INFORMAZIO BIOLOGIKOA WEB SEMANTIKOAREN BIDEZ KUDEATZEKO DISEINU PATROIAK

Mikel Egaña Aranguren, PhD
http://mikeleganaaranguren.com

GAUR EGUNGO INFORMAZIO BIOLOGIKOAREN KUDEAKETA

Biologia molekularrak organismo bizidunei buruzko datuak ekoizten ditu. Giza genoma proiektuan, esaterako, biologia molekularraren teknikak erabiliz giza genoma osoa sekuentziatu zen. 1980ko hamarkadan teknika horietako asko plazaratu ziren, "biologia molekularraren iraultza" deritzon prozesuan. Iraultza horren ondorioz, biologoek datu biologikoak ekoizteko duten ahalmena etengabe handitu da.

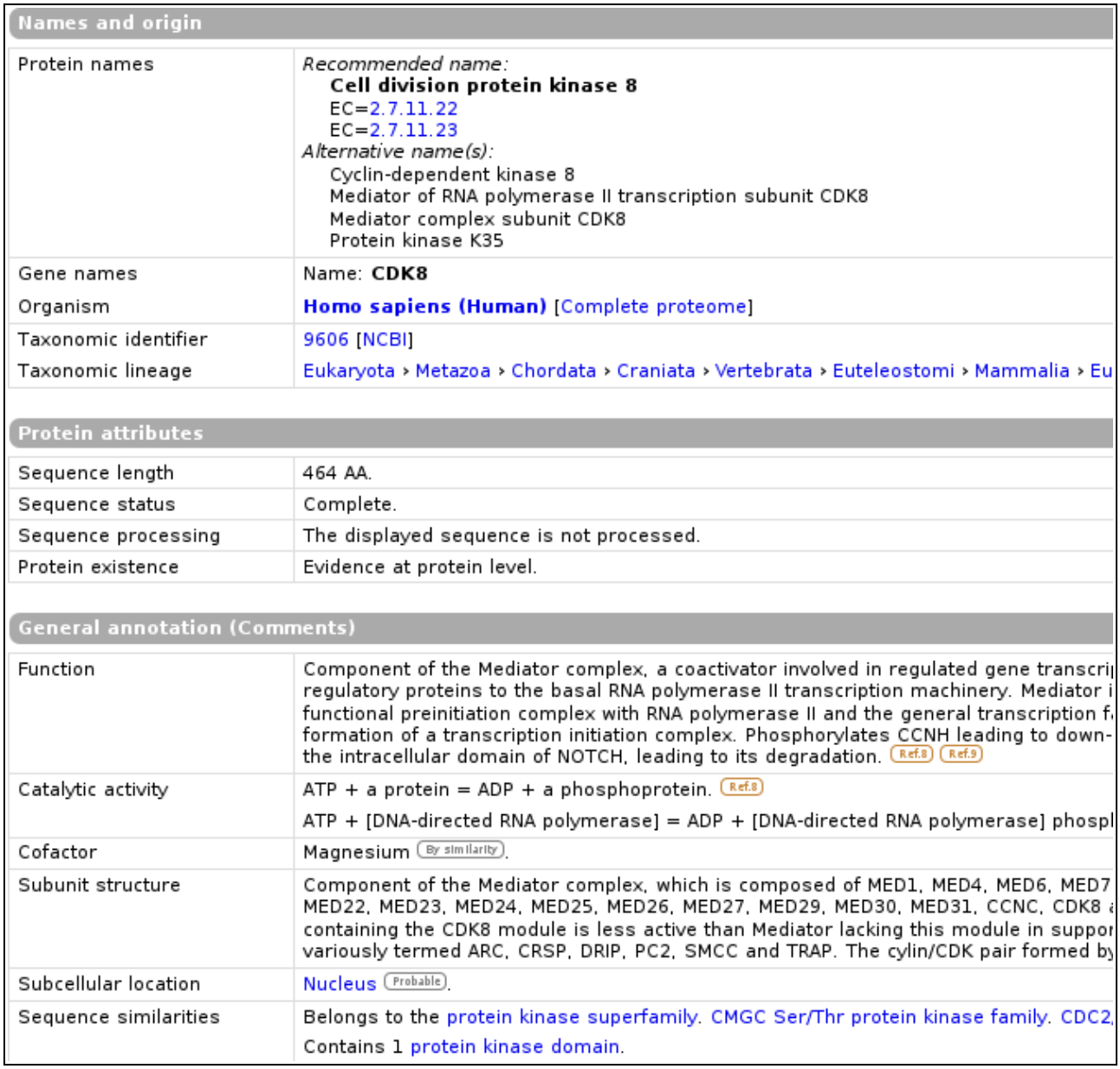
Datu biologikoetatik abiatuta informazio biologikoa lor daiteke. Adibidez, proteina baten sekuentzia laborategian lortuta (datuak), funtzio ezaguneko beste proteinen antzeko sekuentzia baldin badu, antzeko funtzioa izan dezake (informazioa). Informazio biologikoak teknika molekularren bidez lortutako datuei kontestu biologiko bat ematen die, biologoek ondorioetara heltzeko edo hipotesi berriak lortzeko erabil dezaketena.

