Nucleic Acids Research, 2006, Vol. 34, Database issue ***D235–D242***
doi:10.1093/nar/gkj153

BIOZON: a hub of heterogeneous biological data

Aaron Birkland and Golan Yona*

Cornell University, Ithaca, NY, USA

Received August 15, 2005; Revised and Accepted October 31, 2005

ABSTRACT

Biological entities are strongly related and mutually dependent on each other. Therefore, there is a growing need to corroborate and integrate data from different resources and aspects of biological systems in order to analyze them effectively. Biozon is a unified biological database that integrates heterogeneous data types such as proteins, structures, domain families, protein–protein interactions and cellular pathways, and establishes the relationships between them. All data are integrated on to a single graph schema centered around the non-redundant set of biological objects that are shared by each source. This integration results in a highly connected graph structure that provides a more complete picture of the known context of a given object that cannot be determined from any one source. Currently, Biozon integrates roughly 2 million protein sequences, 42 million DNA or RNA sequences, 32 000 protein structures, 150 000 interactions and more from sources such as GenBank, UniProt, Protein Data Bank (PDB) and BIND. Biozon augments source data with locally derived data such as 5 billion pairwise protein alignments and 8 million structural alignments. The user may form complex cross-type queries on the graph structure, add similarity relations to form fuzzy queries and rank the results based on analysis of the edge structure similar to Google PageRank, online at Biozon.org.

INTRODUCTION

The vast amount of biological knowledge today is distributed among many specialized databases. For example, databases such as SwissProt (1) and PIR (2) focus on protein sequences, while Protein Data Bank (PDB) (3) stores protein structures, Genbank (4) stores DNA sequences, BIND (5) and DIP (6) specialize in protein–protein interactions and databases such as KEGG (7) and MetaCyc (8) focus on cellular pathways. This is a partial list out of hundreds of biological databases that are in existence today. Each database contains detailed information on the entities that are stored within. However, the entities in these databases are strongly related and mutually dependent on each other, and to study one entity it is necessary to know what other entities are related to it. For example, to specify the function of a gene one has to know its extended biological context: the set of interactions it forms, the pathways it participates in, its subcellular location, and so on. Similarly, the cellular function of a pathway is a function of its constituent genes. In this view, integration of data from different resources and aspects of biological systems becomes a necessity in biological data analysis.

Traversing the available information on a specific entity is a difficult task, as it is presently fragmented and stored in multiple sources, not necessarily well connected. To obtain the knowledge about the broader biological context of an entity one has to query multiple databases and unify this knowledge into a consistent and non-redundant set, a task that requires certain skills and can be time consuming. And while possible manually on a very small scale, data integration is a major challenge when faced with the millions of entities that are stored on digital media today. Most importantly, the lack of a unified integrated schema makes it difficult to query this wealth of data in ways that can benefit and exploit the mutual dependency between entities.

This challenge becomes especially difficult because of the methods used to identify entities in biological databases. Existing databases use explicit references such as accession nos, and while some databases relate and cross link elements from other databases based on these identifiers, this information is partial and is not readily available in some. Moreover, these cross links are not necessarily maintained in coordination with the other rapidly changing linked databases and this leads to problems of consistency.

While there exist some advanced index-based search engines such as SRS (9) and Entrez (10) as well as mediator-based search engines that distribute and unify query results from heterogeneous sources such as K2 (11), Kleisli (12), TAMBIS (13)

*To whom correspondence should be addressed at Department of Computer Science, Technion, Israel. Tel: +972 4 8294356; Fax: +972 4 8293900; Email: golan@cs.cornell.edu

and DiscoveryLink (14), Biozon is unique in a number of respects. First, Biozon gathers its data into a single warehouse, with obvious performance benefits for searches that comprise data that originate from multiple sources. Second, entities are merged into a set of non-redundant objects in Biozon based on their physical properties rather than based on identifiers. This has the benefit of disambiguating explicit cross-references, and mapping all relationships into a single, consistent form that is straightforward to query without knowing the specific semantics and caveats of heterogeneous nomenclatures. Lastly, the Biozon graph is augmented with derived data, such as protein–protein alignments that are computed on the entire integrated graph structure and are maintained accordingly as the graph changes, to reflect updates in the constituent sources.

