Demo script for ‘Knowtator: a Protégé plug-in for annotated corpus construction’

The demo for Knowtator will be geared towards people who are interested in building annotated corpora and will feature live software demonstrations of key features of the tool. I will emphasize Knowtator’s ability to leverage a conceptual model defined in Protégé for use as an annotation schema. As a case study, I will demonstrate how Knowtator has been employed for actual annotation projects at the University of Colorado Health Sciences Center (UCHSC) in Aurora. Knowtator is being used there to capture mentions of protein transport events. A typical transport event involves a single protein moving from one part of a cell (i.e. cellular component) to another with the help of another protein (i.e. a transporter). A typical demo will consist of annotation schema creation and text annotation. Knowtator is also being used for a text annotation project at the Mayo Clinic in Rochester, MN. Details of this project may optionally provide additional examples.

Annotation schema creation- The Protégé knowledge-base editor can be used to create new class, instance, slot, and facet frames for defining the annotation schema. Figure 1 shows the creation of a subclass of *cellular component* in progress using the Protégé class editor. There is a taxonomy of *cell component* classes that represent the common locations within a cell that a protein is transported to and from. Figure 2 displays the class definition for *transport* and shows its slots and the constraints on those slots (e.g. a *transport origin* must be a *cell component*). Figure 3 shows a class definition for *protein* which is a kind of macromolecule. The only slot of the class *protein* is a simple attribute that accepts an integer value corresponding to an accession number in a protein database called Entrez.

