A knowledge based interface for distributed biological databases

Paolo Bresciani[+] and Paolo Fontana[*] and Paolo Busetta[+]

{brescian,pfontana,busetta}@itc.it.

(+)ITC-irst (TRENTO) and (*)IASMAA (San Michele a.A.)

with the collaboration of Giorgio Valle and Stefano Toppo

CRIBI - University of Padua

Outline of the Talk

- Motivation for new approaches in biological DB access
- The current state of the art (2 examples)
- Our Knowledge Based approach:
  - an example of interaction
  - some technical details
- Extending to multiple DBs

Biological Database access

The formulation of the *intended* query for retrieving the desired data is a problem for every database user.
As a simple example in Biology, consider the task of searching for *KDEL receptor*:

- what does the user exactly mean with *KDEL receptor*? Is she looking for the description of that functionality; or for any protein with that functionality; or for any genomic sequence that is expressed in such a protein?
- moreover, does the user really know all the consequences of looking for all (let's say) the protein having *KDEL receptor* functionality?

Biological Database access *cont'*

It may be very useful to know some relevant limitations on the form of the query, when already some constraints are imposed:
E.g., KDEL receptor function **can NOT** be exhibited by any protein in the cell nucleus.

Biological Database access *cont'*

It may be very useful to know some relevant limitations on the form of the query, when already some constraints are imposed:
E.g., KDEL receptor function **can NOT** be exhibited by any protein in the cell nucleus.

⟶ *"protein located in the nucleus and with KDEL receptor function"*
is inconsistent: submitting it to any biological DB results in a useless interaction (loss of time and money).

Main sources of errors in queries

Current query systems do not provide any support to avoid (or limit) the source of *conceptual errors* in queries.

Many sources of errors:

- Lack of knowledge on the domain

Main sources of errors in queries

Current query systems do not provide any support to avoid (or limit) the source of *conceptual errors* in queries.

Many sources of errors:

- Lack of knowledge on the domain
- Limited knowledge on some parts of the domain

Main sources of errors in queries

Current query systems do not provide any support to avoid (or limit) the source of *conceptual errors* in queries.

Many sources of errors:

- Lack of knowledge on the domain
- Limited knowledge on some parts of the domain
- Terminology disagreement

Main sources of errors in queries

Current query systems do not provide any support to avoid (or limit) the source of *conceptual errors* in queries.

Many sources of errors:

- Lack of knowledge on the domain
- Limited knowledge on some parts of the domain
- Terminology disagreement
- Little understanding of the domain *representation* inside the database: *terminology*, *taxonomy*, *relationships*, *constraints*

The current solutions

The common way to deal with the problem is by

- being as much expert as possible in the domain
- being as much aware as possible of the design and implementation details of the DB.

This may be sometimes interesting (domain knowledge), even if difficult, but also tedious (DB design and implementation details) specially when using several and changing DBs.

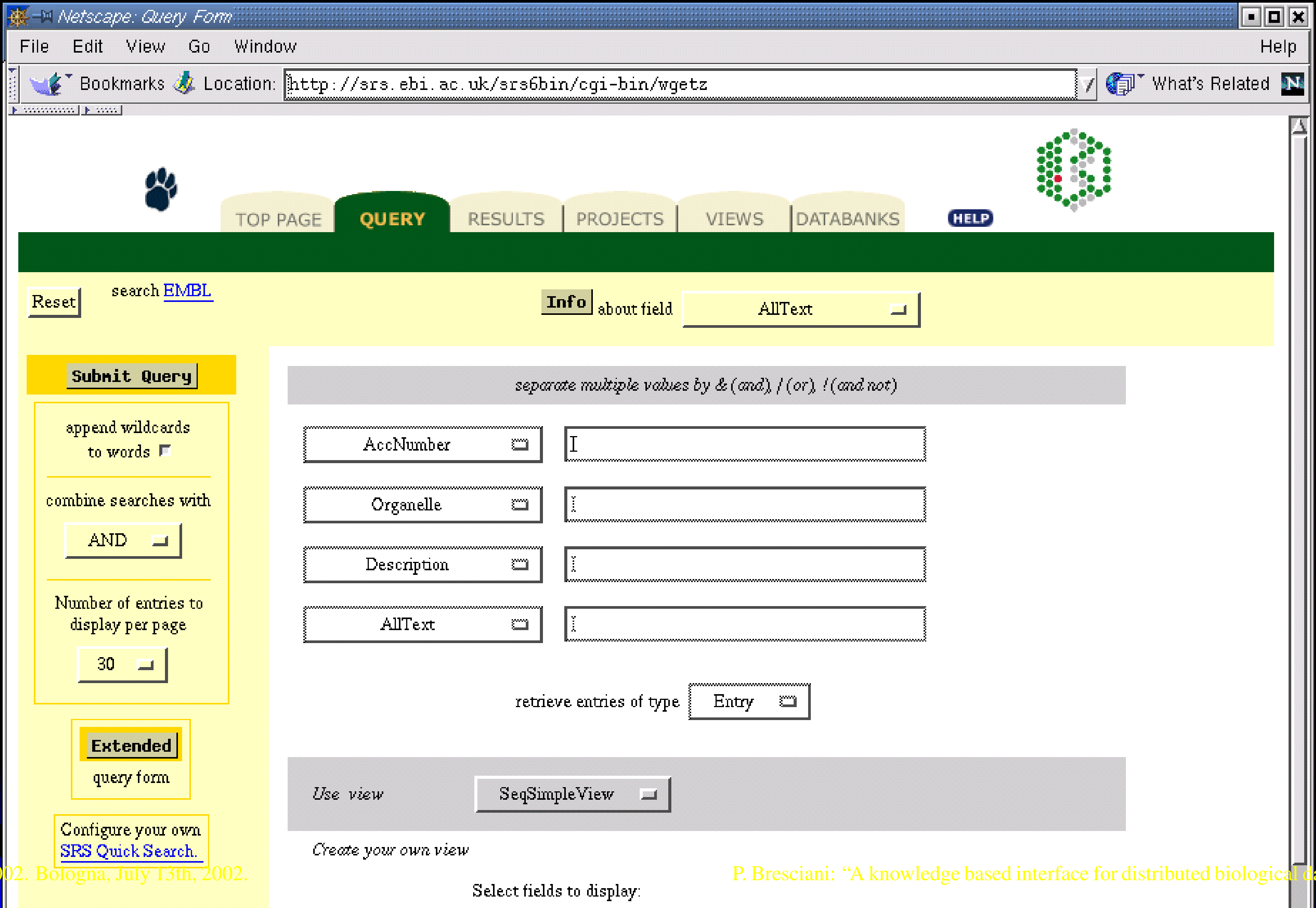
Our solution

We introduce a *concept-demonstrator* of a **knowledge based** Visual Query System. It has been applied in the context of the access to biological databases, with the following advantages for the user:

- allows to interactively and iteratively build **consistent queries only**;
- allows to **interactively explore** the database semantics by gradually browsing only the *interesting* parts of the conceptual model;
- uses simple, but effective, features for query refinement and generalization.

A QBE [Zloof] interface example

(The SRS: Sequence Retrieval System)



Netscape: Query Form

File Edit View Go Window Help

Bookmarks Location: http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz What's Related

Reset search EMBL

Info about field AllText

Submit Query

append wildcards to words □

combine searches with AND

Number of entries to display per page 30

Extended query form

Configure your own SRS Quick Search.

separate multiple values by & (and), | (or), ! (and not)

AccNumber

Organelle

Description

AllText

retrieve entries of type Entry

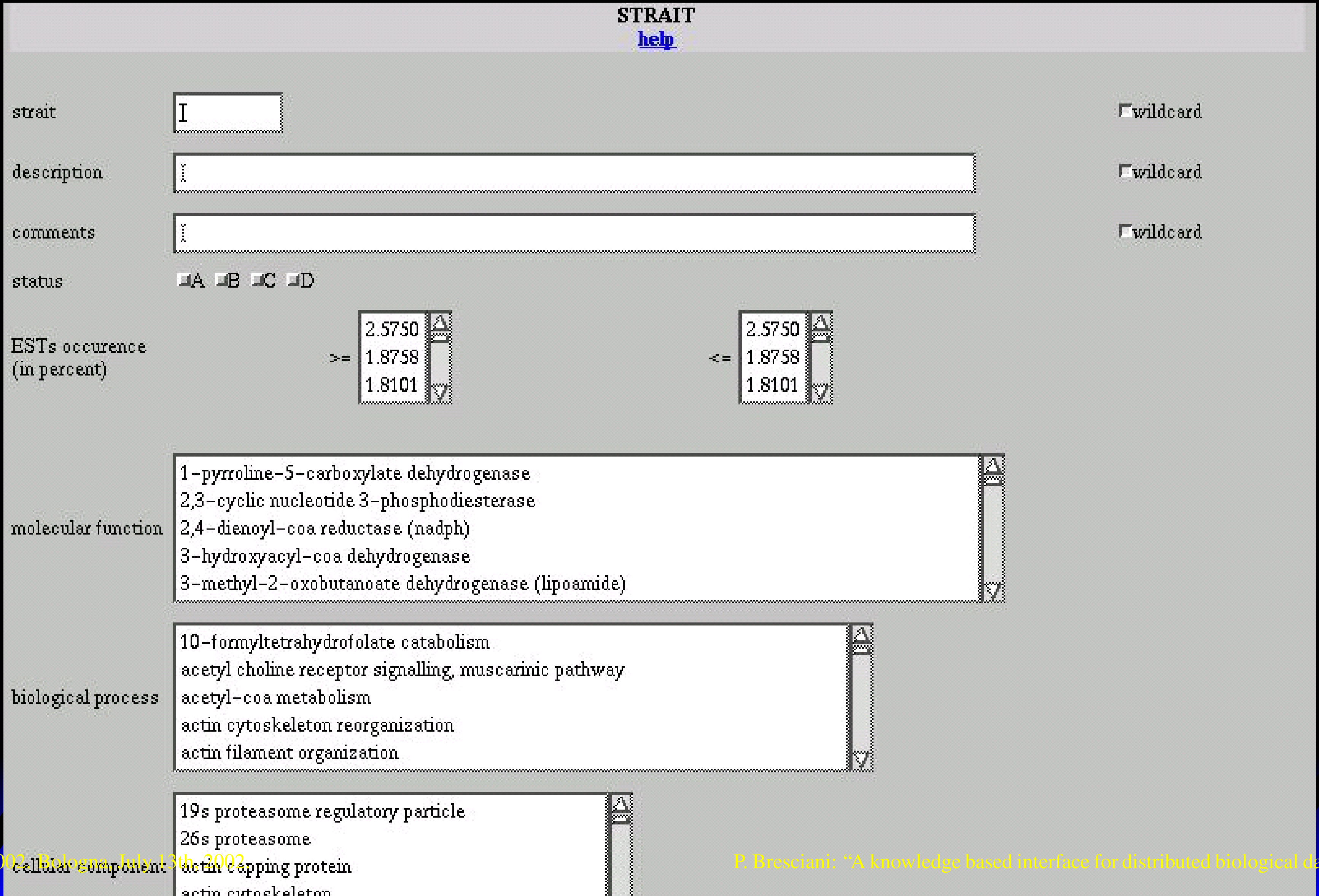
Use view SeqSimpleView

Create your own view

Select fields to display:

A slightly better example

(The muscle-trait DB — CRIBI-UniPD)



STRAIT

help

strait □wildcard

description □wildcard

comments □wildcard

status □A □B □C □D

ESTs occurence (in percent)

>= 2.5750 1.8758 1.8101

<= 2.5750 1.8758 1.8101

molecular function

1-pyrroline-5-carboxylate dehydrogenase
2,3-cyclic nucleotide 3-phosphodiesterase
2,4-dienoyl-coa reductase (nadph)
3-hydroxyacyl-coa dehydrogenase
3-methyl-2-oxobutanoate dehydrogenase (lipoamide)

biological process

10-formyltetrahydrofolate catabolism
acetyl choline receptor signalling, muscarinic pathway
acetyl-coa metabolism
actin cytoskeleton reorganization
actin filament organization

cellular component

19s proteasome regulatory particle
26s proteasome
actin capping protein
actin cytoskeleton

Problems and difficulties

- In the first case, matching strings must be provided

Problems and difficulties

- In the first case, matching strings must be provided
- In the second case, a more “guided” interface is available, but the selection still is among long lists of terms

Problems and difficulties

- In the first case, matching strings must be provided
- In the second case, a more “guided” interface is available, but the selection still is among long lists of terms

In any case no semantic support is provided.

