Modeling and exploration of biological networks using the BioXM™ Knowledge Management Environment

Hilmar Ilgenfritz, Wenzel Kalus, Klaus Heumann and Sascha Losko

www.biomax.com

Biomax Informatics AG
Lochhamer Str. 9
D-82152 Martinsried
Tel. +49 89 895574-0
Fax. +49 89 895574-825

The vast quantities of information generated by academic and industrial research groups are reflected in a rapidly growing body of scientific literature and exponentially expanding resources of formalized data including experimental data from "-omics" platforms, phenotype information, clinical data, etc. For information technologies, the challenge remains to support scientists in identifying relevant information, integrating this information in specific "knowledge bases" and formalizing this knowledge across multiple scientific domains to facilitate hypothesis generation and validation and, therefore, the generation of new knowledge. The BioXM™ Knowledge Management Environment efficiently models such complex research environments. This platform is designed for the aggregation of information and the semantic modeling of scientific processes. Any particular area of scientific interest can be modeled as a structured network of related objects. Thus, the BioXM system allows an efficient modularization and abstraction of knowledge.

The actual definition of "knowledge" is indistinct and multifaceted. One aspect of it, certainly, is the awareness of a validated interconnection of details, which are of lesser value when used in an isolated environment. In the BioXM system, knowledge is conceptualized as relationships between semantic objects representing "elements of a scientific domain" (such as genes or drugs). Those relations are supplemented by the annotation of evidence providing validation. For the related objects, further validated relations with other "elements of a scientific domain" (such as cell types or diseases) may exist, expanding the knowledge network. Specific parts of the knowledge model may be organized in sub-network contexts (such as a particular signal transduction pathway in an organism of interest), allowing a hierarchical structuring of knowledge. The organization of information in specific projects provides a further efficient mechanism of distinction between separate parts of the knowledge network.

The conceptualization of entire areas of interest in ontologies allows one to use the inherent inference relationships for the exploration of knowledge networks. Entities from external public or proprietary databases accessible by the embedded BioRS™ Integration and Retrieval System can serve as "virtual semantic objects" in the knowledge network. They can also be used as "read-only" annotation of the "real" semantic objects. All semantic objects (such as elements, relations, contexts, ontology instances or BioRS database entries) can be annotated with additional information. Annotations are form-based and support hierarchical organization of information.

Using the BioLT™ Literature Mining Tool, relationships based on co-occurrence can be extracted from the literature. Curated dictionaries of semantic categories are used for the search, allowing semantic objects in BioXM to be referenced by the search results. This enables the inclusion of the revealed relationships in the existing knowledge network. Search result details, such as sentences and PubMed links, can be attached to the relation as annotation of evidence.

The BioXM system provides graphical browsing through the network. An advanced query builder allows a flexible exploration of the knowledge with complex queries using a natural-language-like syntax. Flexible reporting allows specified sets of information relevant to particular semantic objects to be displayed in one view. A versatile data management system permits one to modify and expand the information networks without the need for additional programming. Thus, research projects can be modeled and extended dynamically.