The Biozon database is accessible at biozon.org. The graph structure of Biozon lends itself to many unique features that are publicly available, among which are:

- Browsing and navigating capability that shows the biological context of each object in its own ‘profile page’.
- Interface for building complex and fuzzy queries. By using a step-by-step procedure, users can add objects and relationships to create a search topology, as well as define all search constraints.
- Ranking of search results using methods similar to Google PageRank (15) that takes into account the edge structure of the Biozon data graph.
- User accounts system that allows one to ‘attach’ comments to objects and to store the results of complex queries. Queries can be saved, run in the background and their results may be downloaded.
- Online analysis tools for user-supplied data. Currently, Biozon offers BLAST comparisons of submitted sequences with Biozon proteins, EST analysis, expression profile similarity and domain structure prediction.

DATA MODEL

The information in Biozon is logically represented as a graph in which nodes represent some unit of data, and edges indicate a relationship between two nodes (Figure 1). Each graph node or edge is of a certain class. The different classes are organized into a hierarchy of document and relation types (Supplementary Figure S1). This hierarchy serves to refine the data model and increase its expressiveness and simplifies maintenance and expandability. For example, at the top of the document hierarchy we make a distinction between descriptors and objects. Objects are the backbone of the Biozon data graph and the descendant subclasses of the object type are the equivalent of physical objects (such as protein sequences and DNA sequences) and sets thereof (such as interactions and pathways). Descriptors on the other hand contain human-readable information that makes these entities meaningful. Each descriptor is associated with one object and contains annotations, measurements and other pieces of evidence that refer to the object they describe.

The *object-centric* approach of Biozon has implications on integration as well as on searches and emergent graph structure. Biozon employs vertical integration (16) where identical objects are mapped to the same node in Biozon. Each constituent database or dataset is mapped on to some subset of the Biozon graph (Figure 2) and instances of fundamental biological objects that are gleaned from these sources are mapped to object nodes and are required to be non-redundant. As such, physical objects, such as protein sequences, can be viewed as the actual identifiers of the biological entities they represent. Any other information that exists in the source regarding the physical entity is stored in a descriptor document that is linked to the object. Due to the vertical integration, each object might be associated with multiple descriptors that originate from multiple sources, all describing the same entity. As a result, our graph ends up becoming highly connected and centered around hubs of such objects. This connectivity allows for efficient formulation and execution of complex queries that span multiple data types. The use of potentially unstable or inconsistent identifiers (such as accession numbers) to indicate relationships or cross-references is eschewed in favor of using explicit relations between physical non-redundant Biozon objects. [The great variety of identifiers that are used by different databases and the lack of coordination between the databases pose a major problem for biological data integration. Our approach, that utilizes physical properties rather than arbitrary identifiers, consolidates the information from multiple sources while ensuring a faithful representation of the source databases in Biozon. However, since two entities are considered to be the same object in Biozon only if they are 100% identical (where the notion of identity depends on the object type), this approach might result in an over-fragmented graph (e.g. when the representations of the same sequence in different databases differ only slightly). Nevertheless, such sequences are related through similarity relations, thus are easily accessible from a single access point at Biozon.]

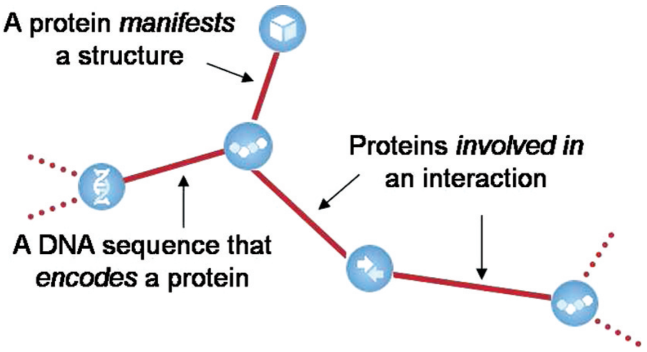

A protein *manifests* a structure
Proteins *involved in* an interaction
A DNA sequence that *encodes* a protein

Figure 1. A schematic representation of a section of the Biozon data graph at the instance level. Each node or edge is an instance of a class in the class hierarchy (Supplementary Figure S1). This particular section of the graph centers around a protein that has a known structure, is coded by a known DNA sequence, and is involved in an interaction. On the complete graph, all types (e.g. pathways, EC families, domains, etc.) are represented.