Beraz, informazio biologikoa gaur egungo ikerketa biologikoaren muina da, eta urteetan zehar informazio hori biltzen duten hainbat datu base sortu dira. Adibidez, CDK8 proteinak UniProt datu basean sarrera bat du [1], bertan CDK8ri buruzko informazioa aurki daitekeelarik (Ikus 1 irudia). Sarrera horrek CDK8ren gaineko informazio gehiago (elkarrekintzak, proteinen egiturak, eta abar) deskribatzen duten beste datu baseetara doazen estekak ditu. Datu baseotan milaka proteinei buruzko informazio elkar erlazionatua aurki daitekeenez, informazio biologikoa oso konplexua eta hedapen handikoa da. Kontuan hartu behar da datu baseok proteinei buruzko informazioaz gain informazio gehiago gordetzen dutela, konplexutasuna areagotuz.

Biologo batentzat informazio hori guztia era efizientean erabiltzea oso zaila da. Demagun biologo batek CDK8rekiko interesa daukala, eta galdera hau ebatzi nahi duela: "CDK8ren ortologoren batekin fosforilazioz elkarrekiten duen proteinaren bat nukleoan kokatua dago?" Galdera horren erantzuna datu baseetan dago, baina datu base desberdinen informazioa konbinatu behar da. Biologoak datu base batetik bestera jo beharko du, informazioa eskuratuz, prozesatuz, eta hurrengo datu base batera joz informazio berriarekin (Ikus 2 irudia). Gainera, jatorrizko galderan proteina bat baino gehiago izanez gero, edo informazio oso desberdina behar izanez gero, prozesua asko konplikatzen da. Beraz, biologoek datu baseei ezin diete galdera konplexurik zuzenean egin, ikerketa biologikoak aurrera arinago egitea oztopatuz.

Arazo honen muina datu baseetako informazioa adierazteko moduan datza: informazioa gizakiak ulertzeko moduan, testu arruntean, soilik argitaratzen da, ordenagailuek "uler" dezaketen modu batean publikatu beharrean. Informazioa ordenagailuek uler dezaketen formato batean argitaratuz, biologoek ordenagailuek informazio konplexua prozesatzeko duten ahalmena erabil dezakete. Hau da, informazio kudeaketa ordenagailuaren baitan delega daiteke, teknologia egokia erabiliz gero.

Teknologia horri Web Semantikoa deritzo, eta, teknologia oro bezala, baditu bere zailtasunak. Lan honen xedea biologoei informazioa Web Semantikoan argitaratzea erraztea izan da, hurrengo ataletan azaltzen den moduan.



Names and origin

Protein names	*Recommended name:* **Cell division protein kinase 8** EC=2.7.11.22 EC=2.7.11.23 *Alternative name(s):* Cyclin-dependent kinase 8 Mediator of RNA polymerase II transcription subunit CDK8 Mediator complex subunit CDK8 Protein kinase K35
Gene names	Name: **CDK8**
Organism	**Homo sapiens (Human)** [Complete proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eu

Protein attributes

Sequence length	464 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is not processed.
Protein existence	Evidence at protein level.

