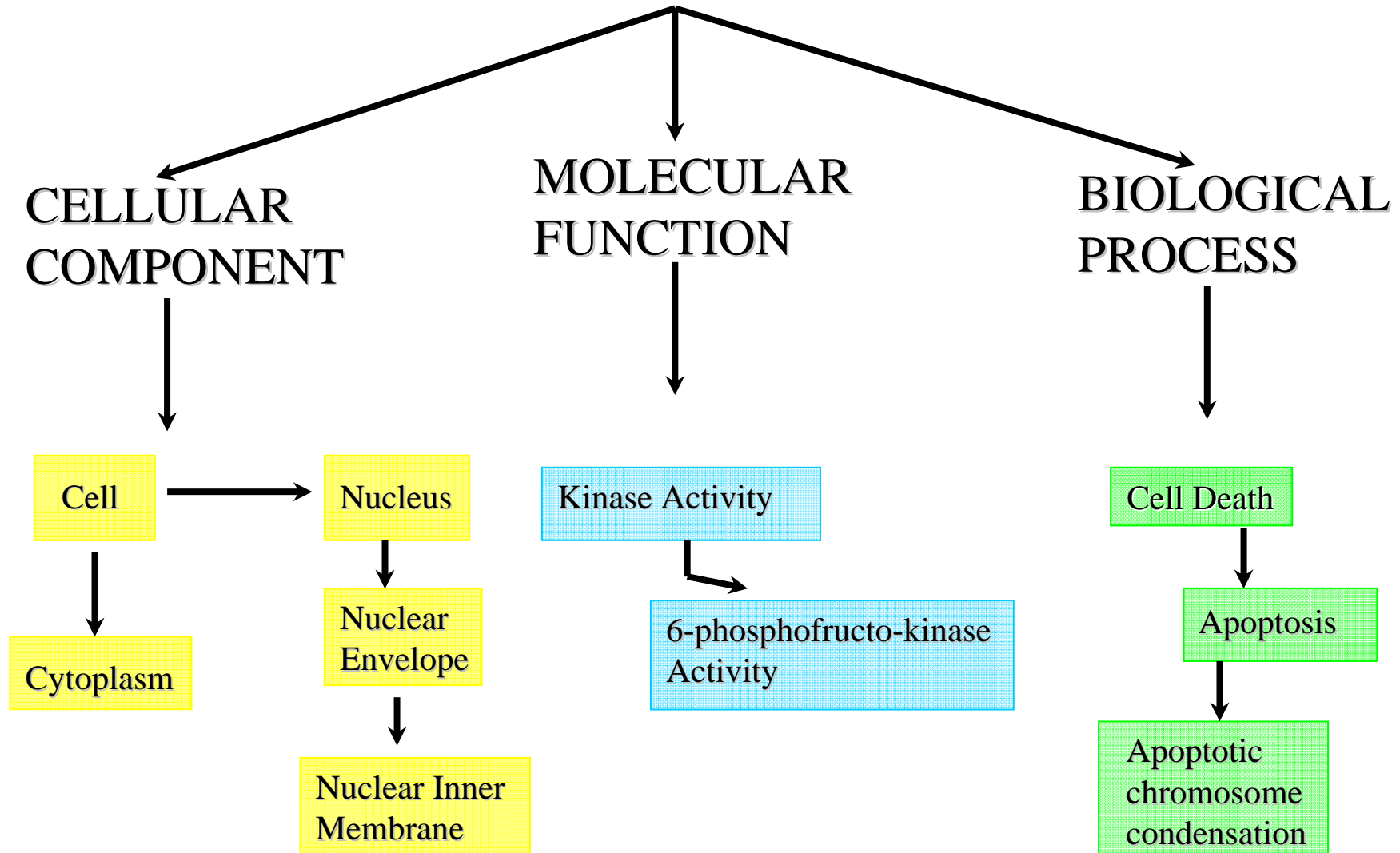


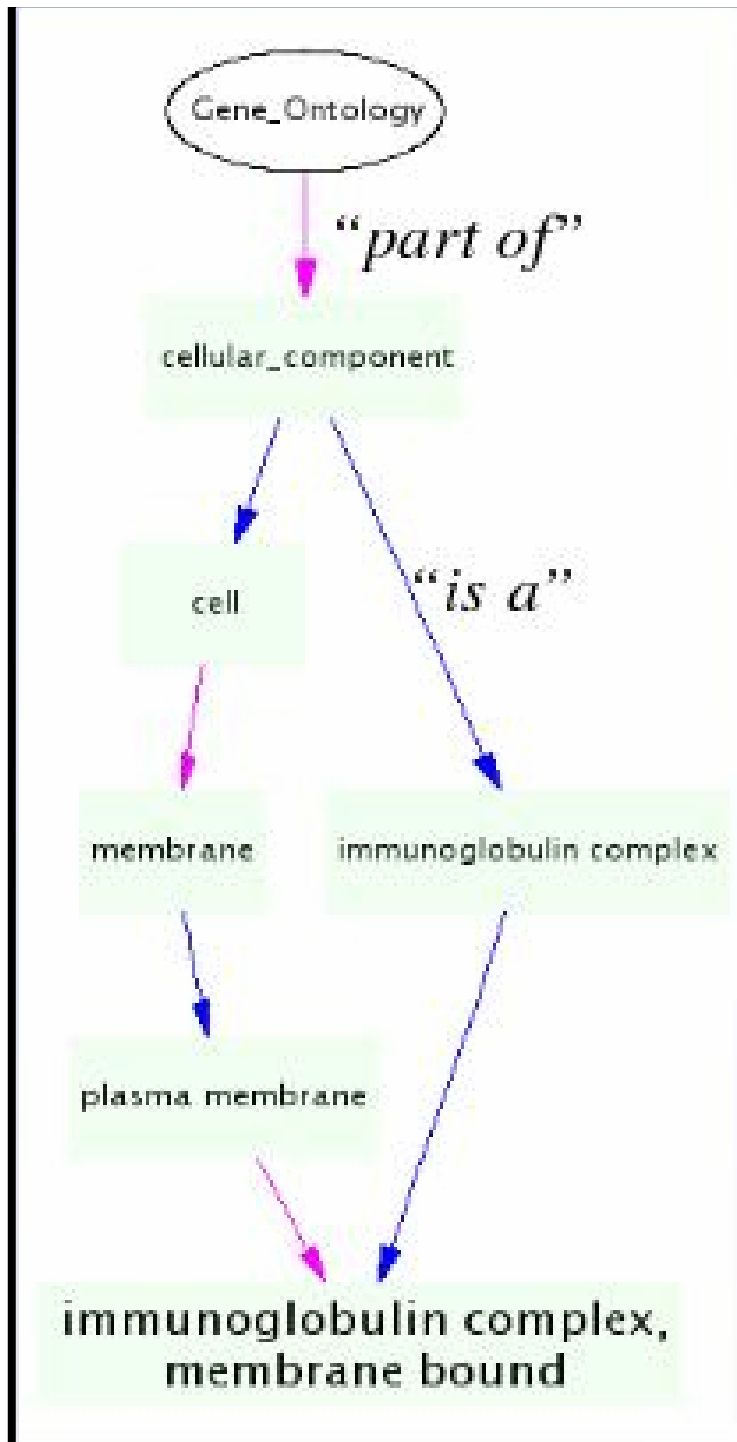
In questo grafo è possibile individuare i geni della serie MCM sia come “ATP dependent DNA helicase” che attraverso i concetti più generali di “ATPasi” o “elicasi” che restituiranno anche altri geni. Ma che è anche possibile attraverso la categoria (nodo) “chromatine binding”.