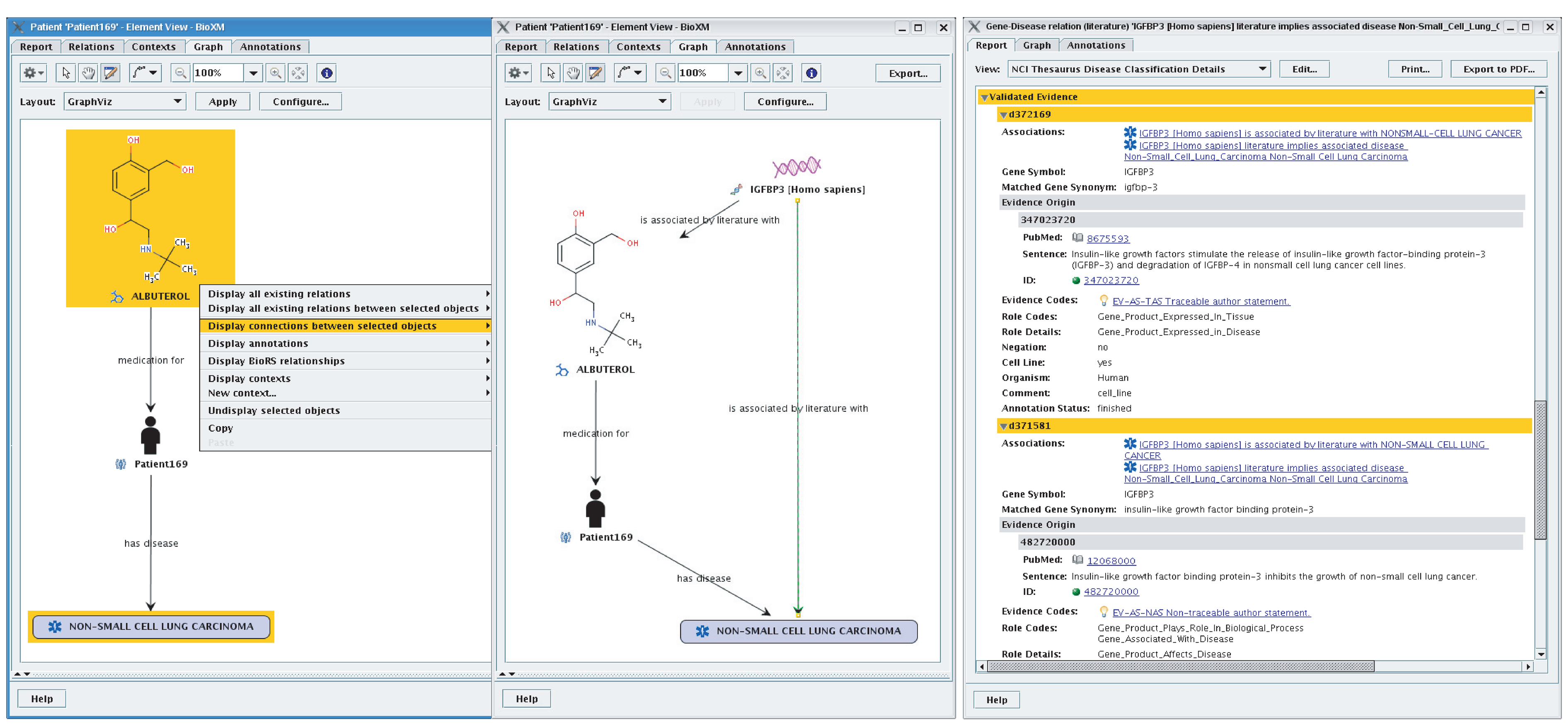

Figure 1. Explore the information network
Left: Connections between objects of a knowledge domain can be found by direct relation or mediated via other objects. The example shows how a gene can be identified that is related to both the selected compound and disease.
Right: The reliability of those connections is easily proven by the annotated evidence validating them.