General annotation (Comments)

Function	Component of the Mediator complex, a coactivator involved in regulated gene transcri regulatory proteins to the basal RNA polymerase II transcription machinery. Mediator i functional preinitiation complex with RNA polymerase II and the general transcription f formation of a transcription initiation complex. Phosphorylates CCNH leading to down- the intracellular domain of NOTCH, leading to its degradation. Ref.8 Ref.9
Catalytic activity	ATP + a protein = ADP + a phosphoprotein. Ref.8 ATP + [DNA-directed RNA polymerase] = ADP + [DNA-directed RNA polymerase] phospl
Cofactor	Magnesium By similarity.
Subunit structure	Component of the Mediator complex, which is composed of MED1, MED4, MED6, MED7 MED22, MED23, MED24, MED25, MED26, MED27, MED29, MED30, MED31, CCNC, CDK8 containing the CDK8 module is less active than Mediator lacking this module in suppor variously termed ARC, CRSP, DRIP, PC2, SMCC and TRAP. The cylin/CDK pair formed by
Subcellular location	Nucleus Probable.
Sequence similarities	Belongs to the protein kinase superfamily. CMGC Ser/Thr protein kinase family. CDC2, Contains 1 protein kinase domain.

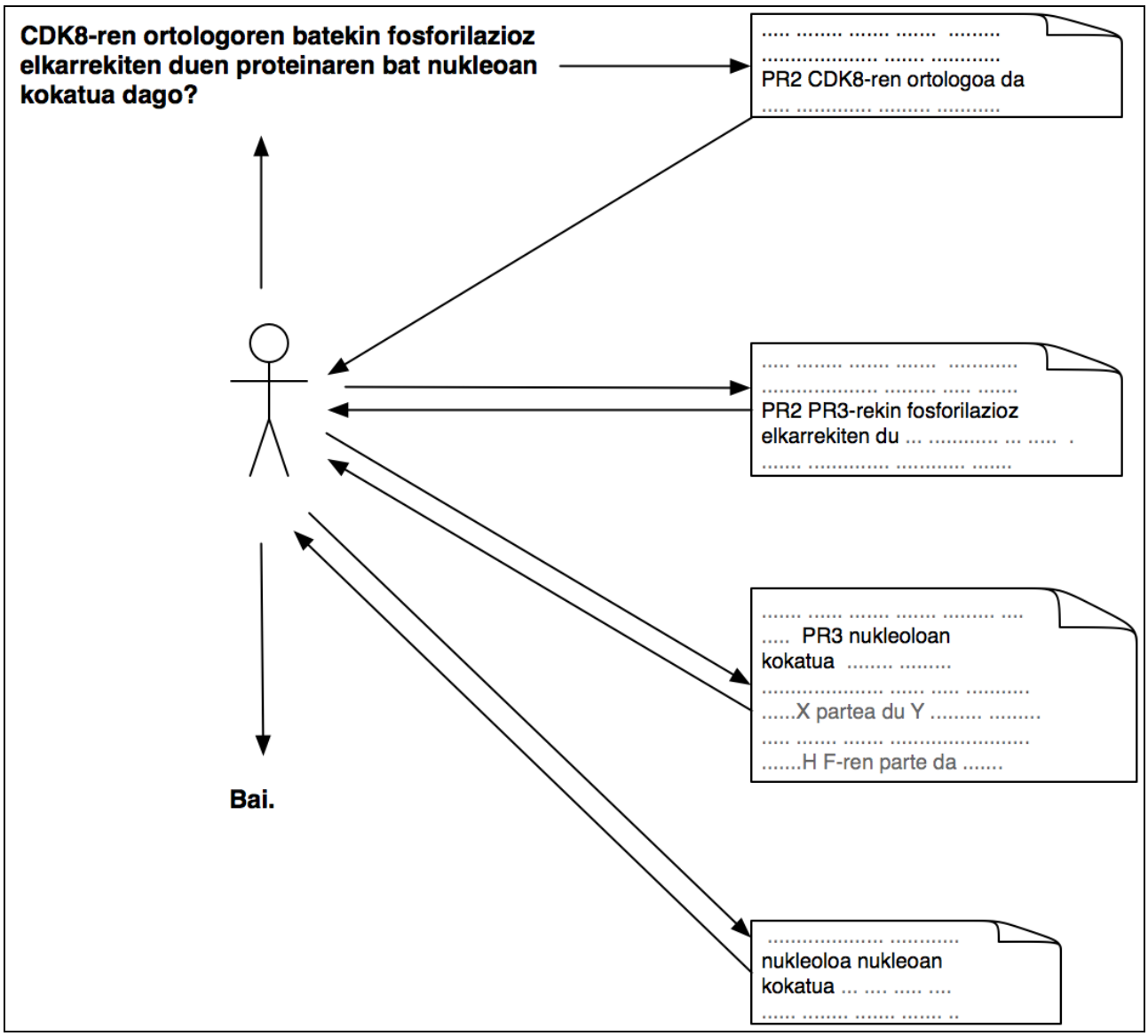
1 irudia. CDK8ren UniProt sarreraren zatia. Sarrerak CDK8ren gaineko informazioa erakusten du.