The “flat files” legacy problem

```
ID   HSA010063  standard; DNA; HUM; 1730 BP.
AC   AJ010063;
SV   AJ010063.1
DT   01-OCT-1998 (Rel. 57, Created)
DT   07-JAN-2000 (Rel. 62, Last updated, Version 2)
DE   Homo sapiens telethonin gene
KW   telethonin gene.
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
RN   [1]
RP   1-1730
RA   Pallavicini A.L.;
RL   Submitted (06-AUG-1998) to the EMBL/GenBank/DDBJ databases.
RL   Pallavicini A.L., Complesso interdipartim. Vallisneri Dipartimento di
RL   Biologia, Universita di Padova, via G.Colombo 3, 35121, ITALY.
DR   SWISS-PROT; O15273; TELT_HUMAN.
FH   Key             Location/Qualifiers
FT   source          1..1730
FT                   /chromosome="17"
FT                   /db_xref="taxon:9606"
```

Terminology standardization

Fortunately some relevant steps ahead have been done in the last few years.

Terminology standardization

Fortunately some relevant steps ahead have been done in the last few years.
In particular **GeneOntology** is one of the most important efforts toward terminology standardization.

Terminology standardization

Fortunately some relevant steps ahead have been done in the last few years.

In particular **GeneOntology** is one of the most important efforts toward terminology standardization.

It aims at providing a support for data-integration and inter-operability among *sequence data* and data from *functional analyses*.

This is crucial for the discovery of the functions of new sequences by comparison with already studied and annotated sequences.

Molecular Function, *Biological Process*, and *Cellular Component* are classified in three hierarchies.

The Knowledge Based Approach

Intelligent Interface

Need of Intelligent Interfaces that must:

- be easy and intuitive to be used
- be “natural” to be used and understood
- require no knowledge of DB technologies or data representation formalisms
- possibly require only little or partial knowledge of the application domain
- be capable to give some semantic advice to the users

⟶ reduce user’s cognitive effort.

Semantics based approach for query formulation support

An approach is needed that:

- is semantically well founded
- is based on a representation of the model/schema of the DB and of the domain (biology)
- support specific kinds of reasoning (terminological reasoning)

Semantics based approach for query formulation support

An approach is needed that:

- is semantically well founded
- is based on a representation of the model/schema of the DB and of the domain (biology)
- support specific kinds of reasoning (terminological reasoning)

$\longrightarrow$ a Knowledge Based approach

The ingredients of our system

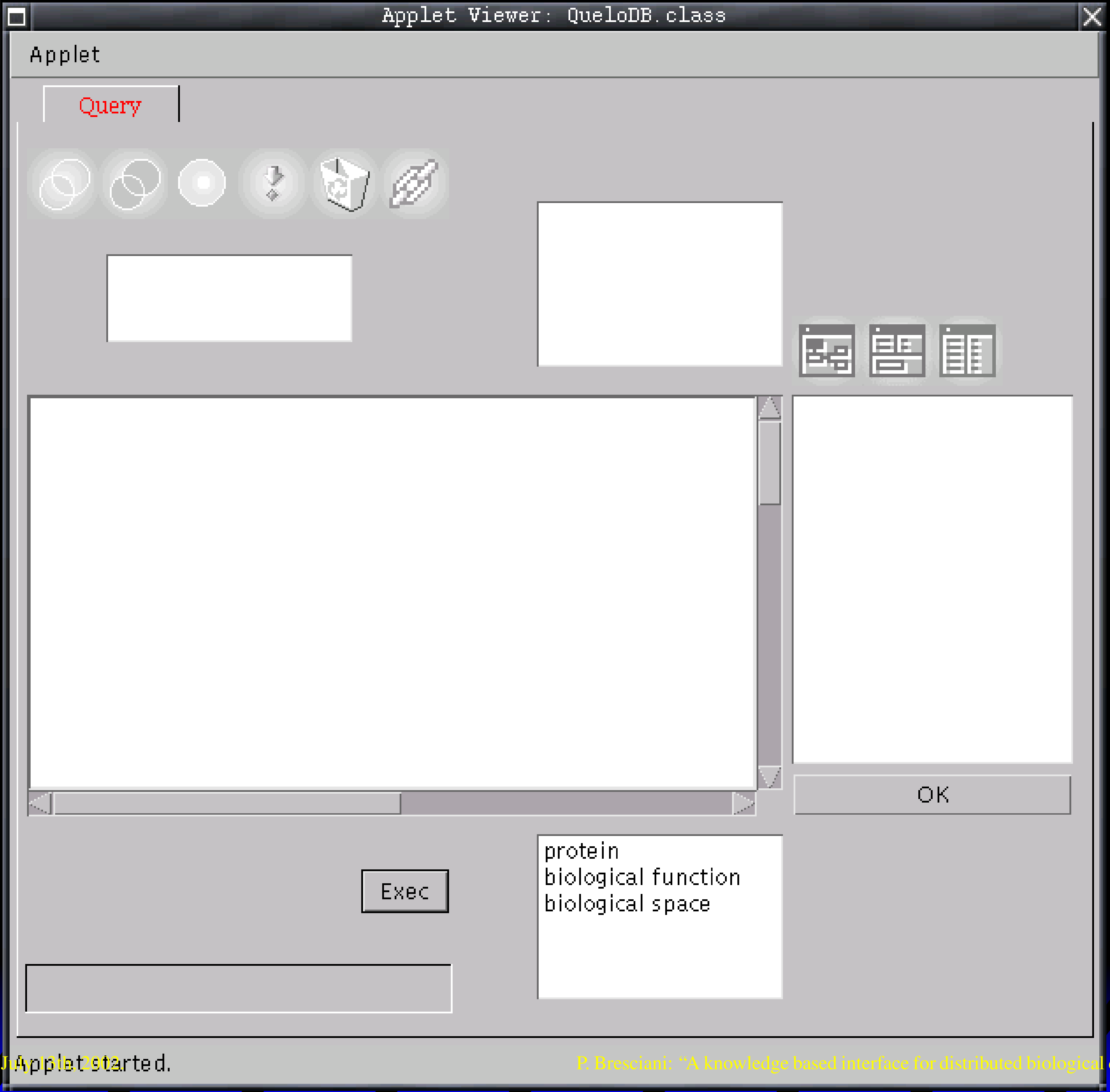
- Visual Interface for:
  - listing domain *concepts*
  - representing queries
  - transforming queries
- A Conceptual Model representation:
  an Ontology (better, a Knowledge Base)
  (**Tambis** + part of **GeneOntology**)
- A reasoner (the Description Logics (DL) Reasoner iFaCt [Horrocks – U.Manchester] + specific code).

Semantic checks

Some important features of the interface:

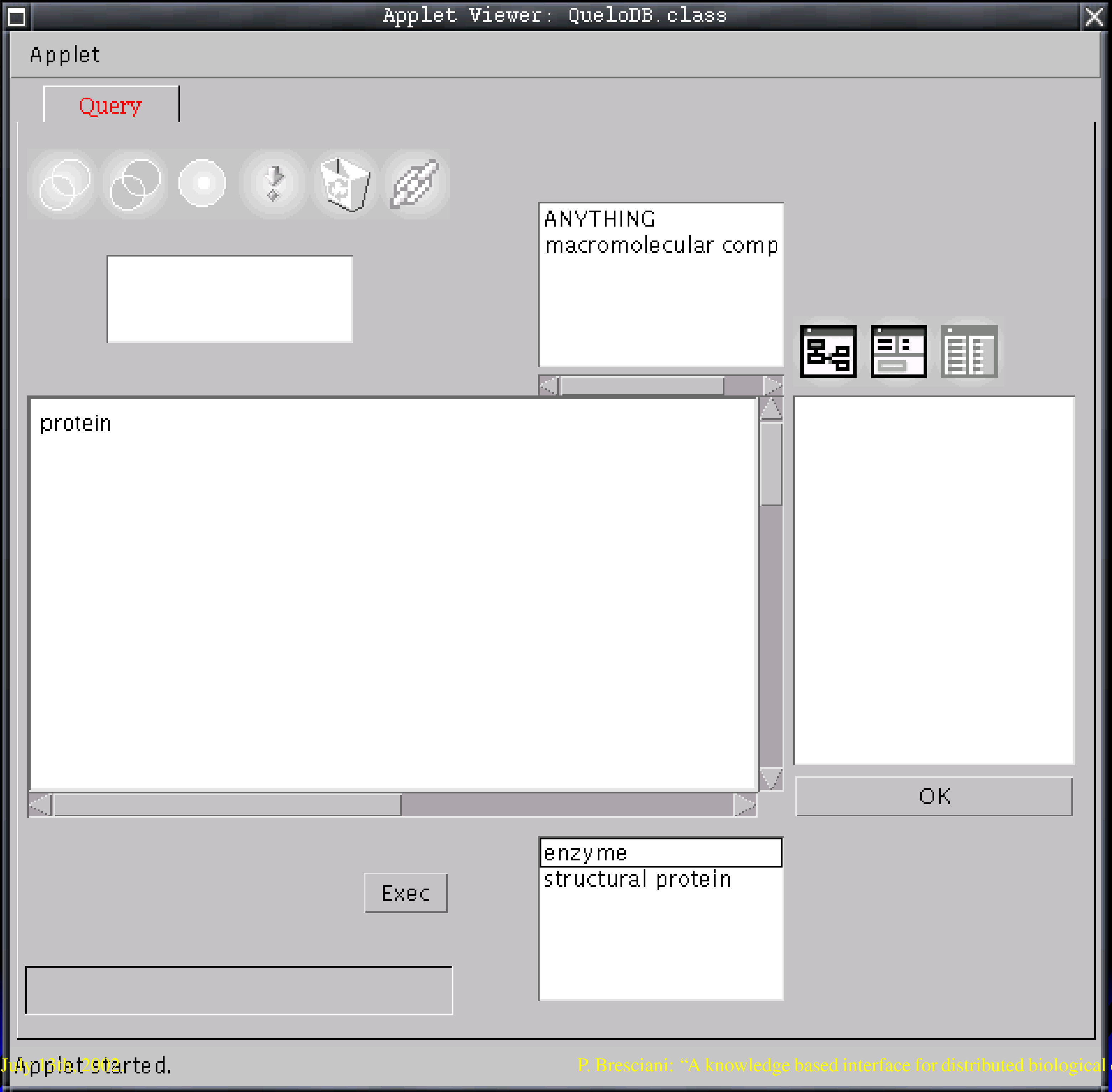
- Only *consistent* actions are allowed (actions that lead to consistent queries).
- Only *relevant* modifications are proposed (transformations that produce queries semantically not equivalent to the original).
- Only *close* modifications are proposed (not all the consistent and relevant modifications are proposed, but only those that lead to queries with semantics close to the original one).