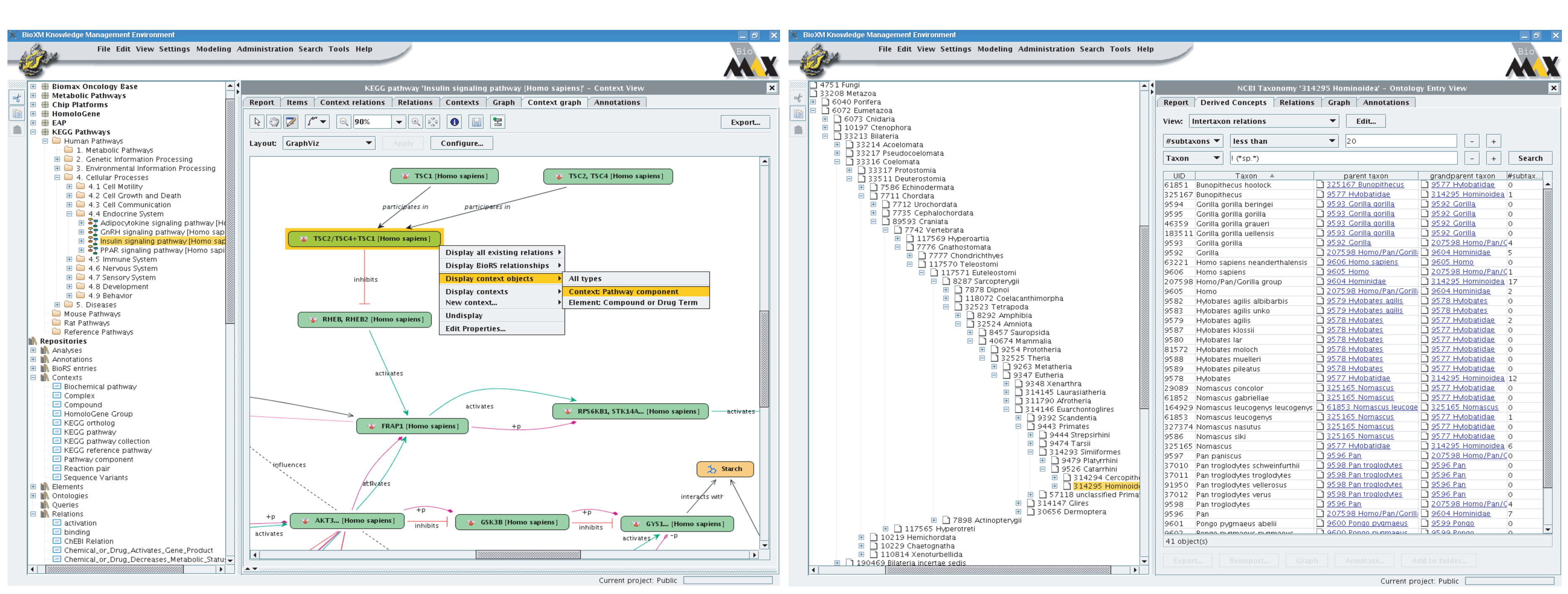

Figure 2. Hierarchical organization of information
Left: Organization of knowledge within project directories allows sequestration of information networks (tree panel). Part of a sub-network ("context") representing a specific human signal transduction pathway is shown in the report panel. The selected object itself is a context representing a protein complex, and the objects participating in this context (constituents of that complex) can be accessed dynamically by context-menu options. The hierarchical semantics of the example shown is "Pathway components TSC2,TSC4 and TSC1 form the protein complex TSC2,TSC4+TSC1 which participates in human insulin signalling pathway which belongs to the endocrine system mediated pathways which belong to cellular processes which are part of Human pathways which belong to the KEGG pathways project".
Right: Ontologies possess an intrinsically hierarchical information structure (tree panel). Each node of the ontology tree with it's derived concepts (i.e., inferred instances) may be explored in the report panel using quick-search and sorting functions and specialized views (e.g., showing parent instances and number of child instances)