INFORMAZIO BIOLOGIKOAREN WEB SEMANTIKOA

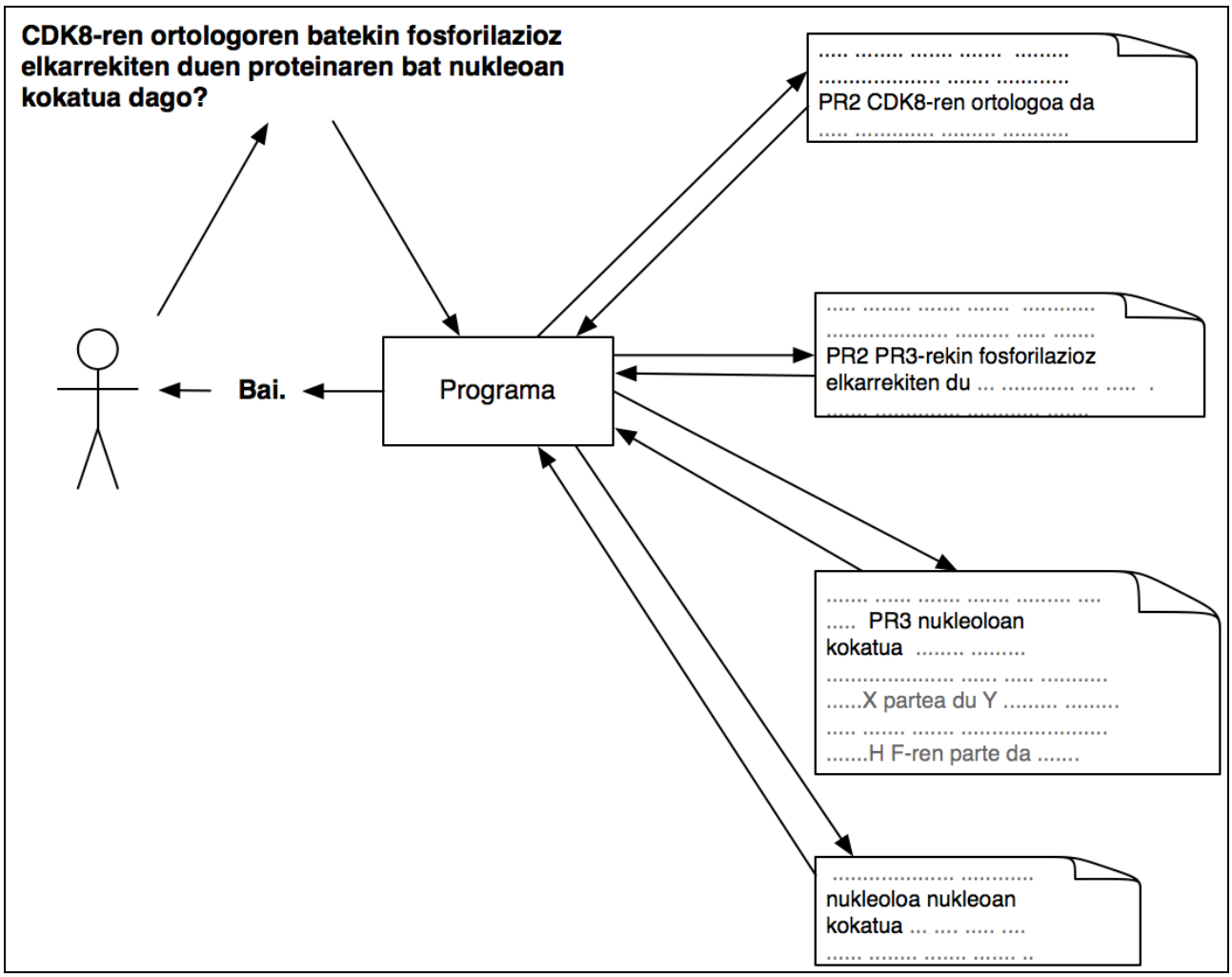
Web Semantikoa gaur egungo interneten balizko bertsio hobetua da, non ordenagailuek interneteko informazioaren esanahia “ulertuko” duten [2]. Web Semantikoaren teknologia informazio biologikoa argitaratzeko eta kudeatzeko erabil daiteke. Hain zuzen ere, 2 irudiko kasura bueltatuz, Web Semantikoan biologoak bere galdera konplexua programa bati egingo lioke, programatik erantzun egokia jasoz. Programak informazioa datu base desberdinetatik eskuratu eta prozesatuko luke, biologoari erantzun egokia denbora laburrean emanez. Web Semantikoa erabiliz ikerketa arintzen da, biologoa ez delako datu base batetik bestera ibili behar, bidean informazioa erauziz eta prozesatuz (Ikus 3 irudia).