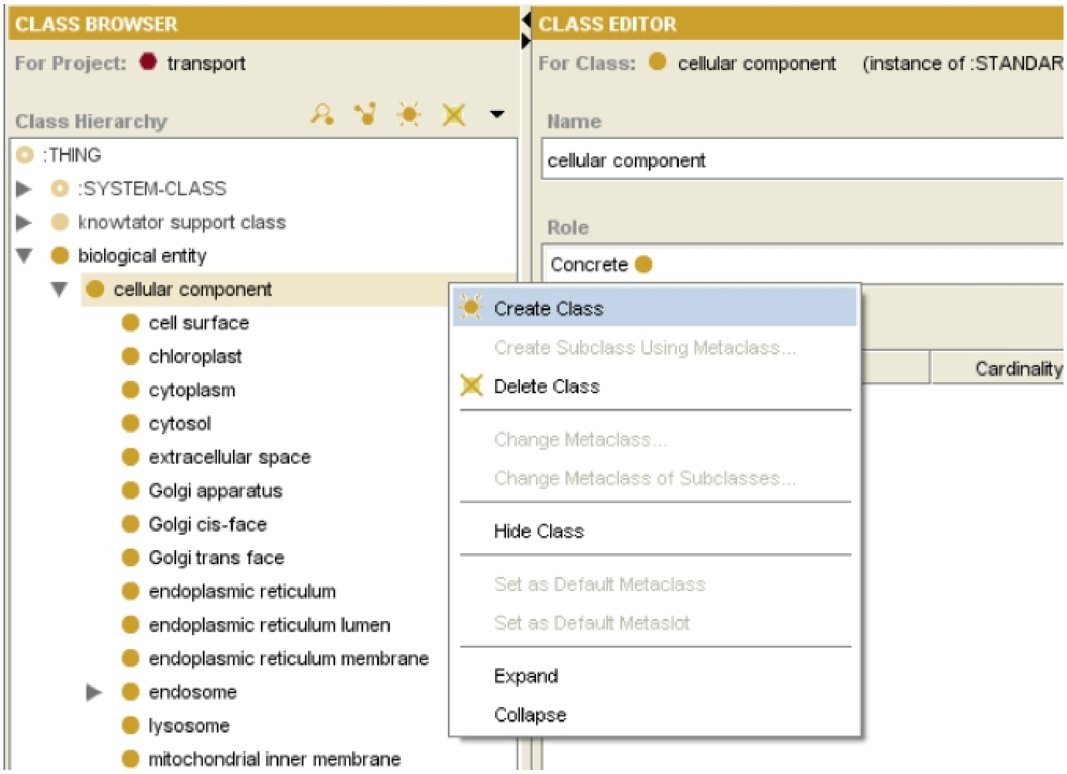

Figure 1

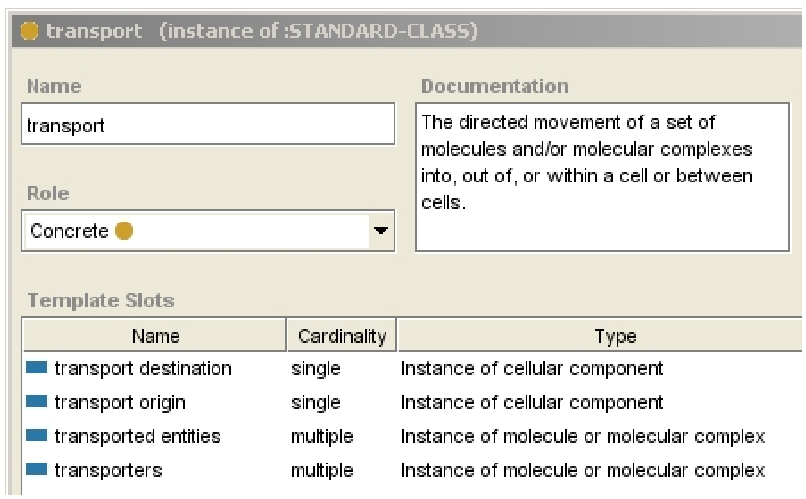

Figure 2

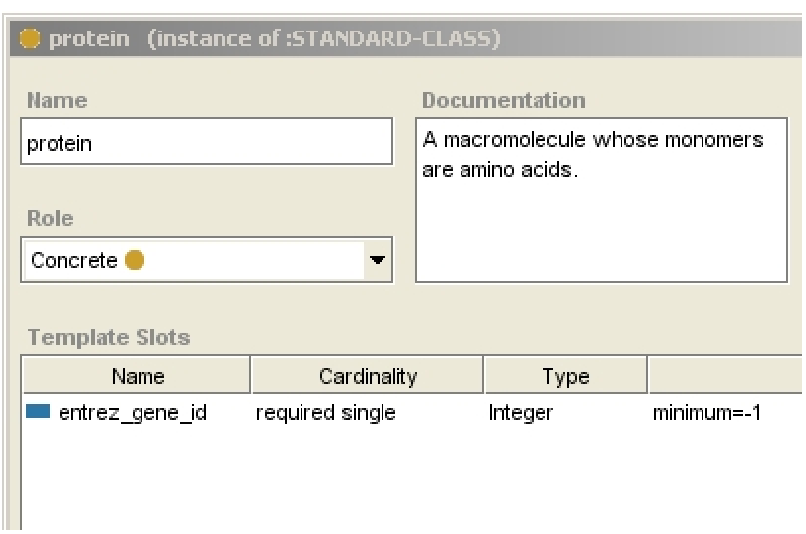

Figure 3

Annotation of text – Once an annotation schema has been created, then it can be immediately used for text annotation. Figure 4 shows a sentence that is going to be annotated. On the left is the subsumption hierarchy of the available annotation types. The text area towards the upper-right corner of the screen shows the sentence that will be annotated. This is a special kind of text called a GeneRIF that comes from the Entrez Gene database. GeneRIFs are short summaries that describe protein function that are attached to protein records in Entrez Gene. The example shown can be found at:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=full_report&list_uids=10945