By enforcing non-redundancy on objects, relationships between the data expressed in different sources referring to the same object become explicit and unambiguous in their graph representation. For example, suppose there is a protein sequence that is defined in SwissProt that has a relation to a domain object from InterPro. Assume this protein is also listed in BIND as a member of an experimentally verified interaction. However, BIND uses GI numbers to identify the proteins that are involved in interactions, and thus the link between the InterPro domain and the interaction can be difficult to

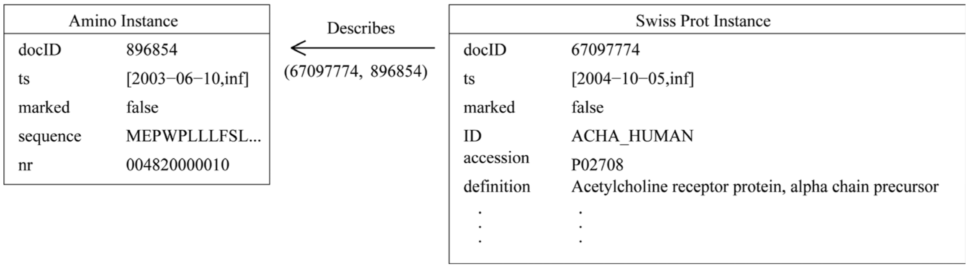


Figure 2. Mapping a SwissProt instance to the Biozon graph. Each dataset maps to a set of nodes and edges in the Biozon graph, each of which is a member of the class hierarchies presented in Supplementary Figure S1. This is the simplest representation in Biozon: one object related to one descriptor containing searchable annotation. For example, a SwissProt record of a protein is transformed into an instance of an amino acid sequence object and a SwissProt descriptor that are related together by the 'describes' relation.

Table 1. Data statistics as of August 2005

Data type	NR record count	Sources
Nucleic acid sequences	42 686 711	GenBank, RefSeq, BIND, UniGene, BodyMap
Protein sequences	2 062 061	SwissProt, TrEMBL, GenPept, PDB, BIND, PIR, RefSeq, SCOP, DDBJ, PRF, PATAA
Protein structures	32 637	PDB
Interactions	155 090	BIND, Biozon (predicted), DIP[a]
Enzyme families	3944	UniProt, PIR, GenPept, SCOP
Pathways	142	KEGG
Unigene clusters	185 543	NCBI
Domain families	181 500	InterPro (includes data from PFAM, PRINTS, PRODOM, PROFILE, PROSITE, SMART, SSF, TIGRFAM), Biozon (predicted)
Sequence alignments	5 000 000 000+	Biozon
Structure alignments	8 250 286	Biozon
Expression similarities	68 138	Biozon
Non-similarity relations	136 972 705	All
Descriptor documents	58 176 040	All
Words indexed	1 627 747 755	All

Numbers and origin of a selection of Biozon objects relations and indexed text. The database will be gradually extended to span both new source data types as well as new derived data. All data from the source 'Biozon' is derived data, either in the forms of predictions (e.g. predicted interactions) or similarity (e.g. sequence alignments).

[a]DIP is not publicly accessible due to copyright restrictions.

ascertain. In Biozon's non-redundant object model, the SwissProt protein and the protein associated with the GI number will be mapped to the same object, and this one object will be linked to both the domain object and the interaction object. In other words, there is now a path in the graph between the domain and the interaction objects. Paths such as these are easy and efficient to mine for when searching for relationships between objects, and their presence is attributable to our non-redundant object model.

Our database model has many benefits in search and analysis as discussed in the next sections. These benefits come at the price of having to address issues of maintenance and consistency. These issues require special attention as they concern the data consistency between Biozon and the individual databases, between representations within Biozon, and with respect to derived relations. To address these issues we developed detailed update protocols that work to preserve consistency with updates. Detailed discussion of all these elements is beyond the scope of this paper and will appear elsewhere, but it is important to emphasize that these were made possible because of our design choices.

It should also be noted that Biozon archives scientific data as no document is ever deleted. Rather, each document is associated with a timeline that indicates its relevance. This property allows time-dependent queries.

Datasets

The data in Biozon are composed of two main types: **source data** that are gleaned from established online databases (such as Swiss-Prot, GenBank, BIND, KEGG and others), and unique **derived data** that is computed in house and includes similarity relationships between objects and predicted information about their functional role. The derived data introduces another level of complexity to our model, but also allows for even more powerful methods of data querying, and manipulation (as is demonstrated in the section that discusses fuzzy searches).