Web Semantikoa posible egiteko informazioa ordenagailuentzat ulergarria den era batean adierazi behar da. Ordenagailuek uler dezaketen hizkuntza bat logika da; logika erabiliz, axiomen bidez, ideiak adieraz ditzakegu, ordenagailuek ideia horiek kudea ditzaten. Logika bidez adierazitako ideia batzuen taldeari “Ontologia” deritzo.

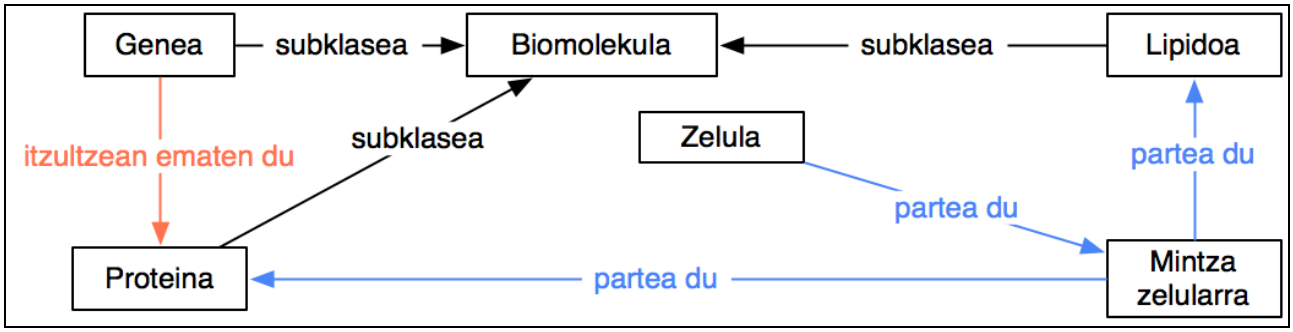
Informazio biologikoa biltzen duten ontologiei bio-ontologia deritze (Ikus 4 irudia). Bio-ontologien erabilera haziz doa, Web Semantikoa informazio biologikoaren kudeaketarako oso lagungarria baita [3].



2 irudia. CDK8ari buruzko informazioa lortzeko prozesua. Biologoak galdera bat du (goian), eta galdera hori erantzuteko datu base batetik bestera doa, pauso bakoitzean datu basetik informazio egokia erauziz eta lehendik zeukan informazioarekin batera prozesatuz.



3 irudia. 2 irudiko egoera bera, Web Semantikoan. Biologoak galdera programa bati egiten dio, eta programak egiten du lan guztia, denbora aurreztuz.



4 irudia. Kontzeptu batzuk eta beraien arteko erlazioak adierazten dituen bio-ontologia. Erlazioek definizio logiko bat dute, ordenagailuak uler dezakeena. Ondorioz, ordenagailuarentzat bio-ontologiaren egitura da garrantzitsuena, izenak ("Zelula", "partea du") gizakiontzat soilik baliagarri direlarik.