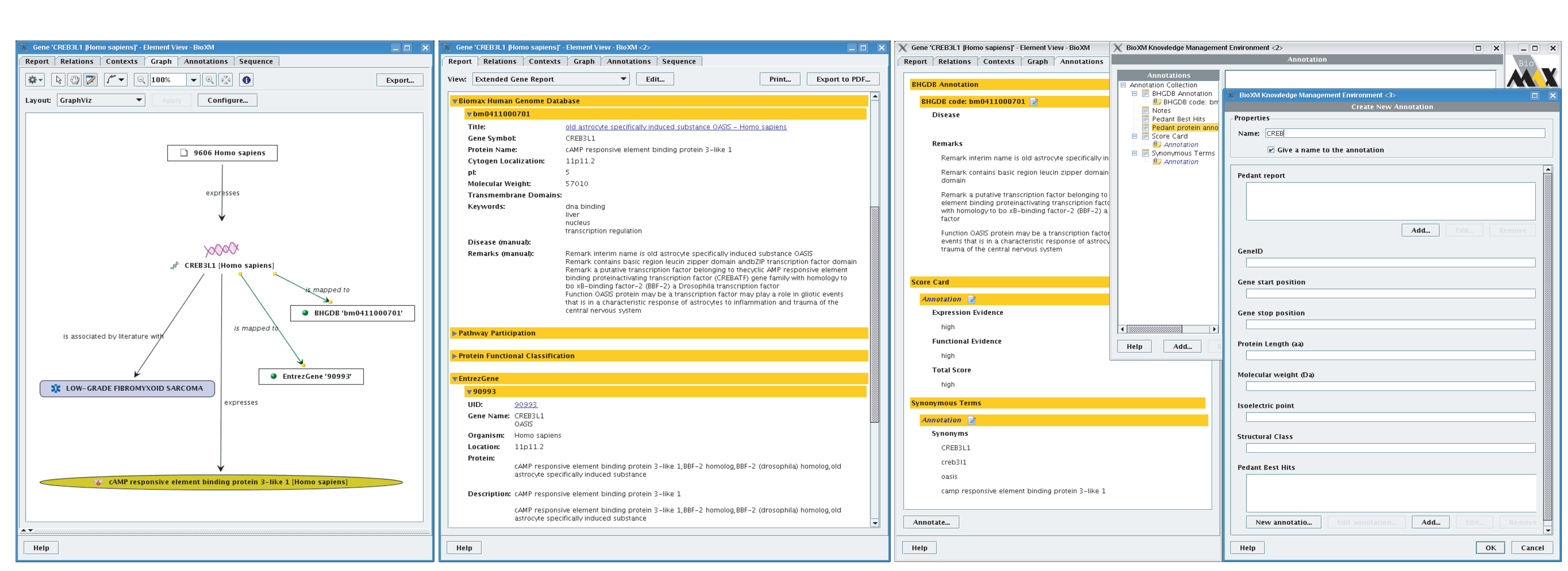

Figure 3. Annotation of information by BioRS databank mapping or by configurable annotation form
Left: A BioXM semantic object can be mapped to entries of external databases accessible via the embedded BioRS retrieval system.
Middle: Information from BioRS database entries can be used as external annotation for the report of the mapped object.
Right: Customary annotation forms can be defined to annotate additional information for semantic objects. Different annotations can be assigned to the same object. The "Annotate" button triggers a dialog which allows one to choose among annotation forms available for the object to create new form-based annotation dynamically.
Annotation can also be created and assigned on larger scales (i.e., for multiple objects) using the data management module (see figure 6).