An example of interaction



protein
biological function
biological space

Applet started.

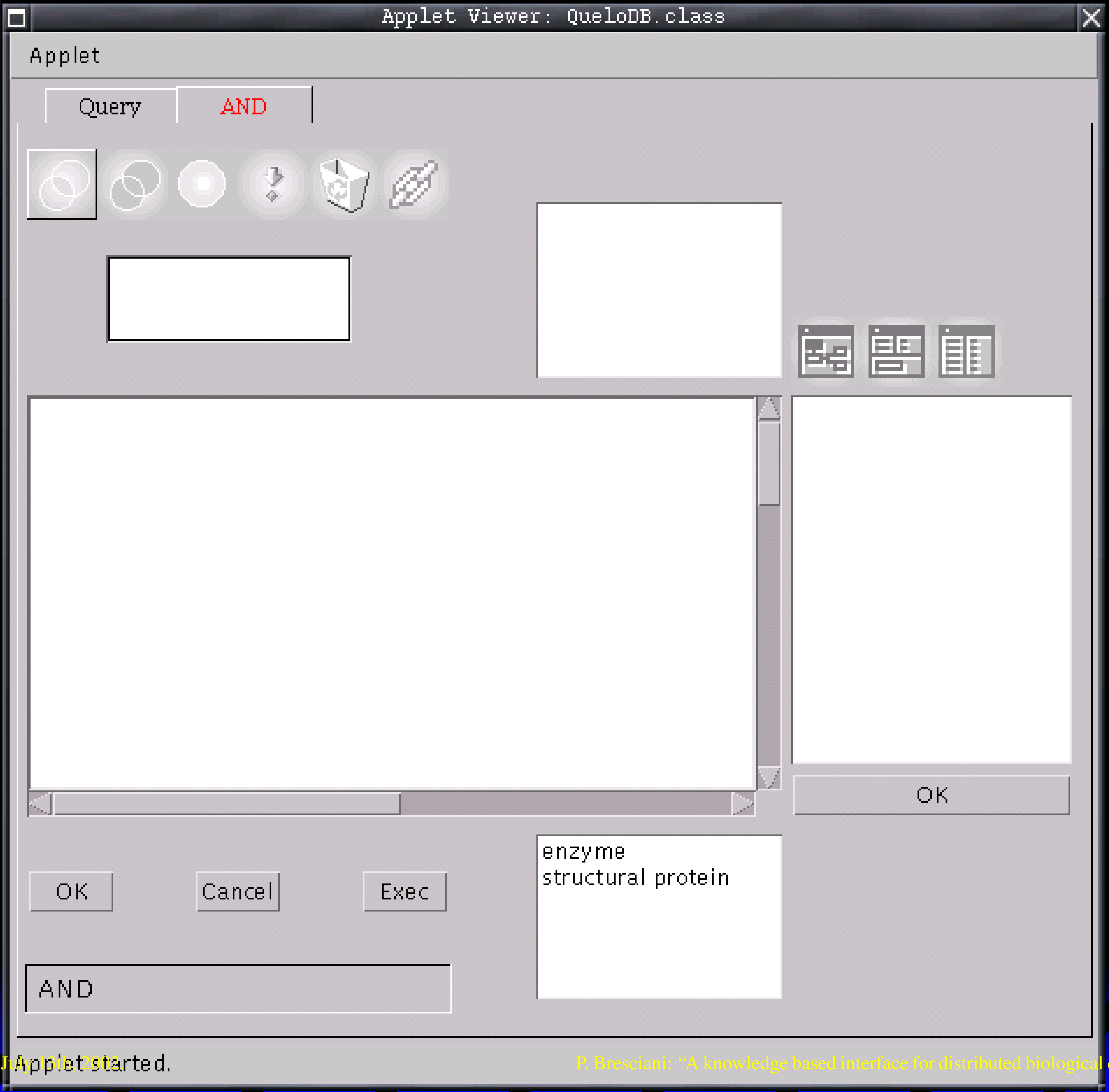


ANYTHING
macromolecular comp

protein

enzyme
structural protein

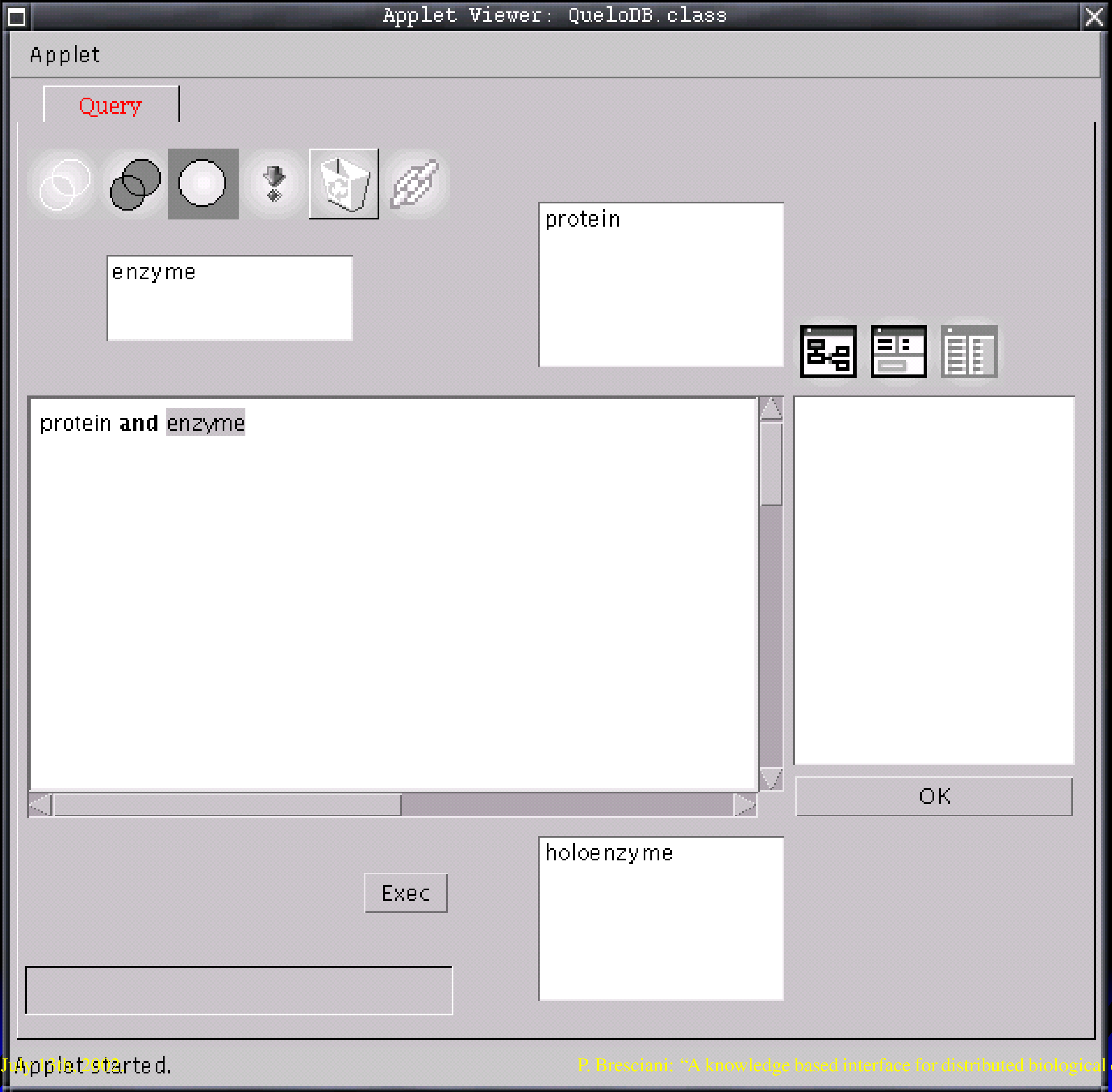
Applet started.



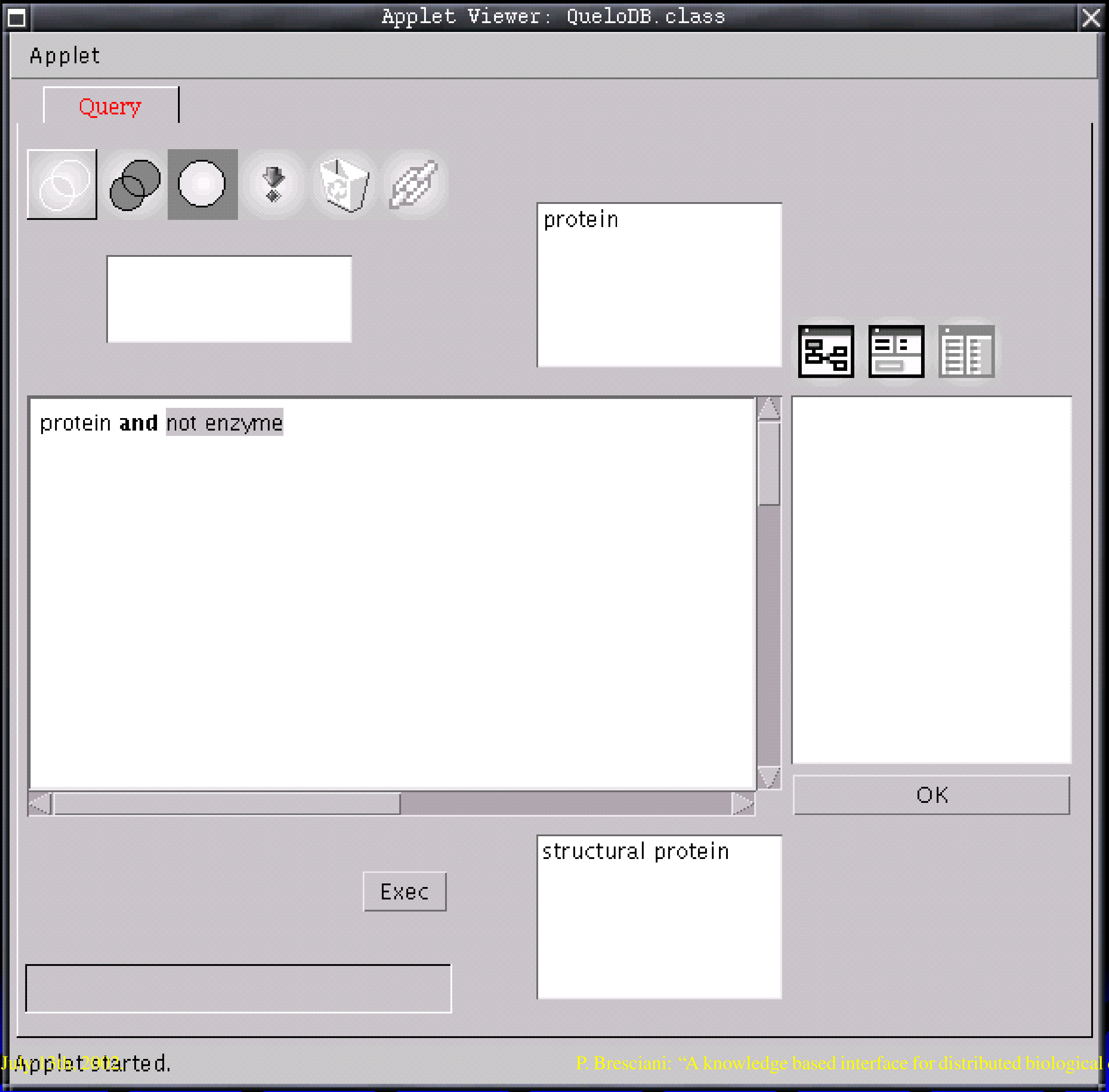
enzyme
structural protein

AND

Applet started.