Currently, Biozon stores extensive information about 50 million objects that is distributed in more than 100 million documents. This data originates from more than 20 different sources, listed in Table 1. It also stores over 5 billion relations between documents, including explicit relations between objects, and derived or computed relations based on sequence similarities, structural similarities and more. [Sequence similarities are computed using BLAST (17) and stored in the database if their E-value <0.1. Structural similarities are computed using CE (18), Dali (19) and URMS (20) and stored if their E-value <1. Expression similarities are computed using the mass-distance function (21) and stored if their E-value <10.]

SEARCH AND ANALYSIS

In parallel to the development of the Biozon graph model we developed tools to browse, visualize and search the entities that are stored within.

Browsing entities in Biozon

Each document in Biozon has a profile page in the web interface. Descriptor profile pages are linked to the profile pages of the objects they describe. Profile pages of objects serve as hubs that link all other documents (objects and descriptors) that make up the broader biological context of the object. Thus, a user is provided instantly with a myriad of relevant information that is otherwise difficult to attain, as it is distributed in multiple disperse sources. One such example is shown in Figure 3 for a protein sequence that is linked to eleven other objects (excluding similarity) and 19 descriptors that originate from at least five sources.

Searching entities in Biozon

The homepage of Biozon at biozon.org presents the user with a schematic representation of the main data types and their relationships. This is a simplified view of the data graph that focuses on the main elements that are stored in Biozon. Each type of data is searchable through a *type-specific* form, accessible by clicking on the corresponding icon. The form displays all attributes that are associated with this data type and were indexed in Biozon. These attributes are collected from multiple sources. Thus, a user can search, e.g. GenBank accession numbers and GO terms using the same form. The attributes are generally divided into four categories: identifiers (accession number, gene names, etc.), descriptions (definition, keywords, GO terms, etc.), physical properties (such as length, or resolution) and taxonomy.

A type-specific search returns a list of entities of that type that match the query, from which the profile page of each entity can be reached and viewed. Another option that the user is presented with in the main homepage is to run a *quick search* that would return *all* entities of all data types that match the query term. While this feature resembles that of Entrez (10), it is the connectivity between data types and the ability to include relations in searches that distinguishes Biozon the most from other sources, as discussed in the next section.

Complex searches

The graph structure of Biozon gives rise to multiple *data topologies*, that is, subgraphs of data types and their relationships. Biozon exploits the graph structure in allowing for complex searches that span multiple data types. The user initiates a search by specifying a graph topology to search for, as well as the constraints on any documents that should be present in matching topology instances as is demonstrated in Figure 4. For example, a valid query could be '*Find all 3D structures of proteins that are involved in phosphorylation interactions and are part of the Prostaglandin and leukotriene metabolism pathway*.' This particular search specifies a topology involving structures, proteins, interactions and pathways. [Currently, the specified topology for this search would involve 'enzyme families' as well. Biozon incorporates metabolic pathways from KEGG, which defines them in terms of enzyme families. Therefore, a path in the graph between proteins and pathways would have to go through the enzyme family node.] The online user-interface allows these queries to be built in a series of simple steps, as is exemplified in the online demonstration at biozon.org/help/. The first step determines the data type the user would like to get as a result (the *target type*). Upon execution, the Biozon graph is searched for graph isomorphism in realtime and all entities of the target type that are members of such topologies are returned as results. [We are developing efficient algorithms that can search for all data topologies in the Biozon data graph that connect the query objects, and future releases will return also the complete list of topologies observed and their constituent instances.]