BIO-ONTOLOGIAK ERAIKITZEKO DISEINU PATROIAK

Ontologiak sortzeko lengoaia desberdinak badaude ere, azken urteotan OWL (Web Ontology Language) lengoaiaren erabilera hazi da [4] (Ikus 5 eta 6 irudiak). OWL oso potentea denez, oso kontzeptu konplexuak deskribatzen dituzten bio-ontologiak sortzeko erabil daiteke. Horrelako bio-ontologiak kalitate handikoak dira, informazioarekin elkarrekintza aberatsa ahalbidetzen baitute. Baina biologoentzat OWLen potentzia erabiltzea zaila da, ideiak adierazteko logika ez delako hizkuntza intuitiboa. Beraz, OWLen erabilera errazten duten teknikak beharrezkoak dira, kalitate handiko bio-ontologiak plazaratu nahi baditugu. Teknika horietako bat bio-ontologiak eraikitzerakoan diseinu patroiak erabiltzea da.

Diseinu patroia bio-ontologiak eraikitzerakoan askotan agertzen den arazo konkretu bat ebazten duen soluzio efizientea da, goitik behera dokumentatua. Hainbat diseinu patroi interneten publikatzen direnez, biologo batek, bere bio-ontologia erakitzean arazo baten aurrean, diseinu patroi aproposa aurkitu eta aplikatu besterik ez du, errezeta bat legez. Beraz, diseinu patroiek OWLen erabilera errazten dute.

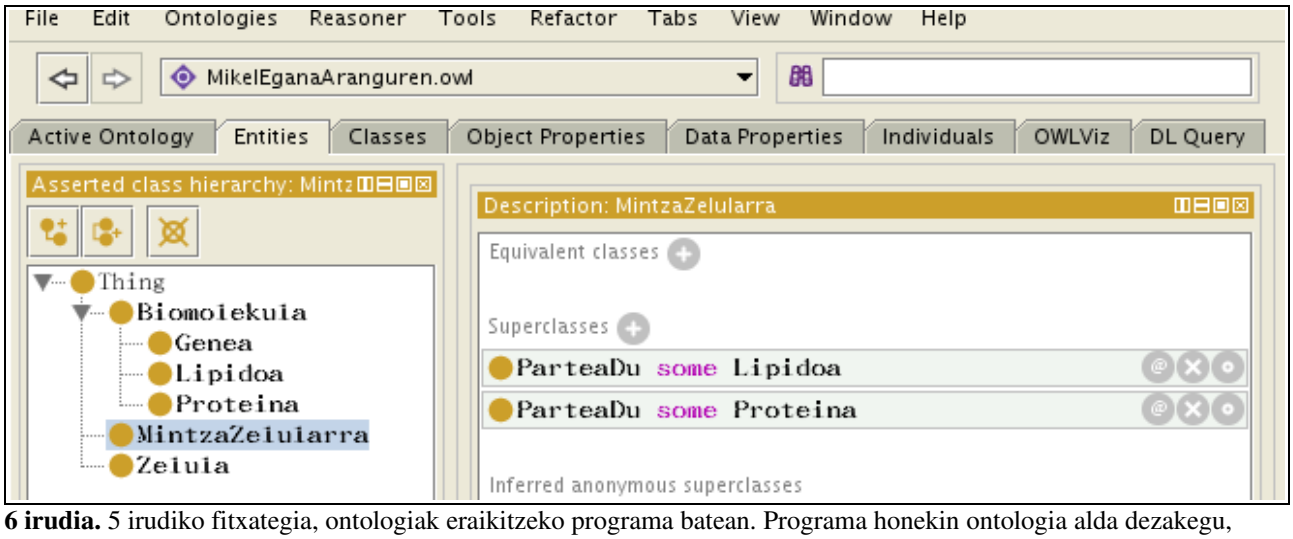
Adibidetzat “Value Partition” (VP) [5] diseinu patroiaren erabilera har dezakegu (Ikus 7 irudia). Baliteke biologo bat, OWL bio-ontologia bat eraikitzen dabilena, erregulazio prozesu baten kontzeptua bere bio-ontologian adierazi nahi izatea. Erregulazio prozesuak bi balio ditu (erregulazio positiboa edo negatiboa), eta gehiagorik ez. Biologoak ez daki OWL erabiliz ideia hori adierazten. Ondorioz, biologoak diseinu patroi desberdinen dokumentazioa irakurtzen du, eta VP diseinu patroiak bere beharrak betetzen dituela konturatzen da (balio jakin batzuk bakarrik dituen propietate bat). VP diseinu patroia bere bio-ontologian aplikatzean, automatikoki axioma oso konplexuak gehitzen dizkio bio-ontologiari, bio-ontologia oso denbora laburrean eta oso modu efizientean aberastuz. Beraz, diseinu patroiak erabiliz, biologo batek kalitate handiko bio-ontologiak eraiki ditzake, esfortzu txikiarekin.