protein **and** enzyme

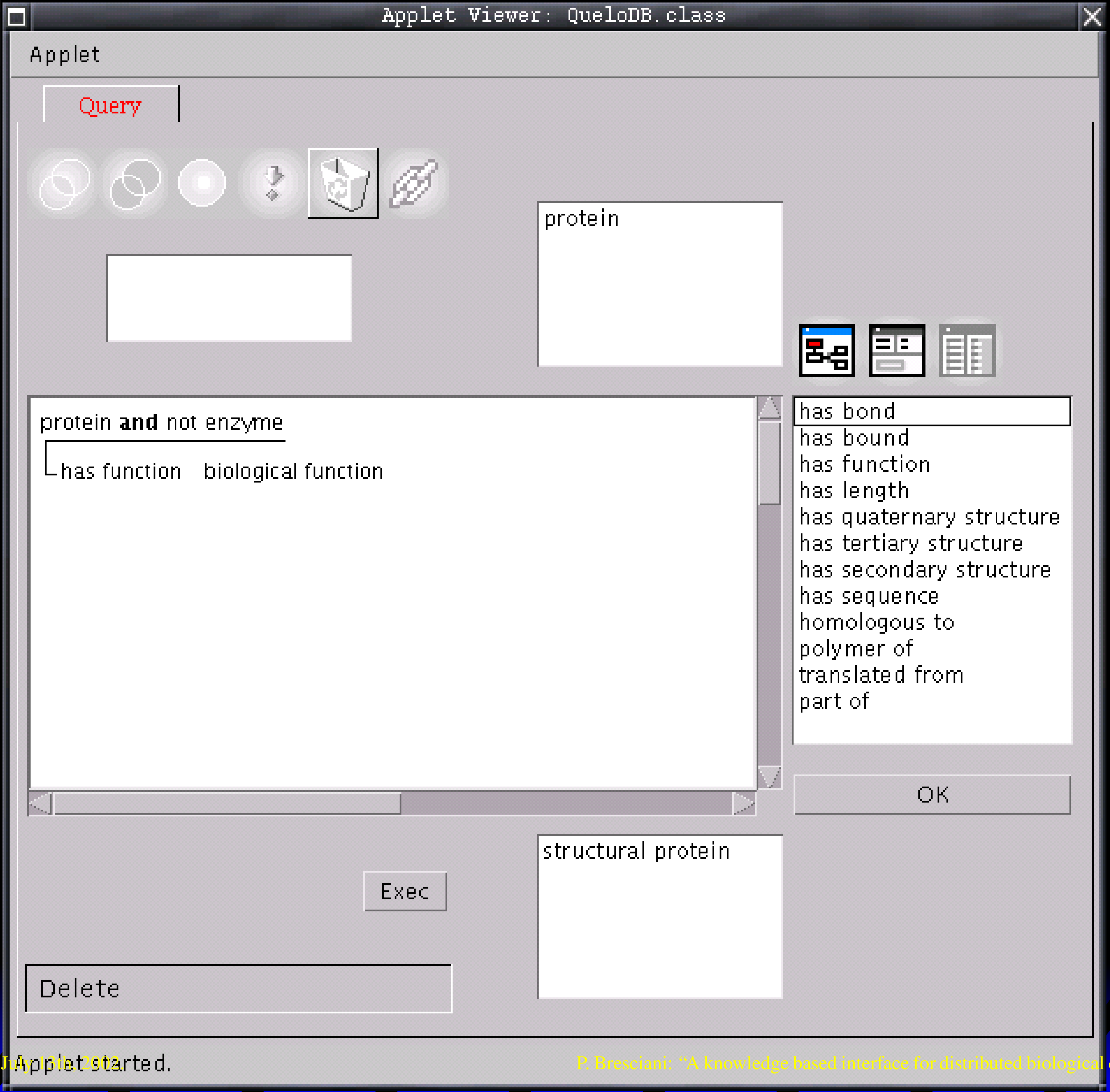


protein

protein **and** not enzyme

structural protein

Applet started.



protein

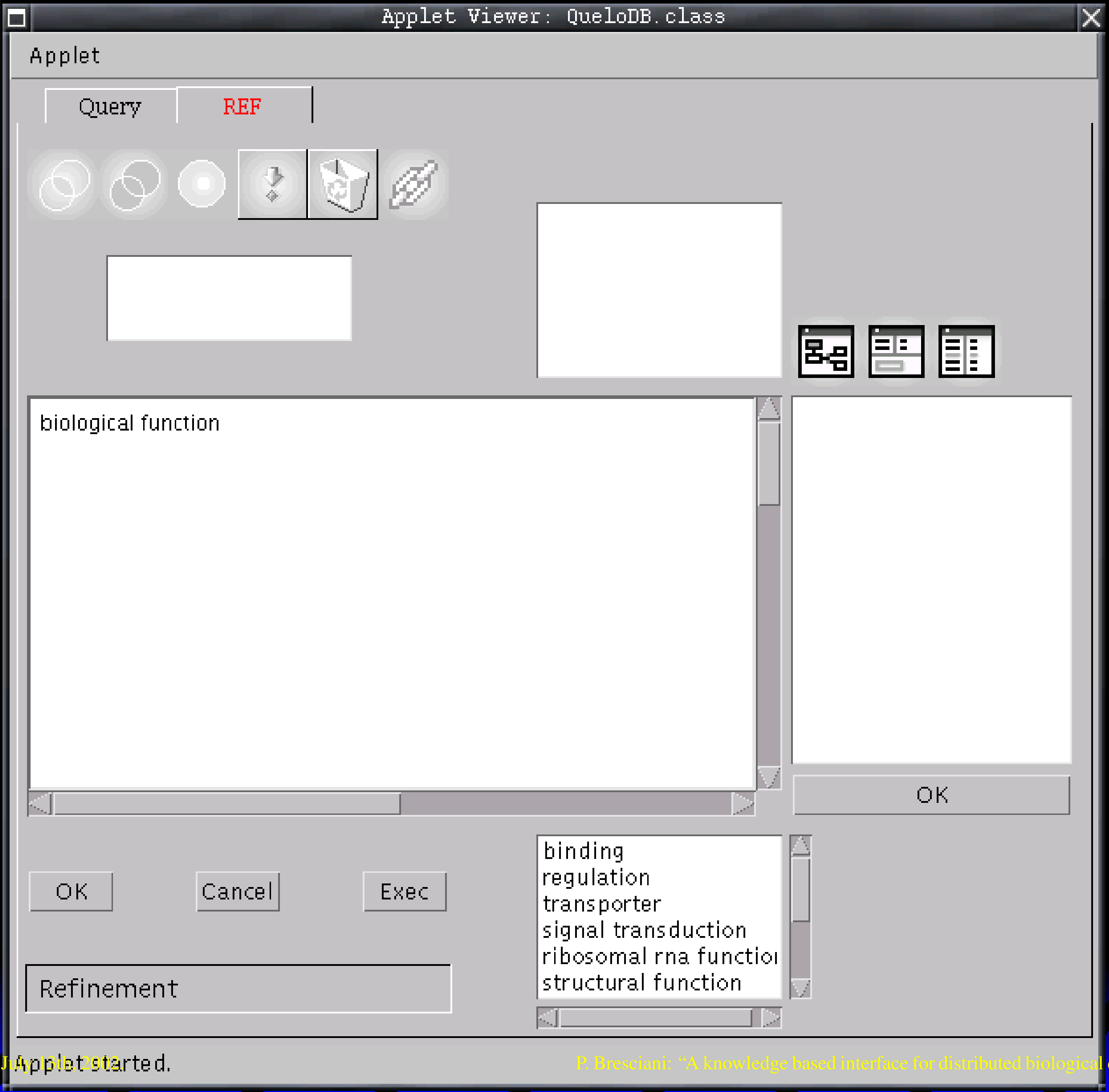
protein **and** not enzyme

has function biological function

has bond
has bound
has function
has length
has quaternary structure
has tertiary structure
has secondary structure
has sequence
homologous to
polymer of
translated from
part of

structural protein

Applet started.