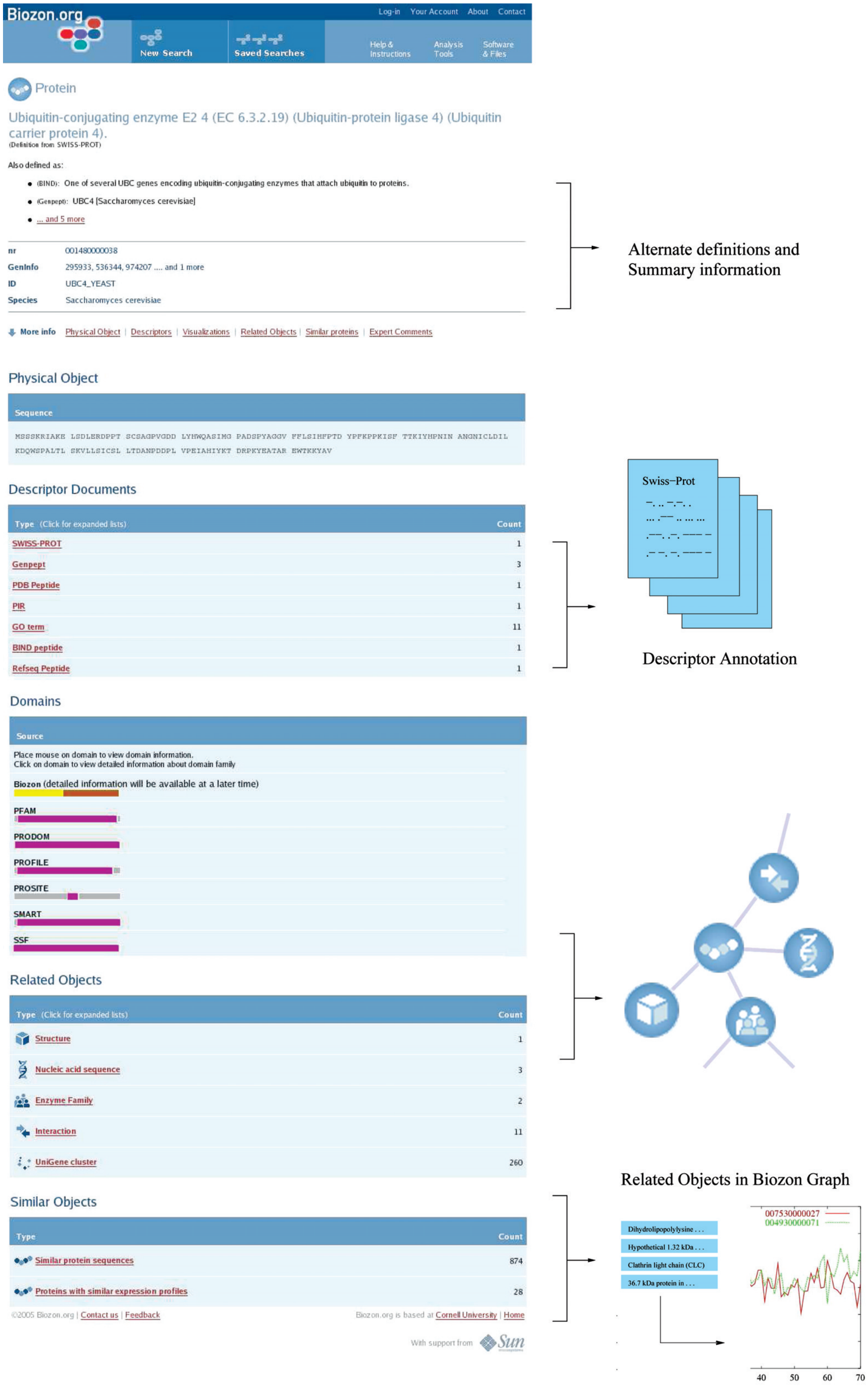
Fuzzy searches

Despite the extensive information that is already stored in biological databases, biological knowledge is partial by its very essence and although great strides have been made in developing technologies that can measure different aspects of biological systems, on many levels molecular and cellular entities are only partially understood. Moreover, because of limited resources, most entities were never studied experimentally. As a result, annotations are often incomplete, and many entities still await analysis.

Much of biological data analysis these days is done *in silico*, and perhaps the most popular of which is prediction by similarity. The similarity relation is one of the most fundamental relations in biology, frequently used for functional inference. To allow large-scale knowledge propagation from well-studied entities to uncharacterized ones Biozon computes and stores extensive similarity indices, based on a variety of comparison methods and representations. Biozon currently includes similarity relations between proteins based on sequence, structure and similarities based on gene expression data. Other similarity measures will be gradually integrated into the system.

Beyond the ability to view entities that are similar to a specific entity, the similarity relations that are stored in Biozon allow for a powerful method of query in the form of *fuzzy searches*. Fuzzy searches extend complex queries to include similar or homologous objects in the search space. Queries may be extended by incorporating the similarity data in any appropriate query step. For example, at one time, querying for structures of proteins that are in enzyme family 1.1.1.1 and are involved in an interaction returned no results. Incorporating similarity into the search transforms the query into one that searches for structures of proteins that are involved in interactions and are members of the 1.1.1.1 family or proteins that are *similar* to these proteins (Figure 5). This query does return significant results from a very large search space in less than minute.