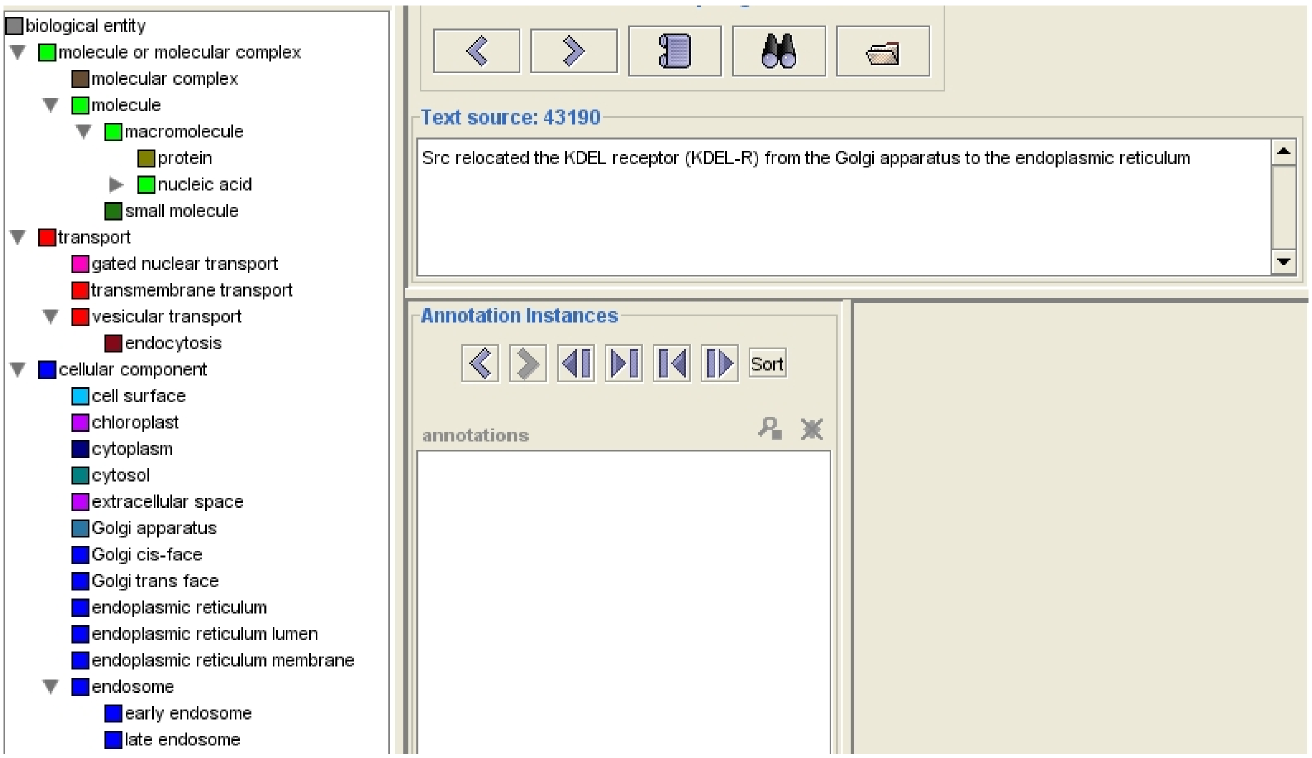

Figure 4

Figure 5 shows a single annotation corresponding to the *protein* class that has been created. The text *Src* was highlighted in the sentence, the class for *protein* was selected, and an annotation was created. The resulting annotation is listed in the section labled 'Annotation Instances.' An entrez gene id can now be filled in for this annotation as shown in Figure 6. Figure 7 shows a more complete set of annotations for the example sentence. There are now three protein annotations, two cellular components and a transport annotation.