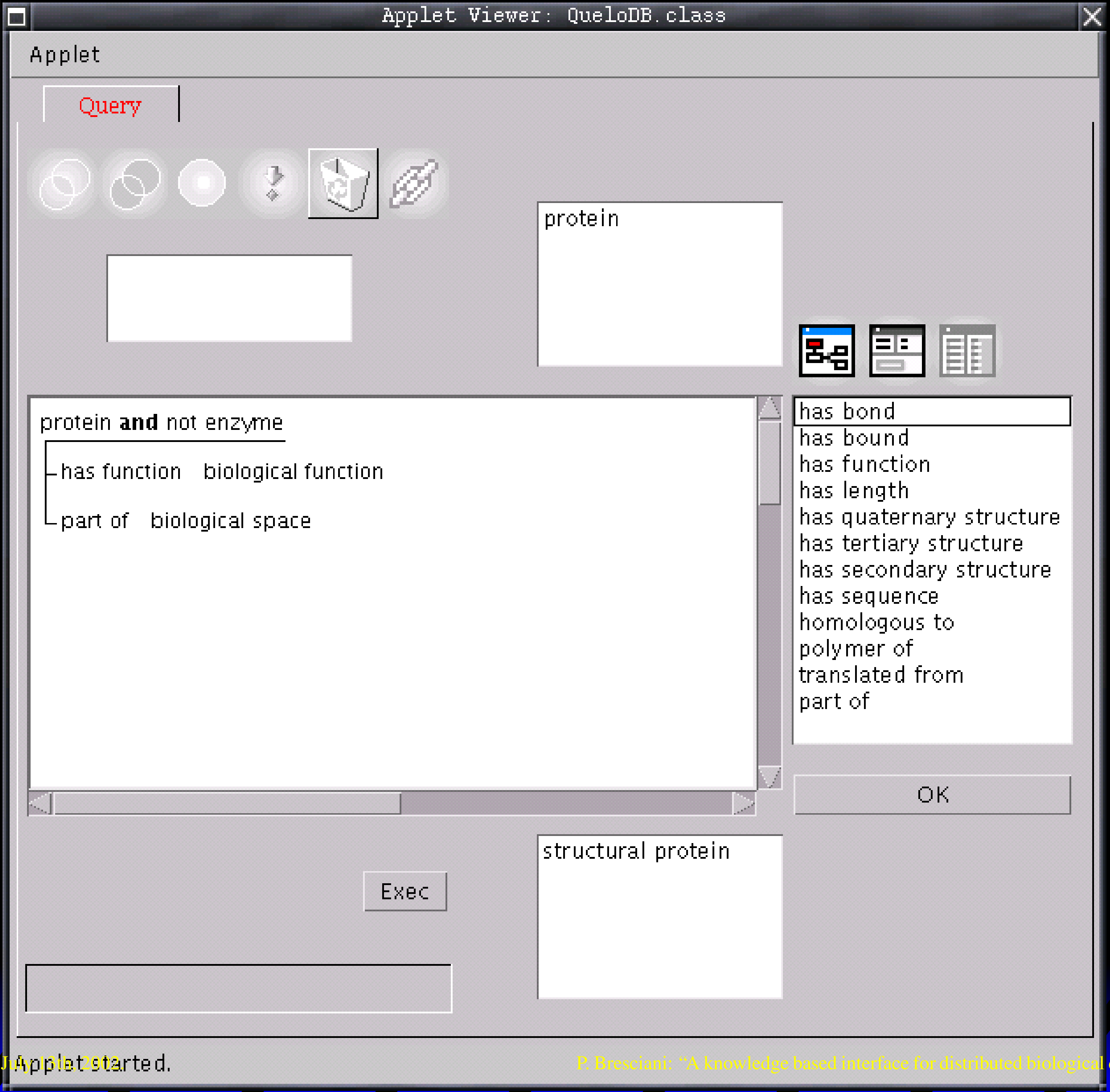


biological function

binding
regulation
transporter
signal transduction
ribosomal rna functioi
structural function

Refinement

Applet started.



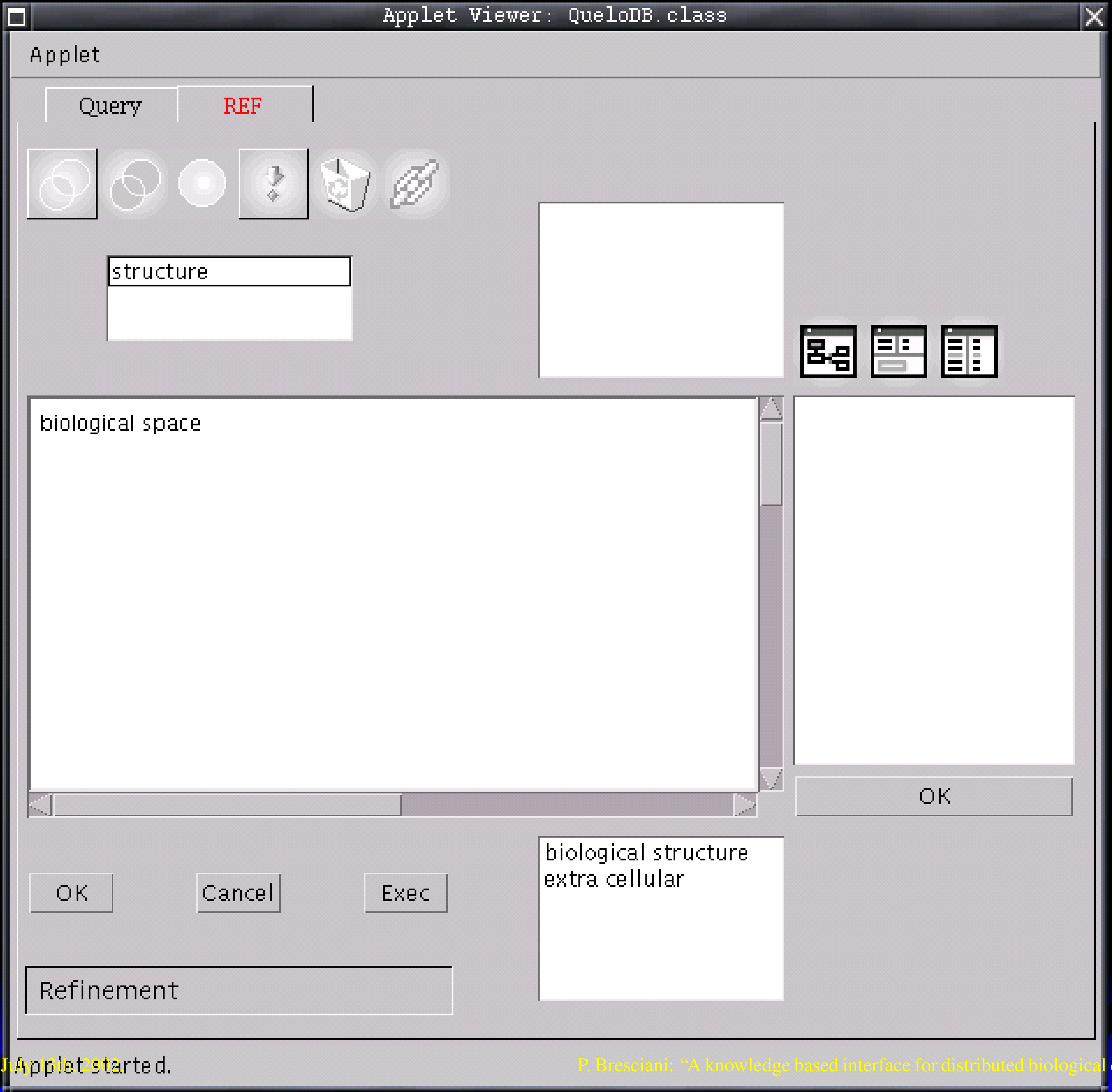
protein

protein **and** not enzyme
- has function biological function
- part of biological space

has bond
has bound
has function
has length
has quaternary structure
has tertiary structure
has secondary structure
has sequence
homologous to
polymer of
translated from
part of

structural protein

Applet started.