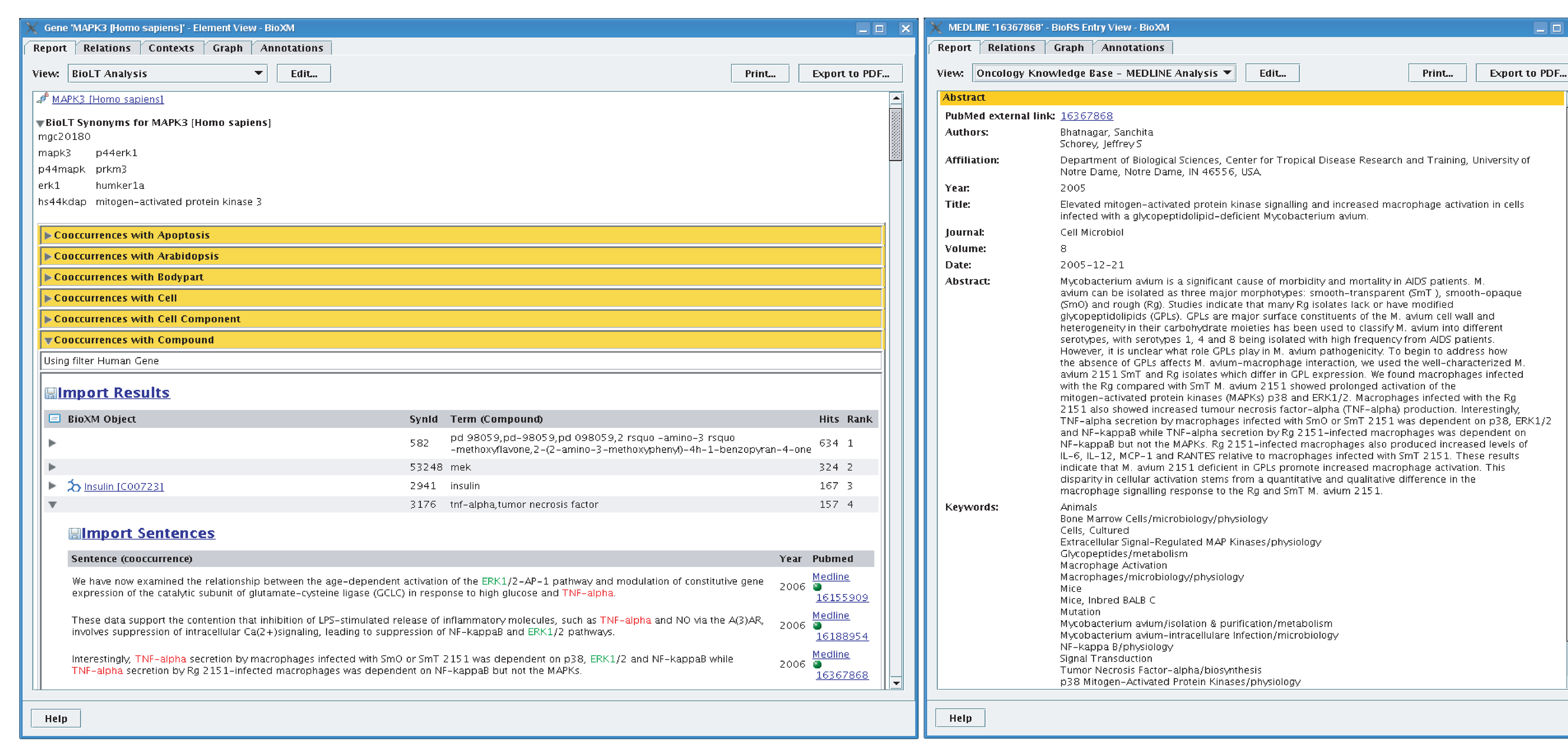

Figure 4. Exploration of literature using the BioLT interface
BioLT uses literature-derived, curated dictionaries of terms and synonyms for different scientific object domains (like "Cell", "Compound", "Disease", "Human Gene", etc.) to mine the literature for co-occurences between objects of one domain with either objects of another domain or a specified text query. BioXM semantic objects can be mapped to those BioLT terms, enabling the user to uncover literature-validated relationships between BioXM objects.
Left: In the example shown, the co-occurrences of the human gene MAPK3 (a list of the synonyms used in the literature is displayed on top) with compound terms are explored. By importing the results, new relations can be established within the knowledge network immediately. For each BioLT co-occurence hit, the corresponding sentences can be viewed and imported as annotation of validation for the relationship. Thus, the knowledge network is expanded dynamically based on literature-validated evidence. Another way to explore literature is to formulate a text query and identify all BioXM objects of a certain type which co-occur with the query text (not shown). This approach uses the advanced query builder (see figure 5) and an evidence reporting similar to the sentence display in the lower part of the shown example.
Right: The PubMed entry corresponding to a BioLT hit is easily accessible as a BioRS-linked object which allows one to view the complete abstract to further validate the scientific context of a co-occurrence.