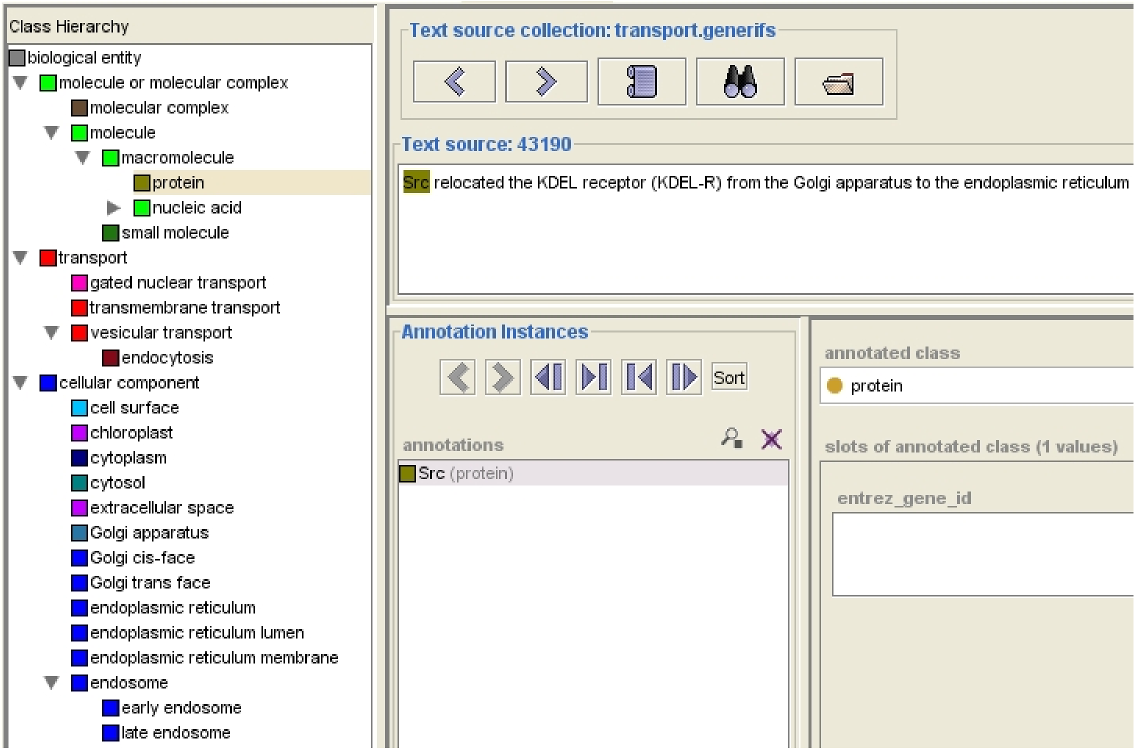

Figure 5

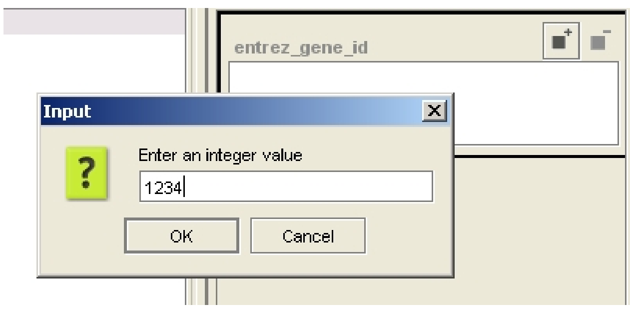

Figure 6

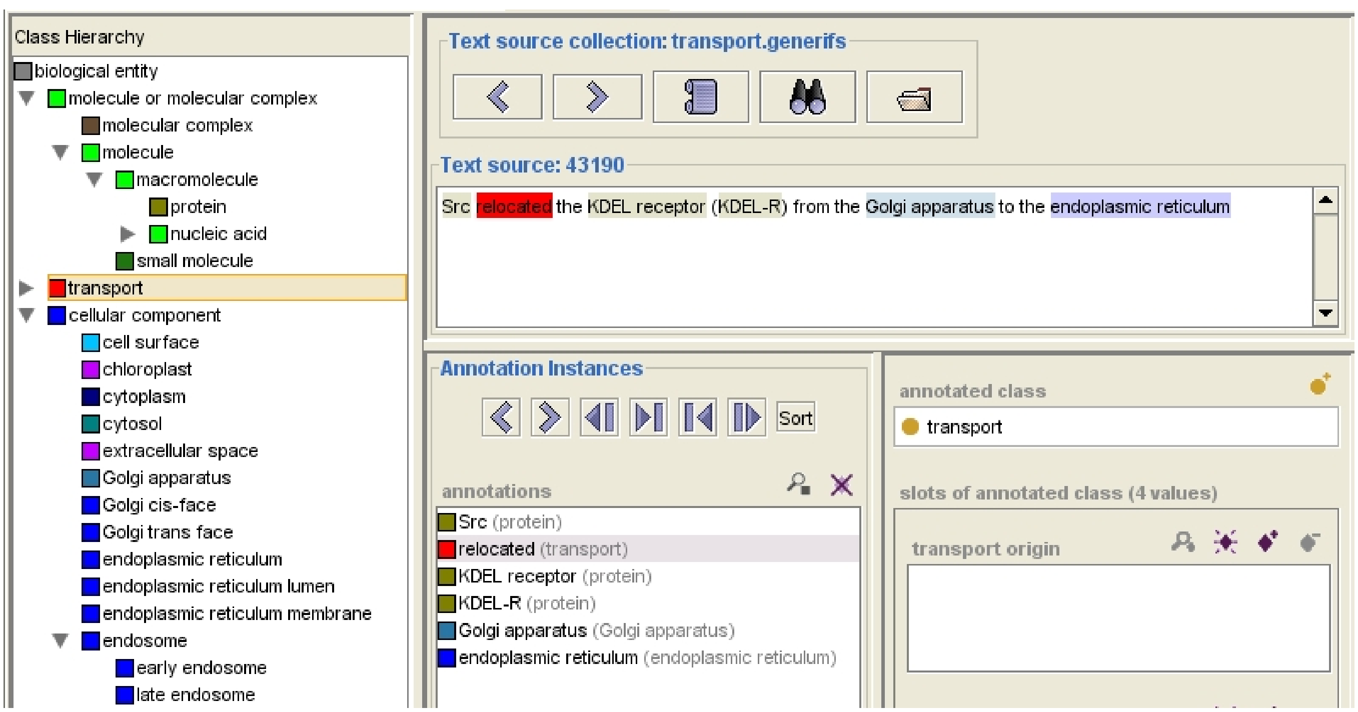

Figure 7

The transport annotation will relate the protein and cellular component annotations with each other via the slots of the transport class. Figure 8 shows the value of *transport origin* being filled in with an annotation of type *cellular component*. This demonstrates the usefulness of a subsumption hierarchy of types because only annotations corresponding to *cellular components* will be made available for selection. This constraint is a direct result of the slot definition of *transport origin* for the class *transport* which requires that the value must be a *cellular component* as shown in Figure 2.