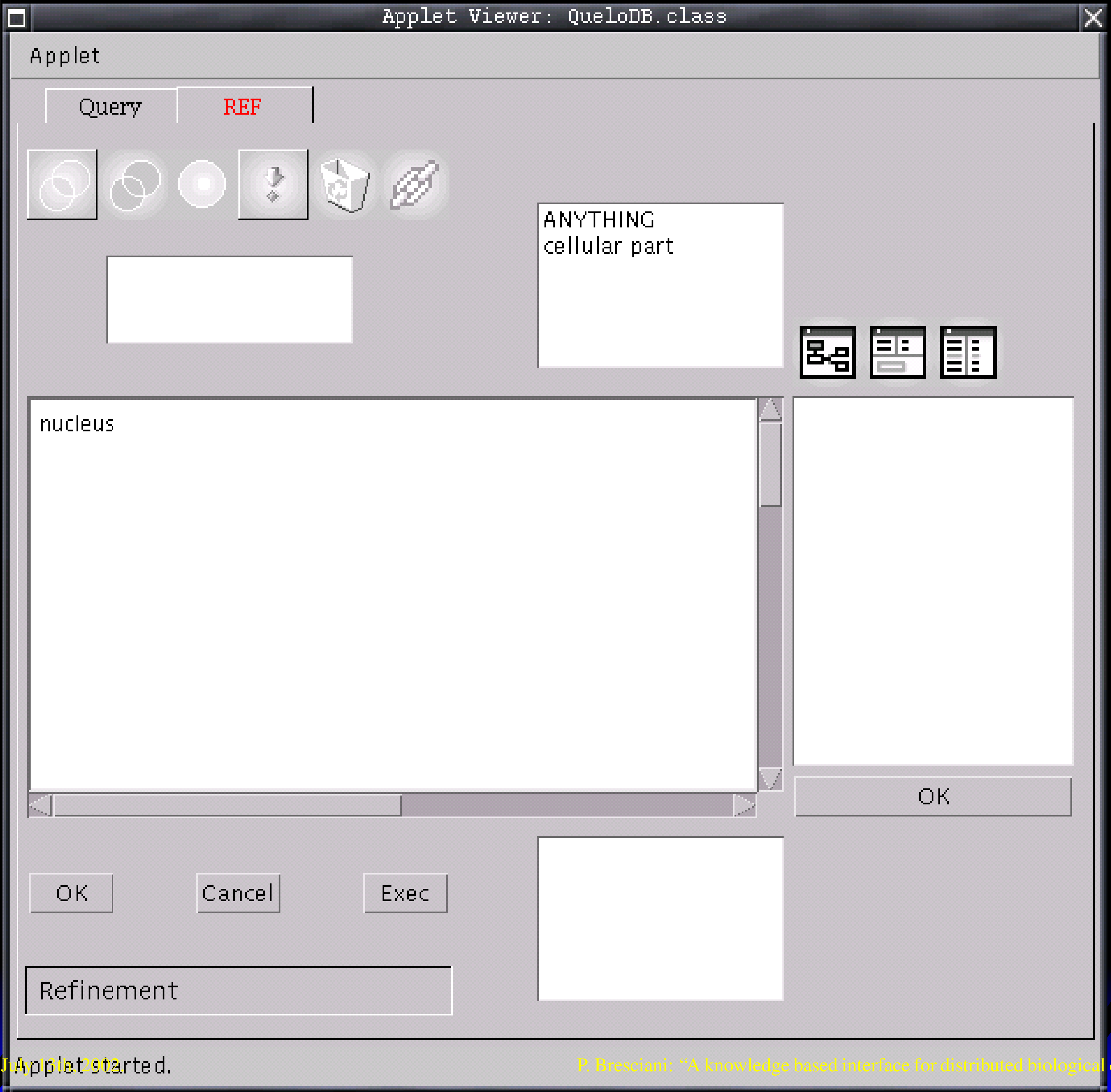
structure

biological space

biological structure
extra cellular

Refinement

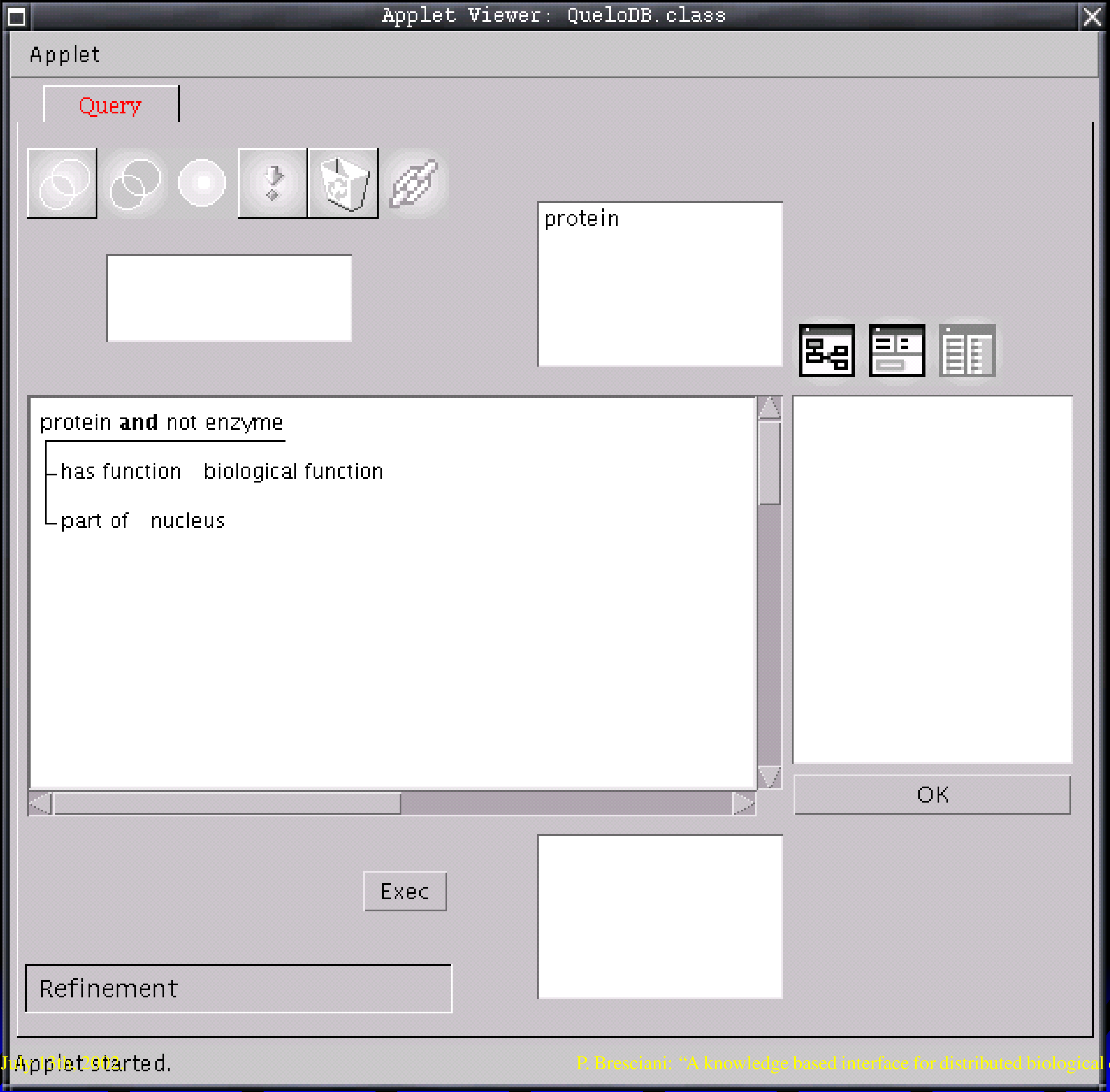
Applet started.



ANYTHING
cellular part

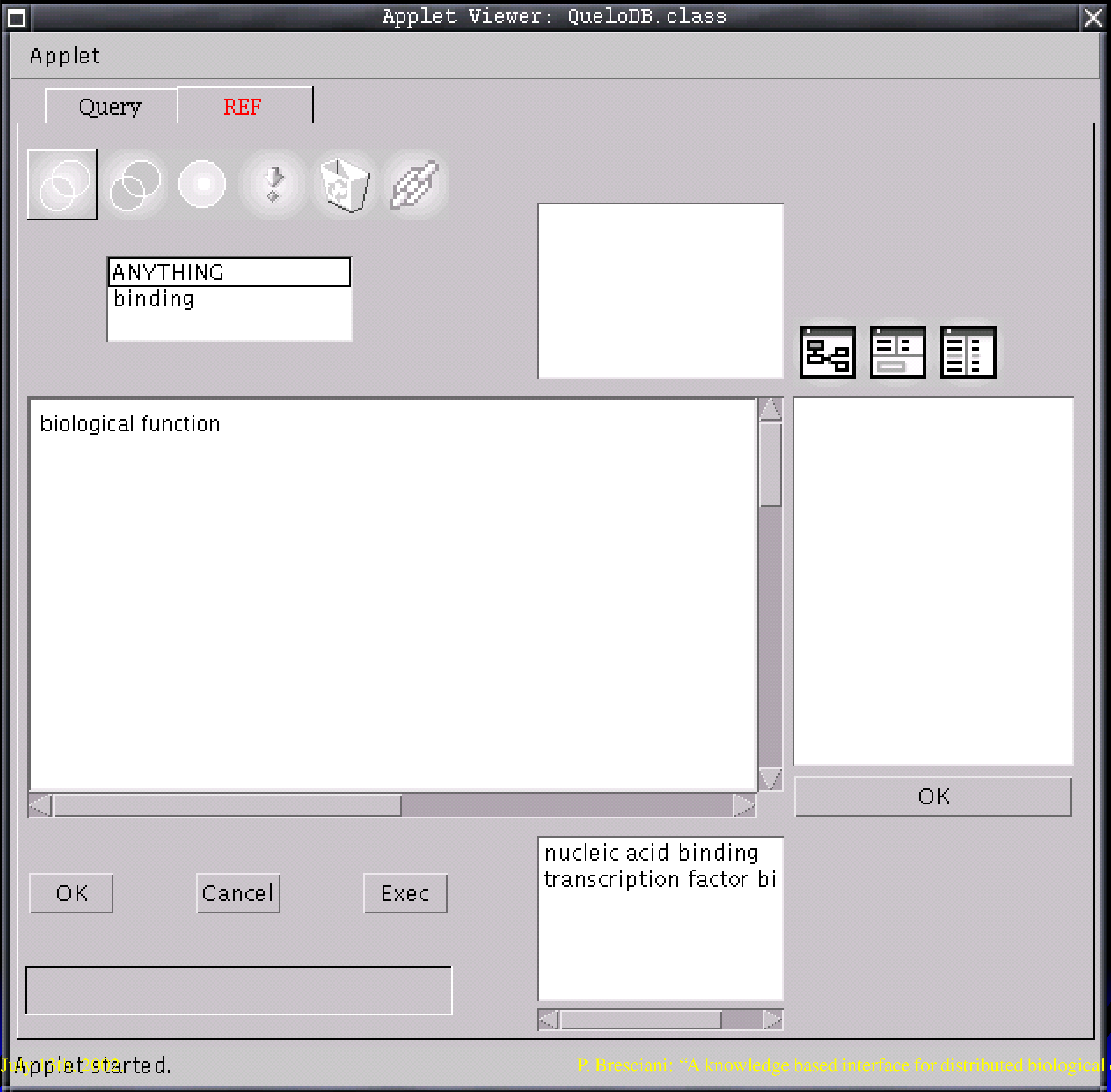
nucleus

Applet started.



protein **and** not enzyme
- has function biological function
- part of nucleus

Applet started.

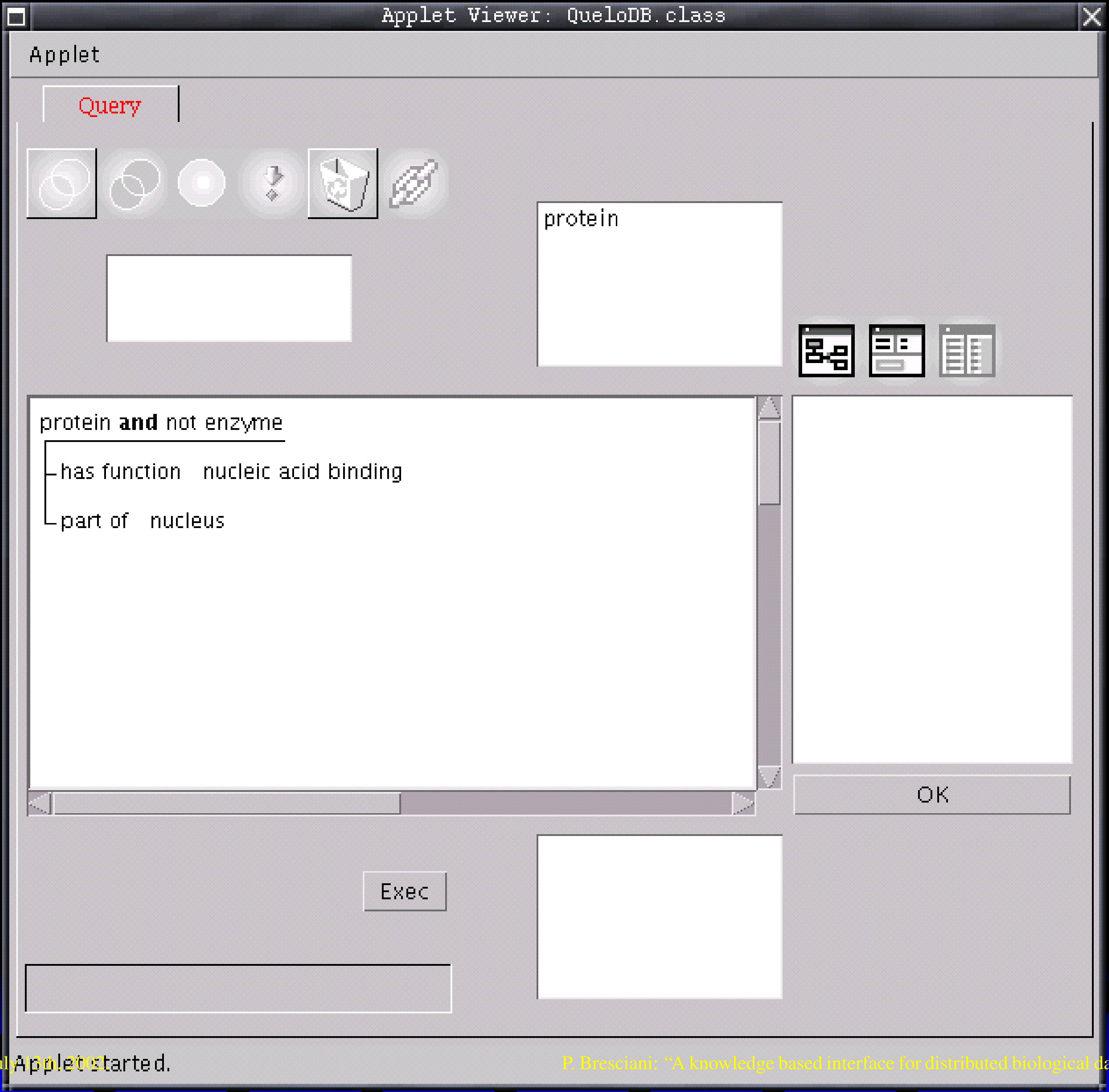


ANYTHING
binding

biological function

nucleic acid binding
transcription factor bi

Applet started.