Currently, fuzzy searches use the results of BLAST, expression profile similarity, and structural similarity of proteins. A user may choose which are incorporated into the results, as well as the similarity threshold (such as *E*-value), where applicable. Because Biozon materializes all similarity data, none of the expensive similarity computations need to be done when executing such a search.



Biozon.org

Log-in | Your Account | About | Contact

New Search | Saved Searches | Help & Instructions | Analysis Tools | Software & Files

Protein

Ubiquitin-conjugating enzyme E2 4 (EC 6.3.2.19) (Ubiquitin-protein ligase 4) (Ubiquitin carrier protein 4).
(Definition from SWISS-PROT)

Also defined as:

- (BIND): One of several UBC genes encoding ubiquitin-conjugating enzymes that attach ubiquitin to proteins.
- (Genpept): UBC4 [Saccharomyces cerevisiae]
- ... and 5 more

nr	001480000038
GenInfo	295933, 536344, 974207 .... and 1 more
ID	UBC4_YEAST
Species	Saccharomyces cerevisiae

More info | Physical Object | Descriptors | Visualizations | Related Objects | Similar proteins | Expert Comments

Physical Object

Sequence

```
MSSSKRIAKE LSDLERDPPT SCSAGPVGDD LYHWQASIMG PADSPYAGGV FFLSIHFPTD YPFKPPKISF TTKIYHPNIN ANGNICLDIL
KDQWSPALTL SKVLLSICSL LTDANPDDPL VPEIAHIYKT DRPKYEATAR EWTKKYAV
```

Descriptor Documents

Type (Click for expanded lists)	Count
SWISS-PROT	1
Genpept	3
PDB Peptide	1
PIR	1
GO term	11
BIND peptide	1
Refseq Peptide	1

Domains

Source

Place mouse on domain to view domain information.
Click on domain to view detailed information about domain family

Biozon (detailed information will be available at a later time)

PFAM

PRODOM

PROFILE

PROSITE

SMART

SSF

Related Objects

Type (Click for expanded lists)	Count
Structure	1
Nucleic acid sequence	3
Enzyme Family	2
Interaction	11
UniGene cluster	260

Similar Objects

Type	Count
Similar protein sequences	874
Proteins with similar expression profiles	28

©2005 Biozon.org | Contact us | Feedback

Biozon.org is based at Cornell University | Home

With support from Sun

Alternate definitions and Summary information

Descriptor Annotation

Related Objects in Biozon Graph

Example of Profile Similarity

Swiss–Prot
Dihydrolipopolylysine . . .
Hypothetical 1.32 kDa . . .
Clathrin light chain (CLC)
36.7 kDa protein in . . .
007530000027
004930000071
40
50
60
70

Figure 3. A profile page displays an overview of the data in Biozon relating to a particular object (such as SwissProt protein UBC4_YEAST). At the top is a brief summary of some of the most generally useful attributes, followed by a physical representation (if applicable), links to relevant descriptor annotation documents, neighboring objects in the Biozon graph, and similar objects. The illustrations on the right side are a schematic representation of the information that is available either one hop or two hops away from the profile page of the object viewed. Similarity relations include similarities based on sequence or based on expression profiles of the corresponding mRNA sequences.

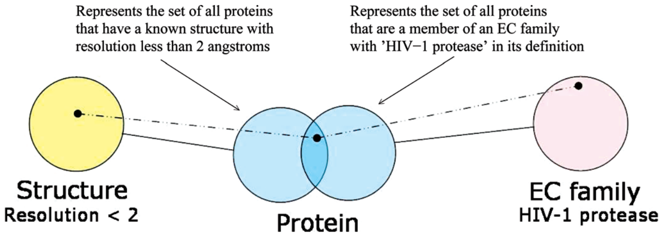

Represents the set of all proteins that have a known structure with resolution less than 2 angstroms
Represents the set of all proteins that are a member of an EC family with 'HIV−1 protease' in its definition
Structure
Resolution < 2
Protein
EC family
HIV-1 protease

Figure 4. Complex search. This is a schematic representation of the complex query: 'Find all structures with resolution less than 2 angstroms of proteins that are in the HIV-1 protease enzyme family'. Results are returned when there are paths conforming to the given topology through objects that satisfy all constraints (instances in the dark-shaded intersection in the middle). In this case, a set of structures will be returned, each one a member of a topology satisfying the complex search parameters.