GENE ONTOLOGY



GENE ONTOLOGY

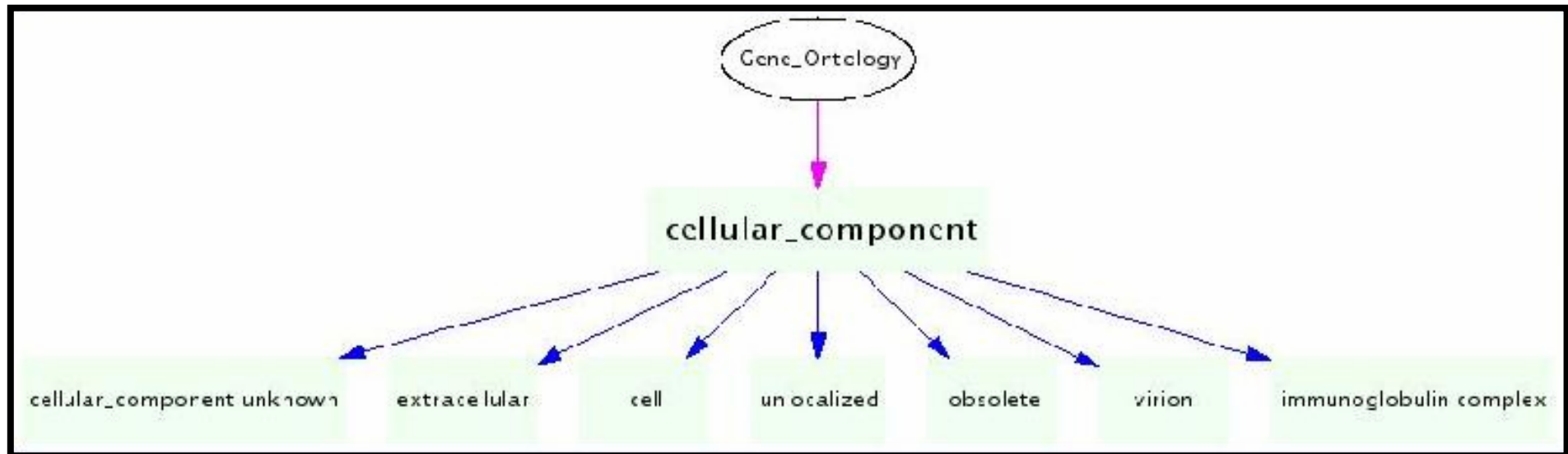




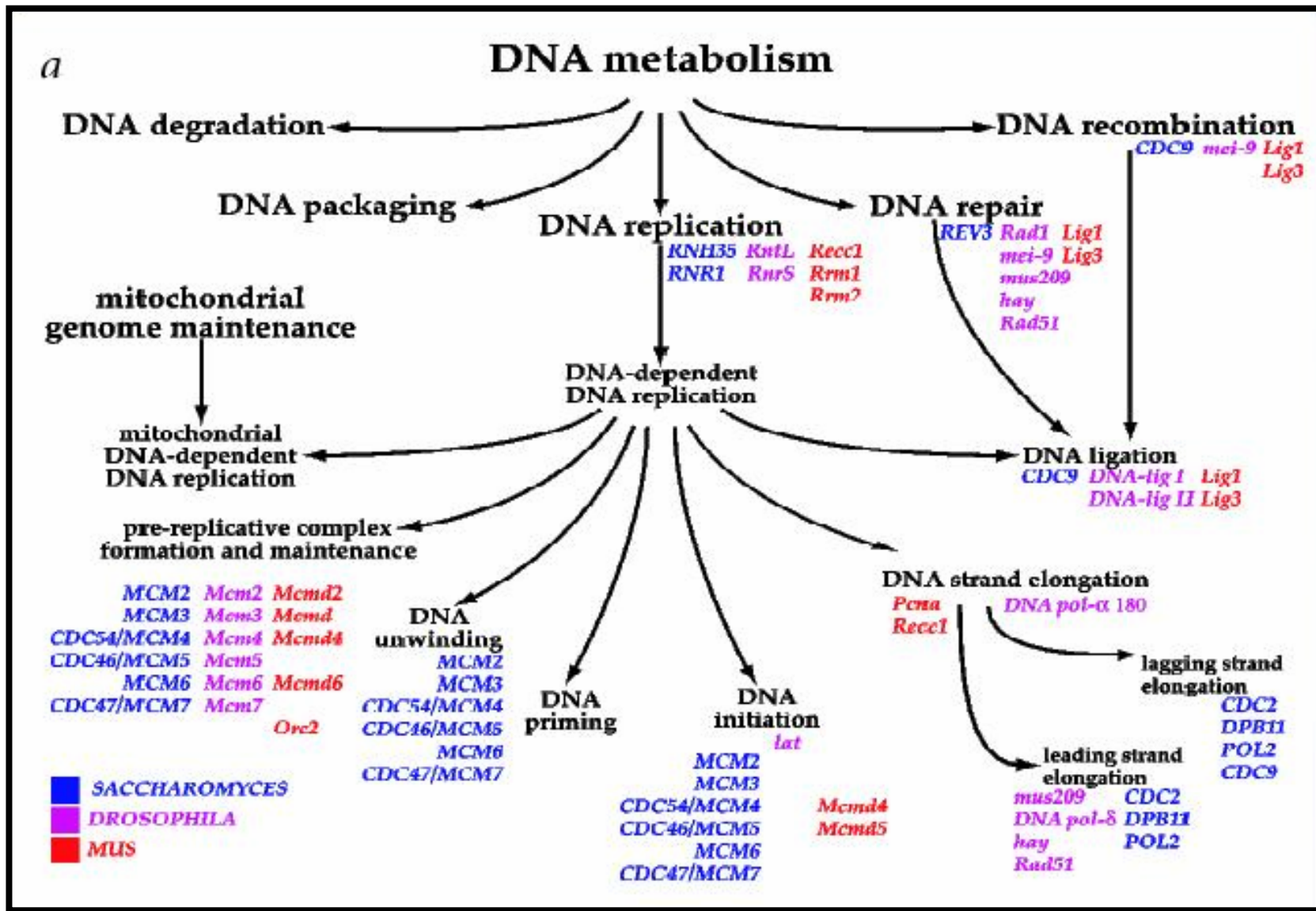
Queste 3 ontologie in GO prevedono l'uso di due tipi di relazioni fra gli oggetti:

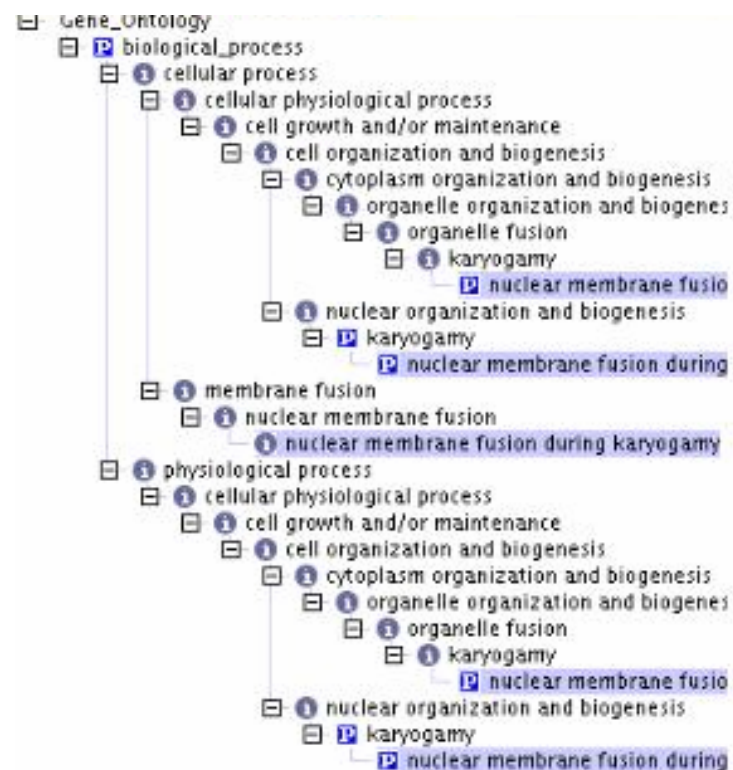
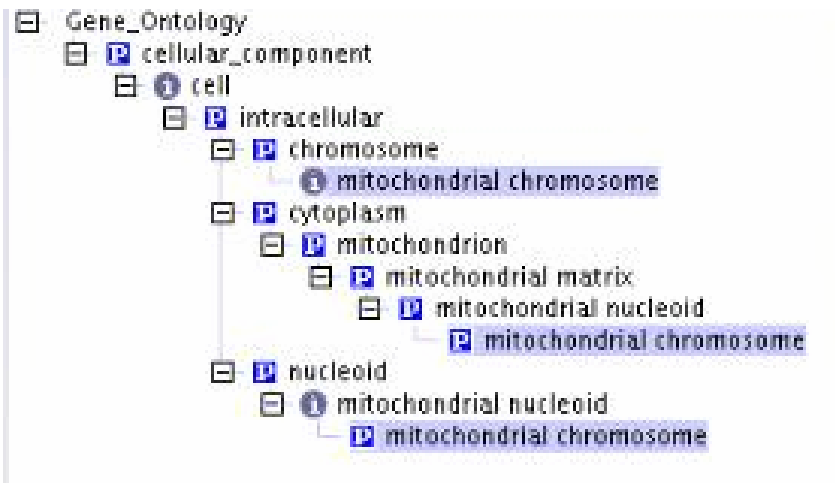