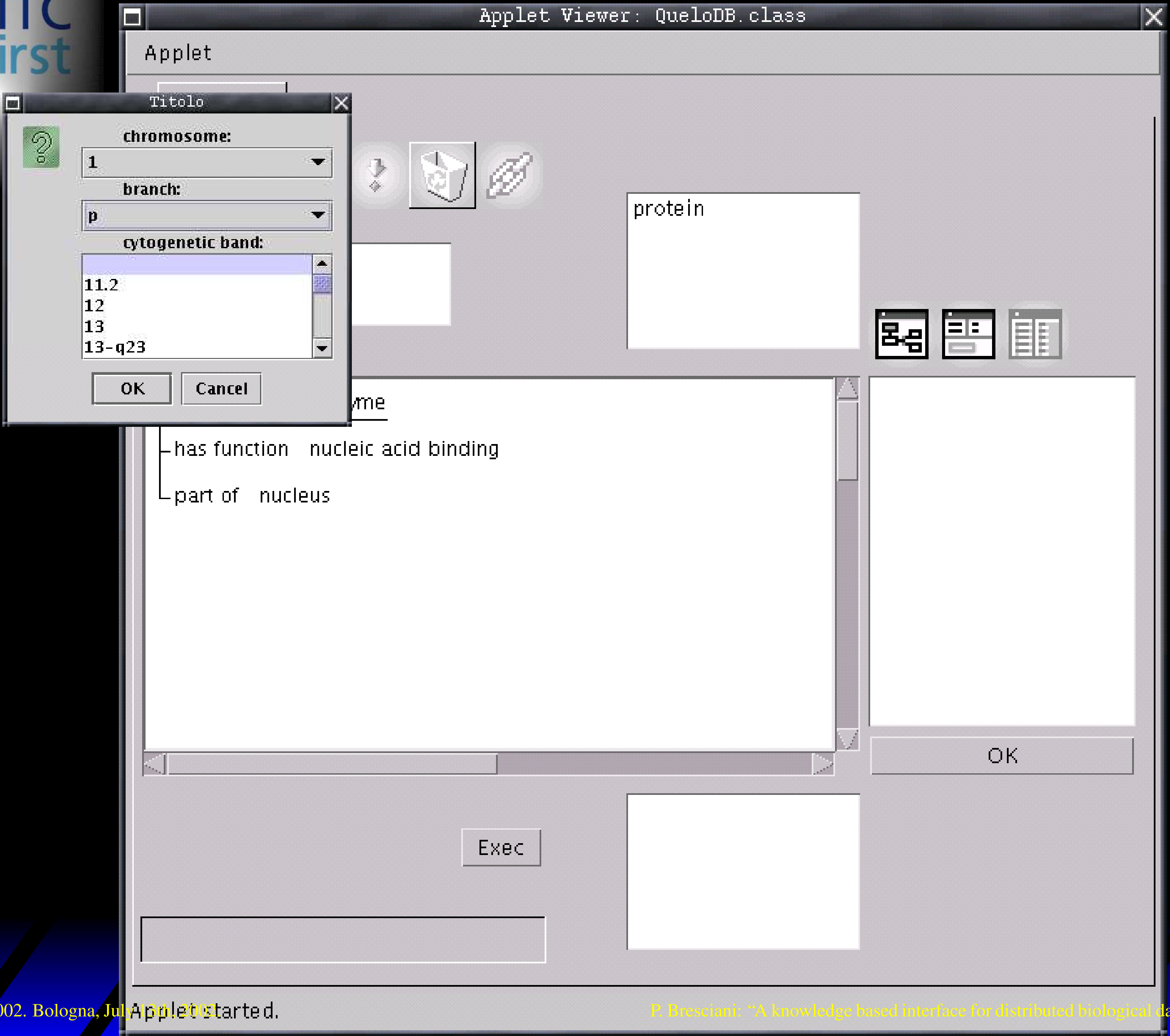


protein

protein **and** not enzyme

- has function nucleic acid binding
- part of nucleus

Applet started.



chromosome: 1

branch: p

cytogenetic band:

11.2
12
13
13-q23

protein

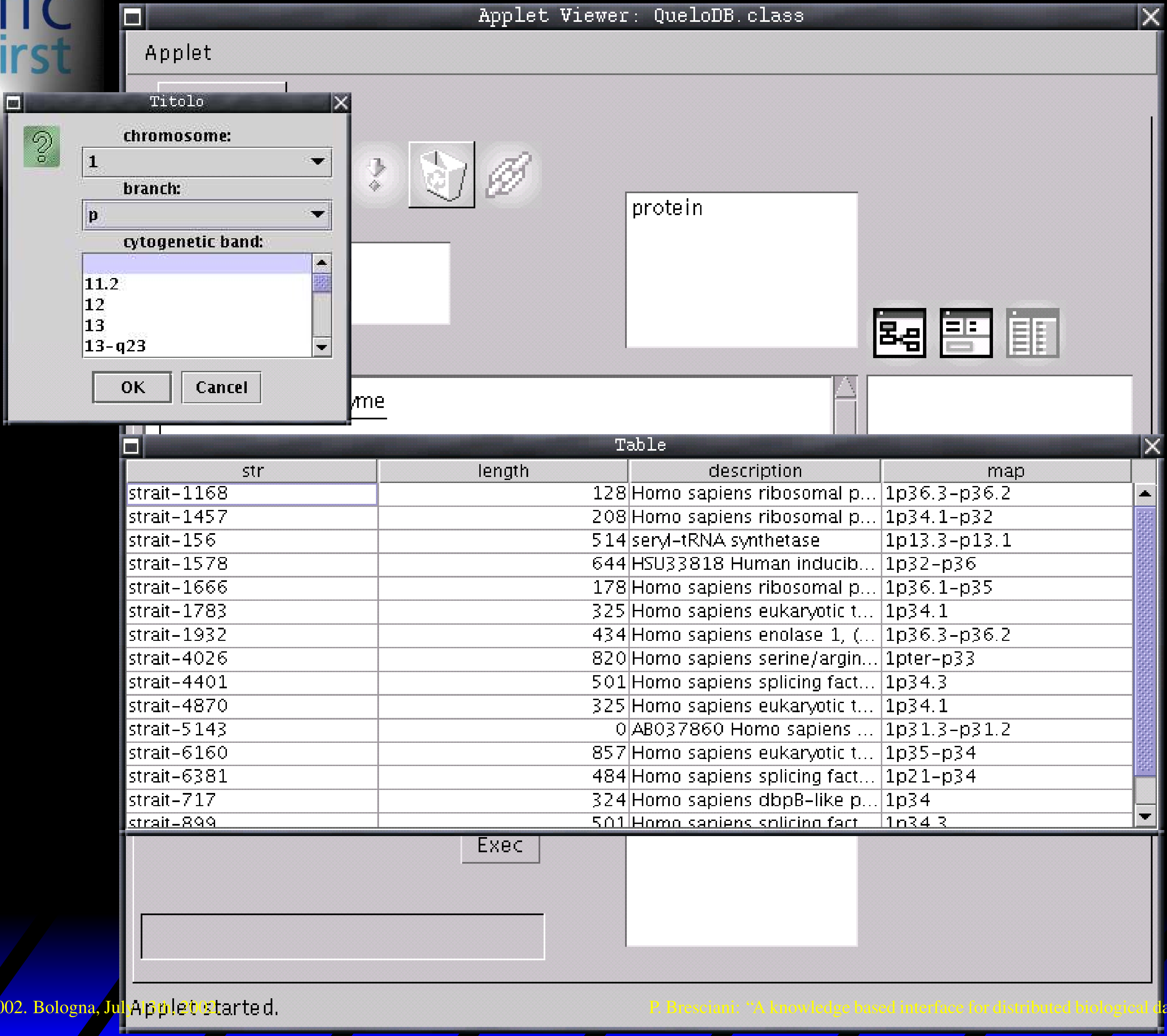
...yme

has function nucleic acid binding

part of nucleus

Applet started.

ITC irst



chromosome: 1

branch: p

cytogenetic band:

11.2
12
13
13-q23

protein

yme

str	length	description	map
strait-1168	128	Homo sapiens ribosomal p...	1p36.3-p36.2
strait-1457	208	Homo sapiens ribosomal p...	1p34.1-p32
strait-156	514	seryl-tRNA synthetase	1p13.3-p13.1
strait-1578	644	HSU33818 Human inducib...	1p32-p36
strait-1666	178	Homo sapiens ribosomal p...	1p36.1-p35
strait-1783	325	Homo sapiens eukaryotic t...	1p34.1
strait-1932	434	Homo sapiens enolase 1, (...	1p36.3-p36.2
strait-4026	820	Homo sapiens serine/argin...	1pter-p33
strait-4401	501	Homo sapiens splicing fact...	1p34.3
strait-4870	325	Homo sapiens eukaryotic t...	1p34.1
strait-5143	0	AB037860 Homo sapiens ...	1p31.3-p31.2
strait-6160	857	Homo sapiens eukaryotic t...	1p35-p34
strait-6381	484	Homo sapiens splicing fact...	1p21-p34
strait-717	324	Homo sapiens dbpB-like p...	1p34
strait-899	501	Homo sapiens splicing fact...	1p34.3

Applet started.

KRR services

Knowledge Based Approach means . . .

Knowledge Based Approach means . . .
Description Logics

DL are . . .

$\mathcal{L} = \{\mathtt{A}, C \sqcap D, \neg C, R, \exists R.C, \ldots, C \sqcup D, \forall R.C, R^{-1}, R^{\geq n}.C, R^{\leq n}.C, \ldots\}$

$I = \langle \Delta^I, \cdot^I \rangle$

semantically sound (and complete) automatic reasoning services as:

- *consistency-checking*
- *subsumption*
- *classification*

KB = $\{C \sqsubseteq D \ldots\}$

DL: FOL based semantics

$$\mathtt{A} \longrightarrow F_{\mathtt{A}}(\gamma) = A(\gamma)$$

$$\mathtt{P} \longrightarrow F_{\mathtt{P}}(\alpha, \beta) = P(\alpha, \beta)$$

Infix	Prefix	Semantics
$\neg C$	$(\mathtt{NOT}\ C)$	$\neg F_C(\gamma)$
$C \sqcap D$	$(\mathtt{AND}\ C\ D)$	$F_C(\gamma) \wedge F_D(\gamma)$
$C \sqcup D$	$(\mathtt{OR}\ C\ D)$	$F_C(\gamma) \vee F_D(\gamma)$
$\forall R.C$	$(\mathtt{ALL}\ R\ C)$	$\forall x.F_R(\gamma, x) \Rightarrow F_C(x)$
$\exists R.C$	$(\mathtt{SOME}\ R\ C)$	$\exists x.F_R(\gamma, x) \wedge F_C(x)$
$\vdots$	$\vdots$	$\vdots$
R^{-1}	$(\mathtt{INVERSE}\ R)$	$F_R(\beta, \alpha)$
$R \circ Q$	$(\mathtt{COMPOSE}\ R\ Q)$	$\exists x.F_R(\alpha, x) \wedge F_Q(x, \beta)$