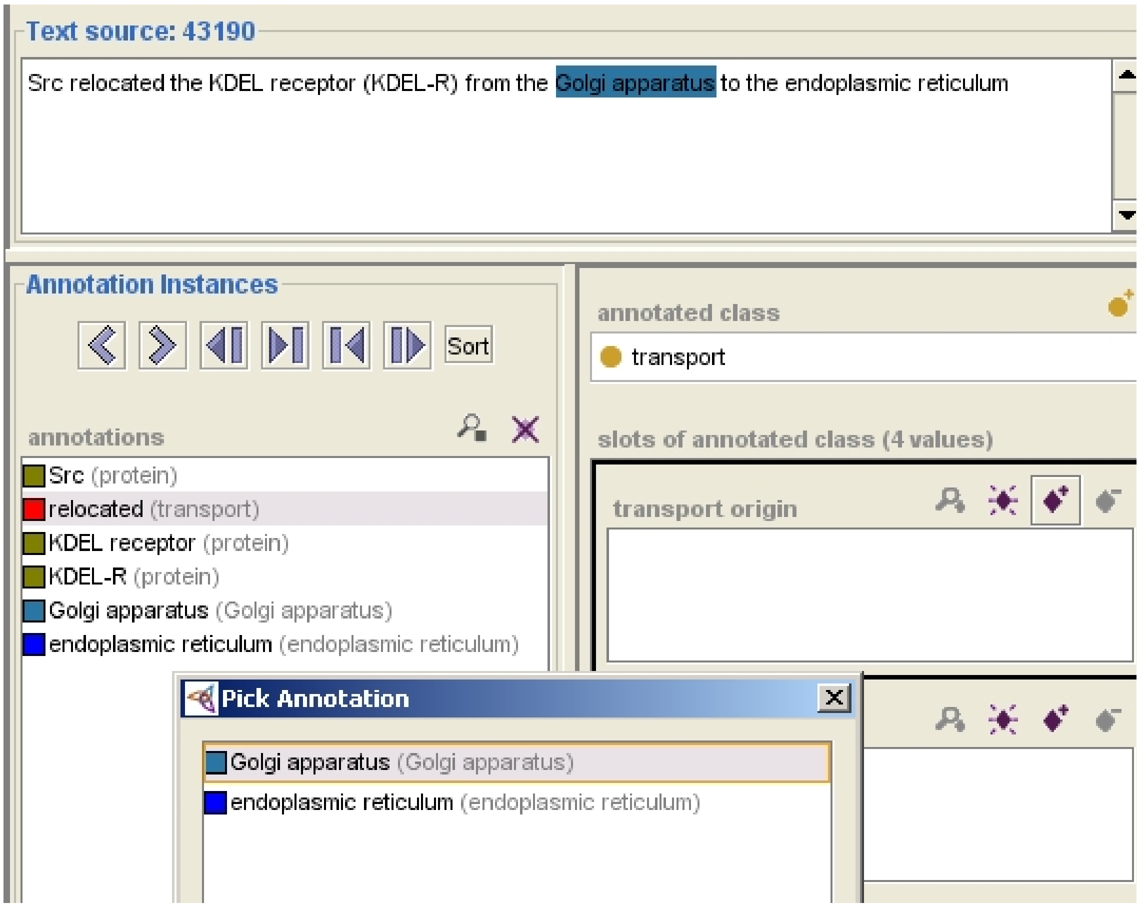

Figure 9

The button with the 'diamond+plus' icon was clicked to add a value for the *transport origin* slot of the *transport* annotation corresponding to the text 'relocated'. A dialog pops up that provides the option to choose one of the annotations related to cellular component. The complete transport annotation is shown in Figure 10.