Lan honek biologoek diseinu patroiak erabil ditzaten ideiak eta teknikak garatu ditu. Lanaren ekarpen nagusia biologoen arteko diseinu patroien ideiaren dibulgazioa da [6]. Halaber, lanak ekarpen konkretuak izan ditu, hala nola, bio-ontologientzako diseinu patroien katalogo bat [7], diseinu patroiak automatikoki aplikatzeko programazio lengoaia sinple bat [8], eta gaur egun erabiltzen diren bio-ontologietan diseinu patroien aplikazioaren adibideak [9,10].

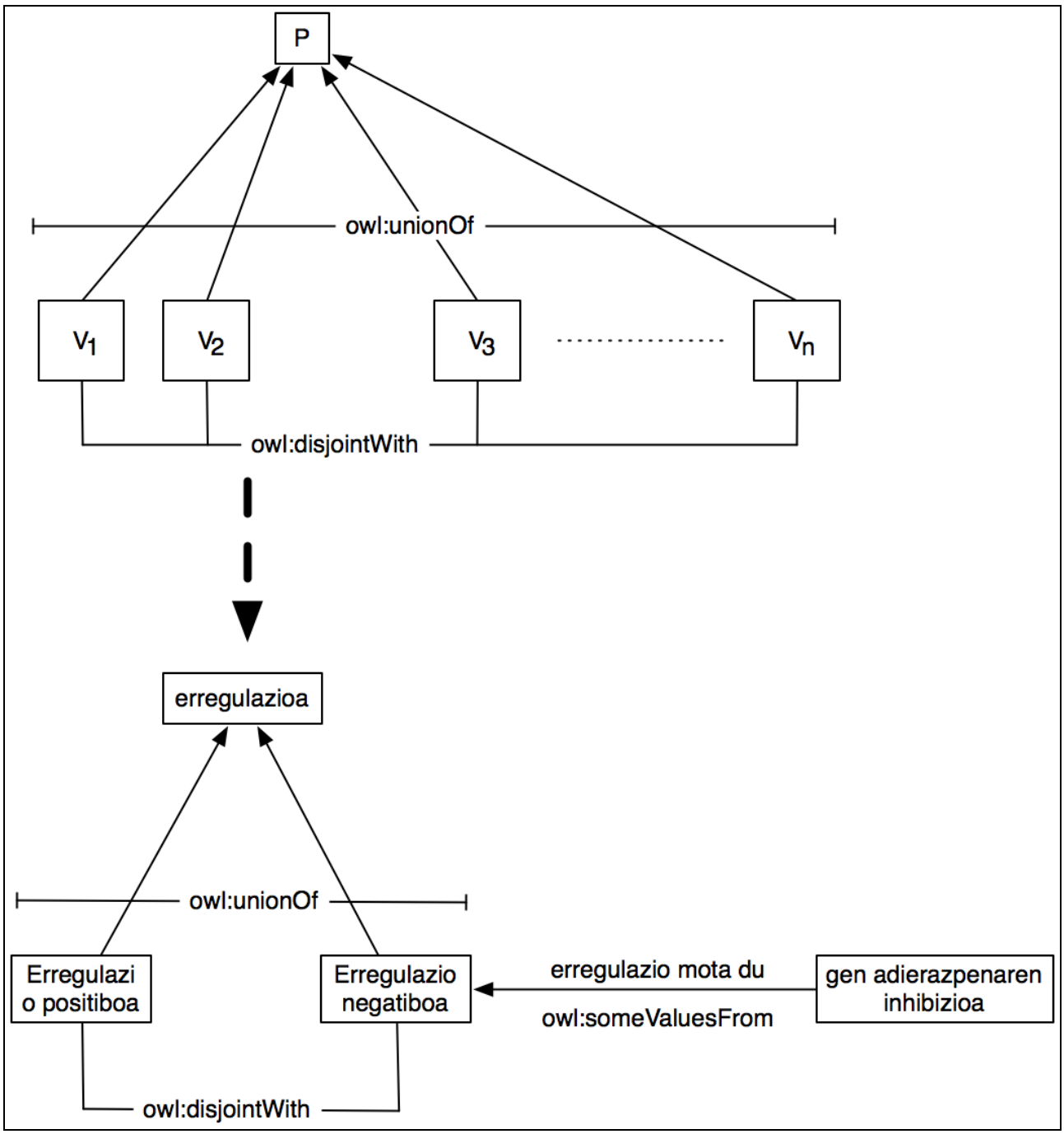
Biologoek bio-ontologia aberatsagoak eraiki ditzakete diseinu patroiak erabiliz, gero eta informazio gehiago Web Semantikoan plazaratuz. Web Semantikoan ordenagailuen potentzia informazio kudeaketa automatikoa burutzeko erabil daiteke, gizakiak baino askoz arinago eta modu efizienteagoan. Minbiziaren kontrako terapietatik eboluzioaren ñabarduretaraino, ikerketa biologikoa informazio kudeaketan oinarritzen da; kudeaketa hori hobetzen duten teknikek, lan honetan plazaratutakoek kasu, eragin nabaria izango dute gaur eguneko biologian.

```
<?xml version="1.0"?>

<!DOCTYPE rdf:RDF [
    <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
    <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
    <!ENTITY owl2xml "http://www.w3.org/2006/12/owl2-xml#" >
    <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
    <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
    <!ENTITY MikelEganaAranguren "http://www.elhuyar.org/MikelEganaAranguren.owl#" >
]>

<rdf:RDF xmlns="http://www.elhuyar.org/MikelEganaAranguren.owl#"
     xml:base="http://www.elhuyar.org/MikelEganaAranguren.owl"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:owl2xml="http://www.w3.org/2006/12/owl2-xml#"
     xmlns:MikelEganaAranguren="http://www.elhuyar.org/MikelEganaAranguren.owl#"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <owl:Ontology rdf:about=""/>

    <owl:ObjectProperty rdf:about="#ParteaDu"/>

    <owl:Class rdf:about="#MintzaZelularra">
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty rdf:resource="#ParteaDu"/>
                <owl:someValuesFrom rdf:resource="#Lipidoa"/>
            </owl:Restriction>
        </rdfs:subClassOf>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty rdf:resource="#ParteaDu"/>
                <owl:someValuesFrom rdf:resource="#Proteina"/>
            </owl:Restriction>
        </rdfs:subClassOf>
    </owl:Class>

</rdf:RDF>
```