The KB

```
;;;;;;;;;;;;;; DISJOINT COVERING ;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
(FACT::disjoint NUCLEUS CELL ORGANELLE CYTOSOL RIBOSOME SPLICEOSOME
                MEMBRANE ENDOPLASMIC-RETICULUM GOLGI-COMPLEX)
(FACT::implies CELLULAR-PART (:OR NUCLEUS ORGANELLE CYTOSOL RIBOSOME
               SPLICEOSOME MEMBRANE ENDOPLASMIC-RETICULUM GOLGI-COMPLEX))
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

;;;;;;;;;;;;;; RELATIONSHIPS RESTRICTIONS ;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
(implies
   (:AND protein (:SOME part-of nucleus))
   (:ALL has-function (:OR nucleic-acid-binding transcription-factor-binding)))
(implies
   (:AND protein (:SOME has-function (:OR nucleic-acid-binding transcription-factor-bi
   (:ALL part-of nucleus))
(implies (:AND protein (:SOME has-function (:OR cell-adhesion signal-transduction)))
(implies (:AND protein (:SOME part-of cell-membrane)) (:ALL has-function (:OR cell-adh
;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;

;;;;;;;;;;;;;; RELATION DOMAIN AND RANGE ;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
(implies enzyme (:SOME has-function enzyme-function))
(implies enzyme (:ALL has-function enzyme-function))
```

Reasoning with DL: Refinement

Refine `Biological-Space`

$Q(x) \leftarrow \texttt{Protein}(x) \wedge \neg\texttt{Enzyme}(x) \wedge \texttt{has-function}(x,w) \wedge \texttt{Nucleic-Acid-Binding}(w) \wedge$

$\texttt{is-in}(x,y) \wedge \texttt{Biological-Space}(y) \wedge \texttt{is-in}(y,z) \wedge \texttt{Cell}(z)$

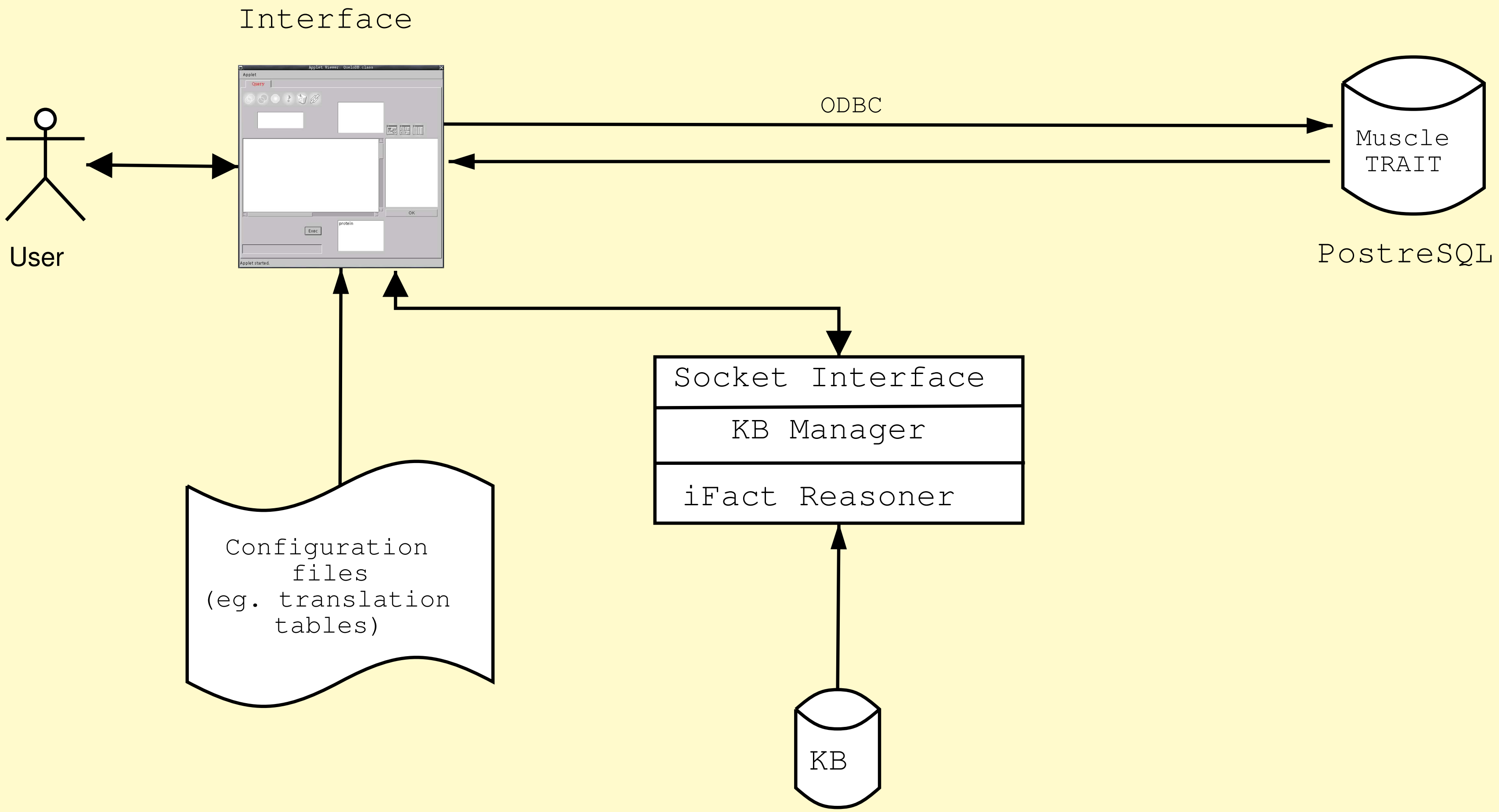
```
(AND Protein (NOT Ezime)
      (SOME has-function Nucleic-Acid-Binding)
      (SOME is-in (AND Biological-Space (SOME is-in Cell)
```

Ref=

$MSS(\{X \mid Q' =$
$(\textsc{Protein} \sqcap \ldots \sqcap \exists\ \text{is-in}.(X \sqcap \exists\ \text{is-in.Cell})) \wedge Q \sqsubset Q' \wedge \nexists Q'' =$
$(\textsc{Protein} \sqcap \ldots \sqcap \exists\text{is-in}.(Y \sqcap \exists\text{is-in.Cell}))s.t.\ Y \sqsubset X \wedge Q \sqsubset$
$Q'' \sqsubset Q'\}),$

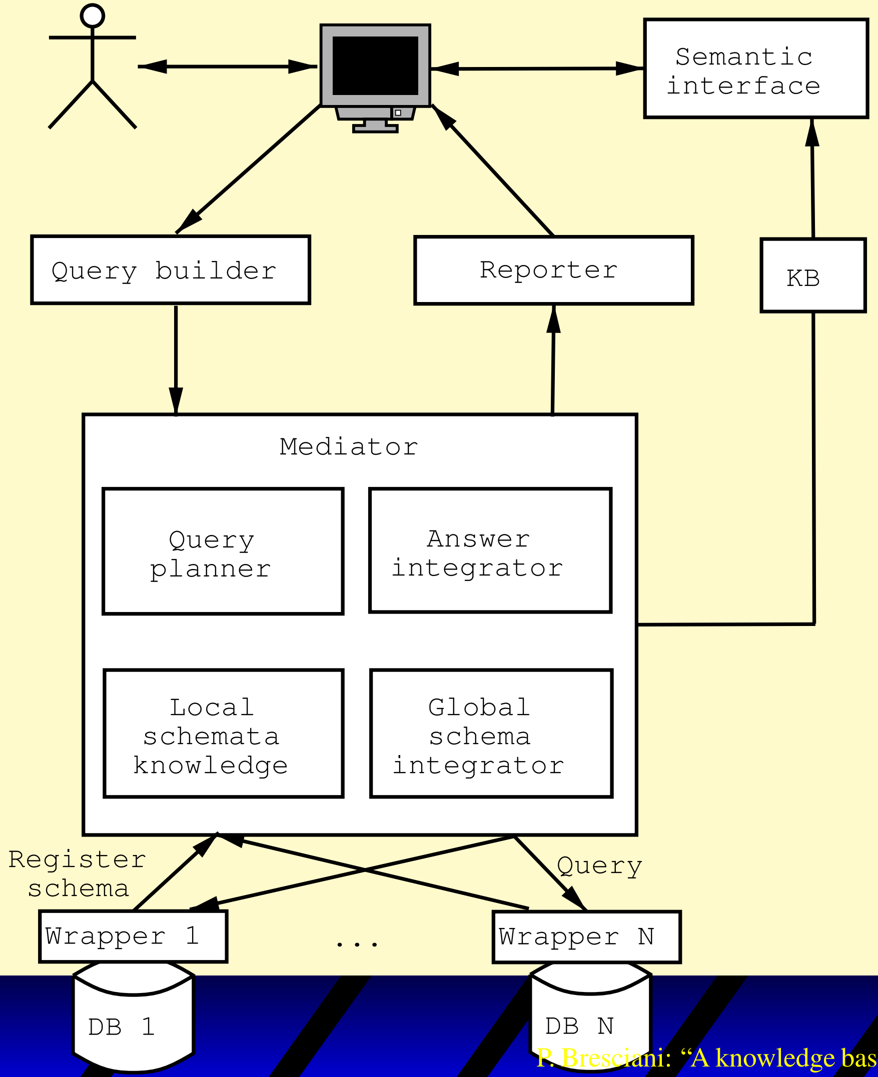
and:

Architecture



Towards accessing more DBs

Towards accessing more DBs



Promising opportunities from Agent Technologies

- accessing multiple DBs with one homogeneous interface
- architectural flexibility (easily extensible)
- robustness (in case of redundancy)
- caching and notification (multiple users, repeated “similar” queries)

Promising opportunities from Agent Technologies

- accessing multiple DBs with one homogeneous interface
- architectural flexibility (easily extensible)
- robustness (in case of redundancy)
- caching and notification (multiple users, repeated "similar" queries)
- standard protocol for schemata interchange and for communication (XML; DAML+OIL)
- wrappers design and maintenance
- query planning

Conclusions and Future Developments

A VQS prototypical system, applied to the access to biological databases, has been introduced.

The use of KR reasoning tools adds interesting semantic services, like:

- consistency checking of the queries;
- intelligent incremental browsing and exploration of the database semantics;
- effective features for relevant query refinement and generalization.

Conclusions and Future Developments

A VQS prototypical system, applied to the access to biological databases, has been introduced.
The use of KR reasoning tools adds interesting semantic services, like:

- consistency checking of the queries;
- intelligent incremental browsing and exploration of the database semantics;
- effective features for relevant query refinement and generalization.

Some preliminary ideas on an agent based architecture for accessing distributed DBs have been presented.

References

- M. M. Zloof. *Query-by-example: A database language*. IBM System Journal, 16(4):324-343, 1977
- T. Cartacci, S.K. Chang, M.F. Costabile, S. Levialdi and G. Santucci. *A graph-based framework for multiparadigmatic visual access to database*. IEEE Transactions on Knowledge and Data Engineering, 8(3):455-475, 1996.
- P. Bresciani, M. Nori and N. Pedot. *A Knowledge Based Paradigm for Querying Databases*. Proc. DEXA 2000, LNCS #1873. Springer.
- P. Bresciani, M. Nori and N. Pedot. *QueloDB: a Knowledge Based Visual Query System*. Proc. IC-AI 2000, vol. III, June 2000. CSREA Press.
- M. Ashburner at al. *Creating the gene ontology resources: design and implementation*. Genome Research, 11(8):1425-1433, Aug. 2001.
- P. G. Baker, C. A. Goble, S. Bechhofer, N. W. Paton, R. Stevens and A. Brass. *An ontology for bioinformatics applications*. Bionformatics, 15(6):510-520, 1999.