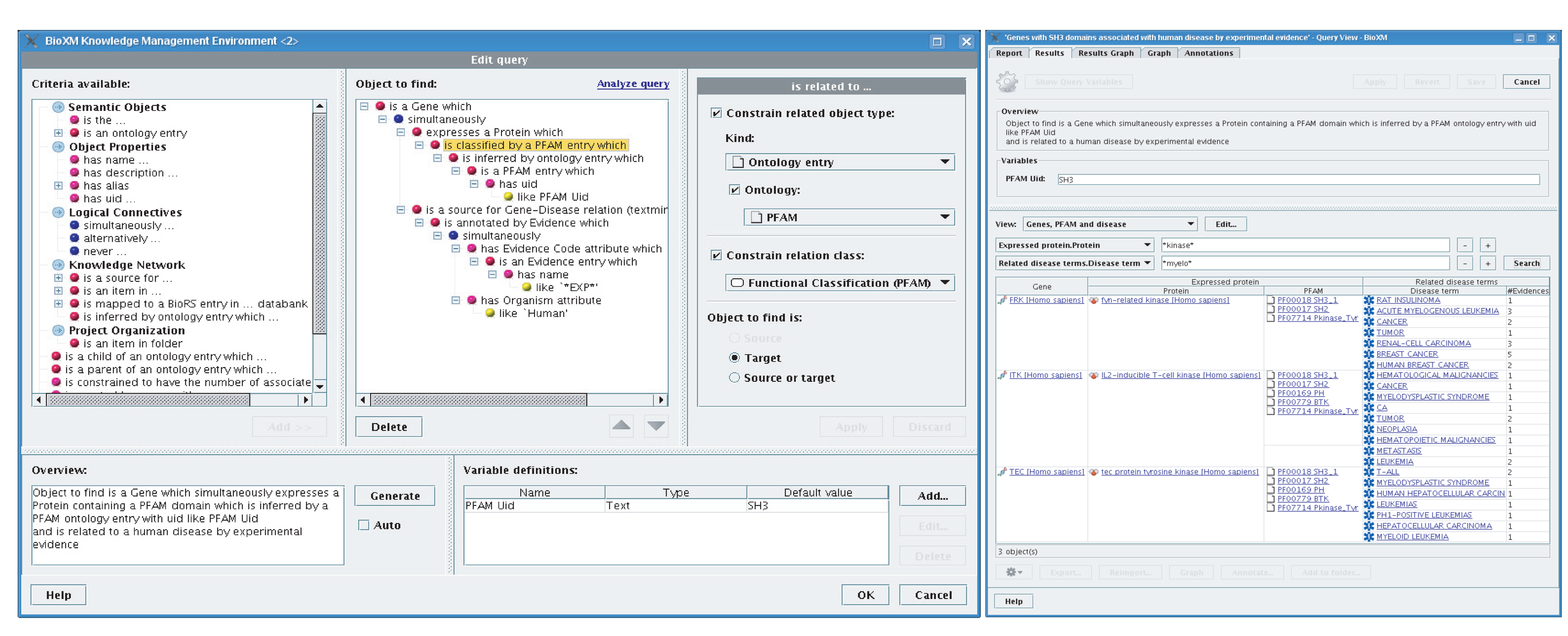

Figure 5. Data retrieval in BioXM
Left: An advanced query builder allows one to construct complex queries in a natural-language-like syntax. The search criteria may reference objects in any distance from the object to be found as long as they are linked in some way. The combination of knowledge network exploration with logical operators, string matching and numerical comparison provides unprecedented flexibility in query construction. Once defined, a query can be stored as template, which allows one to introduce variables for string matching, numerical comparison or object references. Query templates can be used as smart folders for quick retrieval of objects of interest.
Right: Using the query shown on the left as a smart folder, a list of objects was retrieved (note that the query variables may be edited directly in the results report) and further narrowed by the list report's quick-search functionality. Thus, a very specialized set of data is rapidly accessible in an efficient manner.

Figure 6. Data management
The data manager allows one to choose from a number of import and transformation options and combine them into a data modification sequence which is executed for each line of the tabular data document. In this way, huge amounts of data can be modeled in strucured networks of knowledge. Apart from basic import operations, like the creation of objects of a specified type, objects can be retrieved by simple lookup or by query execution, e.g., to create some kind of relation between an object already present in the knowledge network and a newly created one. Annotation can be retrieved and shared with new objects or created for previously retrieved or created objects using the content of specified columns as new attribute values. Data modification sequences can be controlled by the number of results of a preceding operation (e.g., an objekt lookup) or by the content of a specified column. In the example shown here, the creation of a new annotation and its assignment to a previously defined context are carried out only if the preceding query returns no results.