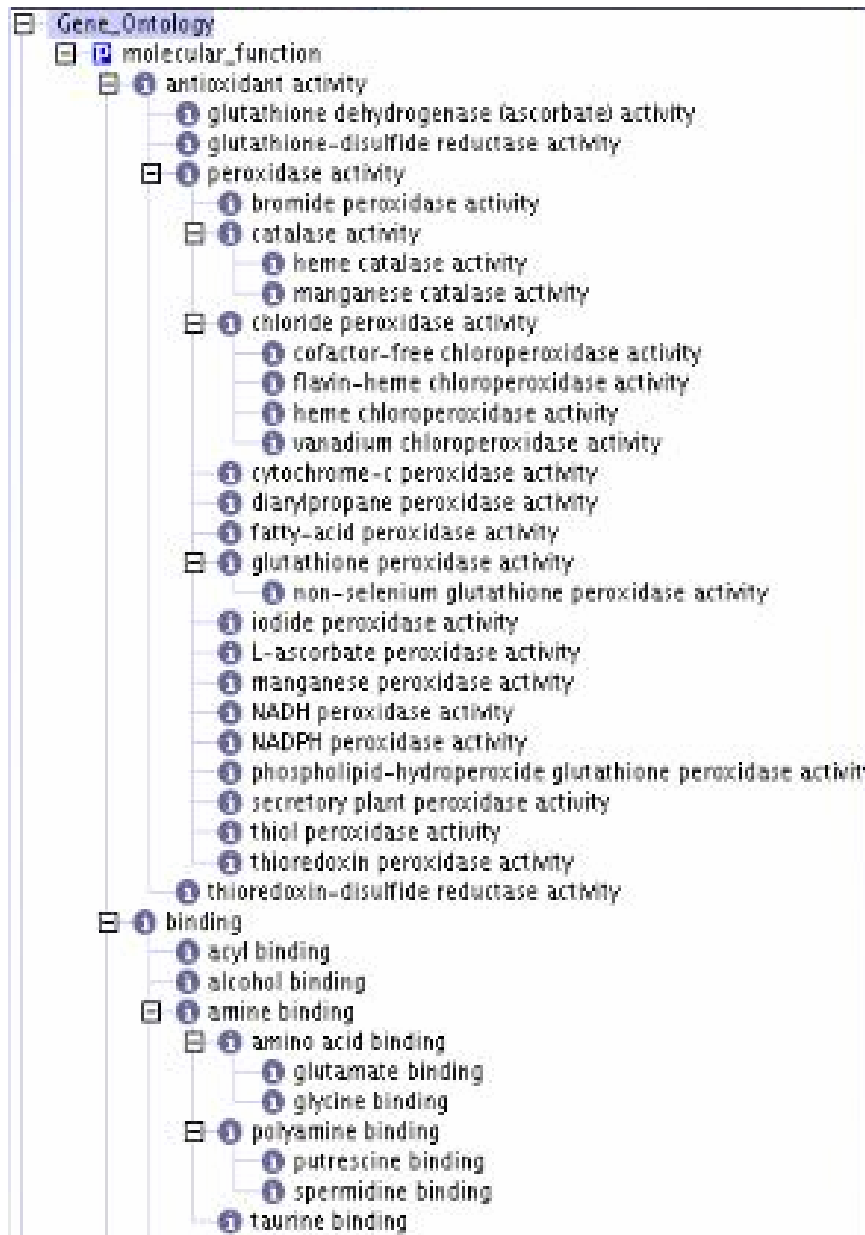
—————→ Istanza, è un