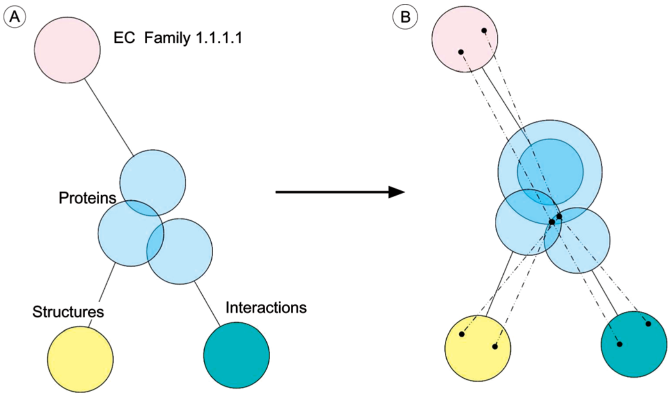


Figure 5. Fuzzy search for all proteins in enzyme family 1.1.1.1 that have a known structure and are involved in an interaction. (**A**) Standard search (as of April 2005) returned no results: There are no proteins that are a member of all three sets (EC 1.1.1.1, have known structure, and in an interaction). (**B**) To form a fuzzy search, the set of proteins in EC 1.1.1.1 is extended to include those *similar* to ones in EC family 1.1.1.1. In this example the fuzzy search returns two results.

Ranking

The graph structure of Biozon and the emergent graph resulting from integration of many sources lends itself well to analysis. Particularly relevant to searches and result sets is the ability to assign ranks to Biozon objects based on the graph structure.

We studied different prominence models that were proposed in the context of web searches, such as PageRank (15) and Hubs & Authorities (22), all based on spectral analysis of the data graph. To assign prominence values to objects Biozon employs a method that resembles PageRank by Google (the detailed methodology will be described elsewhere) and search results may be ordered by ranks. These values reflect the importance of the different entities and are correlated with the amount of information associated with them.

As an example, we contrast the results of searching for proteins containing the term 'cancer' in their definition with and without ranking (Figure 6). When the results are ranked, the top ranked hits tend to be the most highly connected on the graph, as compared to a random ordering. These high ranked objects can direct biologists that study specific systems to other equivalent systems that were studied extensively, thus providing them with myriad of relevant information.

User accounts and data materialization

A major emphasis of Biozon is data and knowledge dissemination, and in that spirit the Biozon web server allows users to open user accounts. These accounts serve multiple purposes. For example, users can save and manage their queries. This feature can become quite useful if one would like to reproduce search results, especially with complex or fuzzy queries that contain many clauses.

Furthermore, users these days often want more than just a list of results. They might want to *materialize* and store the data on their computer for further analysis. Clearly, collecting results from complex queries that involve multiple data types that reside in multiple sources is a very time consuming and difficult task that requires certain expertise. Biozon addresses this problem by allowing users to materialize and download the results of their queries for further analysis, through their user accounts. Moreover, utilizing the time stamp feature of Biozon documents, users can choose to materialize and re-materialize the results as of arbitrary times in the past, thus allowing them to reproduce the same result set that was obtained on the original query date.