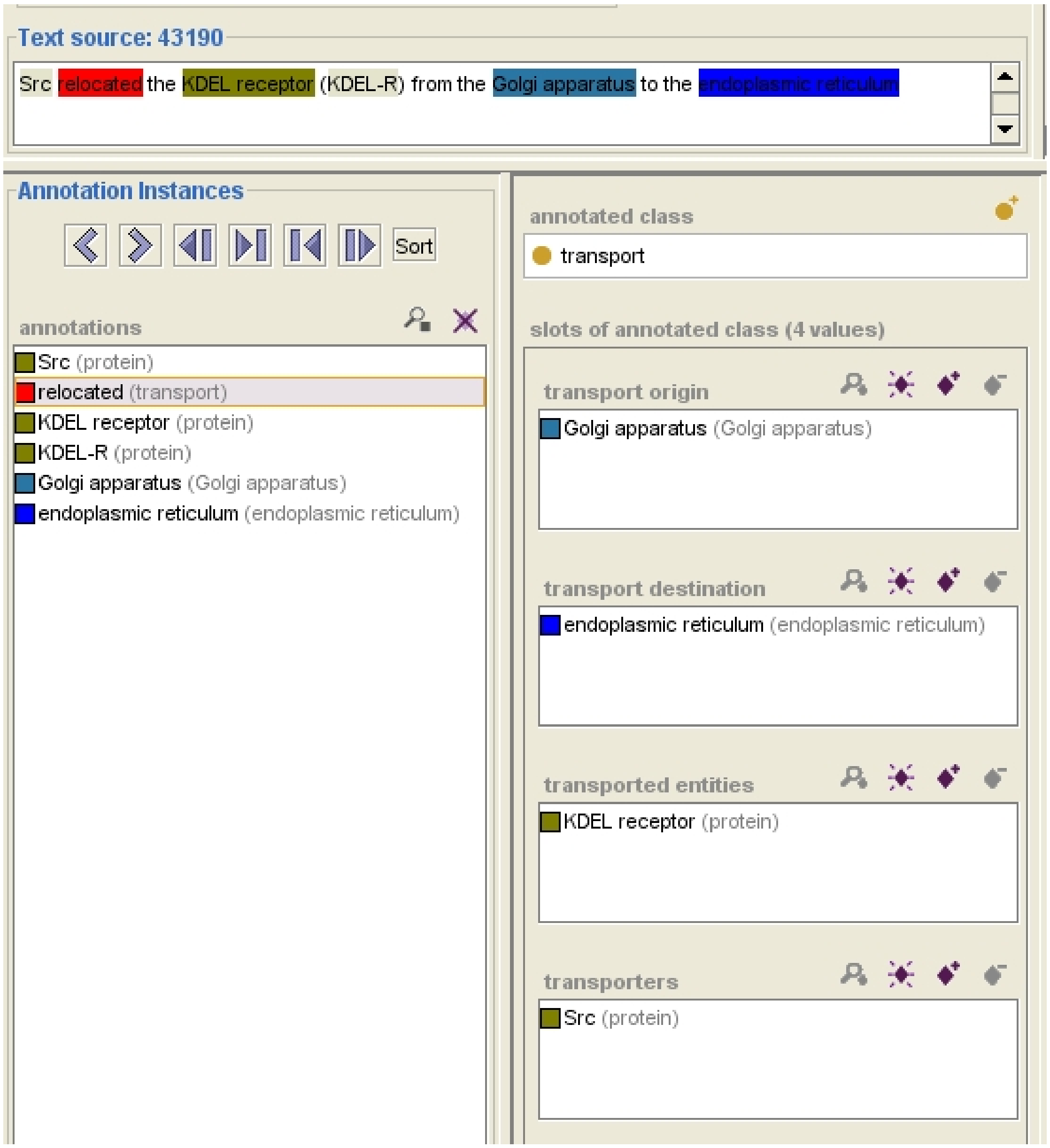
Text source: 43190
Src relocated the KDEL receptor (KDEL-R) from the Golgi apparatus to the endoplasmic reticulum
Annotation Instances
Sort
annotations
Src (protein)
relocated (transport)
KDEL receptor (protein)
KDEL-R (protein)
Golgi apparatus (Golgi apparatus)
endoplasmic reticulum (endoplasmic reticulum)
annotated class
transport
slots of annotated class (4 values)
transport origin
Golgi apparatus (Golgi apparatus)
transport destination
endoplasmic reticulum (endoplasmic reticulum)
transported entities
KDEL receptor (protein)
transporters
Src (protein)

Figure 10