—————→ Parte / Tutto, fa parte di



Qualsiasi concetto espresso da un termine della gene ontology si può rapidamente conoscere il contesto in cui si esprime (nodi con valore concettuale più ampio) o discendere verso concetti con valori più stringenti o parziali





GO category	p-value in 3172	p-value in 662	p-value in 178
regulation of cell proliferation	.00161	Non significativo	.00161
apoptosis	.00272	.0017	.00272
programmed cell death	.00313	.00182	.00286
positive regulation of programmed cell death	.00401	.000161	.00313
positive regulation of apoptosis	.00401	.000161	.00361
induction of programmed cell death	.00401	.000161	.00401
induction of apoptosis	.00401	.000161	.00401
regulation of programmed cell death	.00401	.000161	.00401
cell growth and/or maintenance	.0048	.00355	.0048
regulation of apoptosis	.001	.000445	Non significativo
cell death	Non significativo	.00264	Non significativo
death	Non significativo	.00281	Non significativo

GO non è continuamente aggiornata.

L'unica fonte di informazione biologica aggiornata è
PUBMED.

Informazioni estraibili:

- Il contesto in cui il gene funziona
- Fenotipo
- Correlazione con altri geni

PRIMA METODOLOGIA

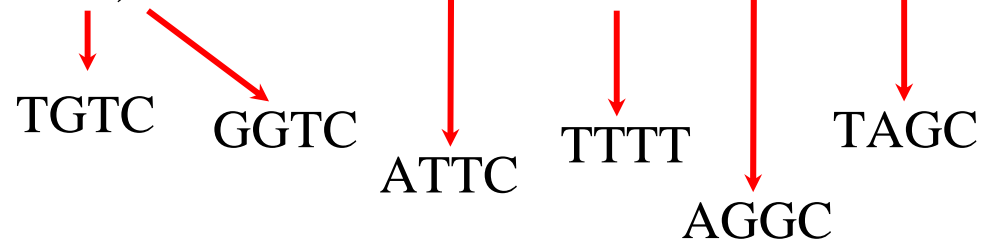
DIZIONARIO

Using BLAST for identifying gene and protein names in journal articles

Michael Krauthammer ^{a,*}, Andrey Rzhetsky ^{a,b}, Pavel Morozov ^b, Carol Friedman ^{a,c}

Conversione di tutte le lettere i simboli i numeri in un codice basato sui 4 nucleotidi

and VEGFR-3), and soluble form of VEGFR-1



ATTCGATCGACGATTTTAGCCCAGCTAGCCAGCTAGCCATAGC
AGTTCCCAGCTAGACAGCTAGC

ARPM2 PRDM16 EGFR LOC401936 KIAA0450

TACAGCCTTCCCACGTTTTAGCTAGAGTCACACAAAGTTTTGC
TAGC

Using BLAST for identifying gene and protein names in journal articles

Michael Krauthammer ^{a,*}, Andrey Rzhetsky ^{a,b}, Pavel Morozov ^b, Carol Friedman ^{a,c}

Abstract

AATC GTGA CGTACAGCAGTACAAA

GenBank

AGTG GTGA ACTACATCCATACAAA

Abstract NOTCH2

Non c'è soluzione

GenBank NOTCH

P=71,7% R=78,8%

Valutazione degli algoritmi

Recall (R) = $TP / (TP+FN)$ (= numero di geni totali dell'articolo)

Precision (P) = $TP / (TP+FP)$ (= numero geni riconosciuti)

F-score = $2 * P * R / (P + R)$

SECONDA METODOLOGIA

COSTRUZIONE REGOLE

**Toward Information Extraction:
Identifying protein names from biological papers**

K. FUKUDA, T. TSUNODA, A. TAMURA, T. TAKAGI

Core-Term

- Src homology (SH) 2 and SH3 domains
- p54 SAP kinase

Feature-Term

- EGF receptor
- Ras GTPase-activating protein (GAP)