5 irudia. 4 irudiko bio-ontologia gordetzen duen fitxategiaren zati bat. Fitxategia OWL lengoaian idatzia dago.



6 irudia. 5 irudiko fitxategia, ontologiak eraikitzeko programa batean. Programa honekin ontologia alda dezakegu, adibidez kontzeptuak edo axiomak gehituz.



7 irudia. VP diseinu patroiaren erabilera. VP diseinu patroiaren egitura abstraktoa, goian, egitura konkretu batean bihurtzen da OWL bio-ontologia batean aplikatzerakoan, behean.

[1] http://www.uniprot.org/uniprot/P49336

[2] http://www.w3.org/standards/semanticweb/

[3] Benjamin M. Good, Mark D. Wilkinson. The Life Sciences Semantic Web is full of creeps! Briefings in bioinformatics, Vol. 7, No. 3. (2006), pp. 275-286.

[4] http://www.w3.org/standards/techs/owl

[5] http://www.gong.manchester.ac.uk/odp/html/Value_Partition.html

[6] Mikel Egaña Aranguren, Erick Antezana, Martin Kuiper, Robert Stevens. Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology. BMC bioinformatics 2008, 9(Suppl 5):S1.

[7] http://odps.sf.net/

[8] http://oppl.sf.net/

[9] Erick Antezana, Mikel Egaña, Bernard De Baets, Ward Blondé, Aitzol Illarramendi, Iñaki Bilbao, Bernard De Baets, Robert Stevens, Vladimir, Mironov, Martin Kuiper. The Cell Cycle Ontology: An application ontology for the representation and integrated analysis of the cell cycle process. Genome Biology 2009, 10:R58.

[10] Mikel Egaña, Alan Rector, Robert Stevens, Erick Antezana. Applying Ontology Design Patterns in bio-ontologies. EKAW 2008, LNCS 5268, pp. 7-16.