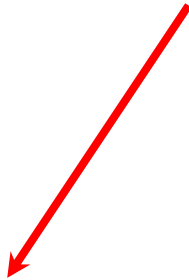


FILTRAGGIO

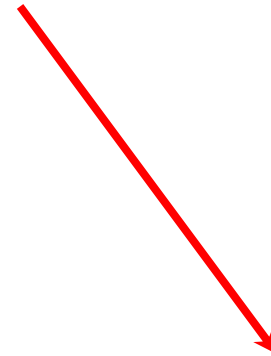
FILTRAGGIO



CONCATENAZIONE



Vicinanza



POS tagger

FILTRAGGIO

- capital letters (e.g., ADA, CMS)
- Arabic numerals (e.g., ATF-2, CIN85)
- Roman alphabets (e.g., Fc alpha receptor, 17beta-estradiol dehydrogenase)
- Roman numerals (e.g., dipeptidylpeptidase IV, factor XIII)
- frequent words appearing in protein names (e.g., myelin basic protein, PI 3-kinase, nerve growth factor)
 - words that have a capital letter in the beginning followed by more than three lower letters (e.g., According, Basically)
 - words that are composed of only capital letters longer than 6 characters. (e.g., KTPGKKKKKGK)
 - only one character other than 'V' (i.e., A, B, ..., U, W, ..., Y, Z)
 - measuring units (e.g., nM, MM, mM, pH, MHz)
 - chemical formulas (e.g., CaCl₂, NH₂, Ca₂, HCl, Mg₂)

ERRORI:

Riferimenti non corretti

Concatenazione

Ras guanine nucleotide exchange factor Sos
→ Ras guanine nucleotide exchange factor Sos

p85 alpha → p85 alpha

the focal adhesion kinase (FAK) → the focal adhesion kinase (FAK)

P= 94,70% R=98,84%

TERZA METODOLOGIA

METODOLOGIE STATISTICHE

Automatic Term Identification and Classification in Biology Texts

Chikashi Nobata, Nigel Collier and Jun-ichi Tsujii

Metodologie di selezione dei termini da classificare:

1. Shallow parsing (EngCG)
2. Alberi decisionali
3. Identificazione statistica



Classificazione

Internal evidence, Naive Bayes

Alberi decisionali: 45 categorie (nomi geni e proteine, linee cellulari, tessuti...)

Combinazione di caratteri

Valutazione dell'algoritmo

Class	class. only	I1	I2	I3
SOURCE	69.9	37.9	53.4	37.1
PROTEIN	70.3	31.9	53.8	33.6
DNA	83.8	42.3	50.0	47.4
RNA	8.2	2.7	4.1	6.6
All	65.8	37.4	58.5	40.1

Table 2: F-scores for **C1** on 100 abstracts with different term identification methods

Class	class. only	I1	I2	I3
SOURCE	77.19–82.30	15.87–24.77	46.33–55.14	22.99–29.74
PROTEIN	83.43–87.53	33.17–40.81	63.37–72.10	42.87–47.95
DNA	17.86–44.59	2.40– 4.20	4.95–14.81	7.29–19.05
RNA	0.00– 0.00	0.00–16.67	0.00– 0.00	0.00– 0.00
All	87.72–90.10	28.93–31.31	56.98–66.24	37.85–42.22

Table 3: F-scores for **C2** on 100 abstracts with different term identification methods

SHALLOW PARSING (EngCG)

Analisi lessicale ed Analisi sintattica

Analisi lessicale

Individuazione classe di appartenenza

Scelta di regole euristiche per chiarire le ambiguità

Analisi sintattica

Individuazione frasi grammaticalmente eterogenee



Estrazione regole sintattiche

Frase non analizzata



Frase diagnostica



Modellare regole

ENGCG

Type one or more English sentences (max. 100 words). For best results, use proper capitalization and punctuation.

Parse Reset Use heuristics

Time flies like an arrow.

(See the description of [morphological tags](#), [syntactic tags](#) and [other notations](#).)

```
"<*time>"
    "time" <*> N NOM SG @SUBJ @NN> @ADVL
"<flies>"
    "fly" <SVO> <SV> V PRES SG3 VFIN @+FMAINV
    "fly" N NOM PL @SUBJ
"<like>"
    "like" PREP @<NOM
    "like" <SVOC/A> <SVO> <SV> V PRES -SG3 VFIN @+FMAINV
"<an>"
    "an" <Indef> DET CENTRAL ART SG @DN>
```

ENGCG

Type one or more English sentences (max. 100 words). For best results, use proper capitalization and punctuation.

Parse

Reset

[Use heuristics](#)

Time flies like an arrow.

(See the description of [morphological tags](#), [syntactic tags](#) and [other notations](#).)

```
"<*time>"
    "time" <*> N NOM SG @SUBJ
"<flies>"
    "fly" <SVO> <SV> V PRES SG3 VFIN @+FMAINV
"<like>"
    "like" PREP @ADVL
"<an>"
    "an" <Indef> DET CENTRAL ART SG @DN>
```