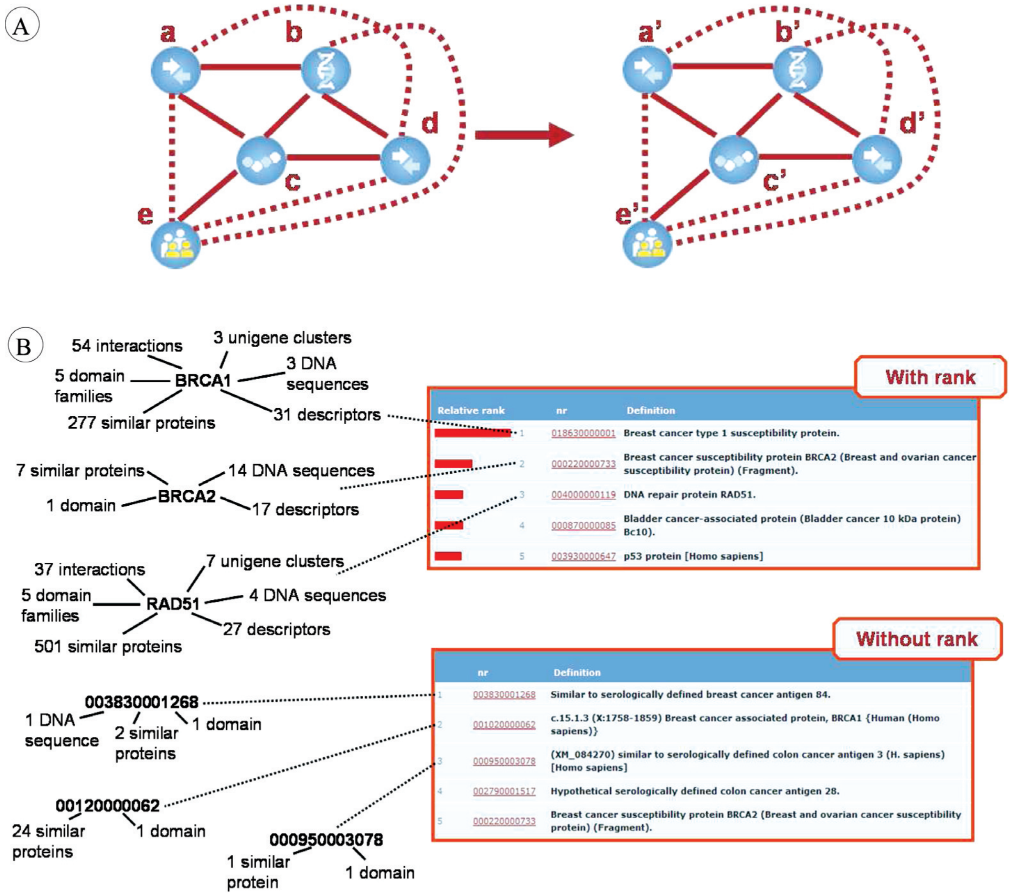
Another purpose of user accounts is to allow users to access private address spaces. This feature becomes especially useful if some datasets are privately owned and cannot be made accessible to all, or if there are restrictions on the use and distribution of the data [as is the case with DIP (6)]. To allow certain groups of users to access protected data that they are privileged to, Biozon defines different classes of users that are mapped to different domains of data. Each user can belong to multiple user classes, thus maximizing flexibility in data accessibility.

Analysis tools

Biozon offers several additional services that are geared toward analyzing user data. For example, users can submit a sequence and search for similar sequences. This service is quite standard today in many other databases. In addition, Biozon offers a domain prediction server that takes a user-supplied sequence and predicts its domain structure using the algorithm of (23). Biozon also offers an EST analysis tool that explores different paths in the data graph to map EST sequences to their protein products, as is demonstrated in Supplementary Figure S2. The currently available services are listed at biozon.org/tools/.

DISCUSSION

In this paper we describe Biozon, a database system that integrates biological information at the macro-molecular level as well as at the cellular level, from a variety of resources. Biozon is constructed as a tightly integrated schema that provides the greatest benefit for performing large scale analysis and integrating subsequent results into the database. The logical graph schema is based around a backbone of interrelated, fundamental biological objects. A key factor in design was the decision that these fundamental objects would be non-redundant. An important consequence of this decision is the



Relative rank		nr	Definition
▬	1	018630000001	Breast cancer type 1 susceptibility protein.
▬	2	000220000733	Breast cancer susceptibility protein BRCA2 (Breast and ovarian cancer susceptibility protein) (Fragment).
▬	3	004000000119	DNA repair protein RAD51.
▬	4	000870000085	Bladder cancer-associated protein (Bladder cancer 10 kDa protein) Bc10).
▬	5	003930000647	p53 protein [Homo sapiens]

	nr	Definition
1	003830001268	Similar to serologically defined breast cancer antigen 84.
2	001020000062	c.15.1.3 (X:1758-1859) Breast cancer associated protein, BRCA1 {Human (Homo sapiens)}
3	000950003078	(XM_084270) similar to serologically defined colon cancer antigen 3 (H. sapiens) [Homo sapiens]
4	002790001517	Hypothetical serologically defined colon cancer antigen 28.
5	000220000733	Breast cancer susceptibility protein BRCA2 (Breast and ovarian cancer susceptibility protein) (Fragment).

Figure 6. Ranking search results (**A**) Computing ranks: Each entity confers its weight on its neighboring entities (solid lines) with probability α, and to a random node selected from all graph nodes (dashed lines) with probability 1 − α, imitating a random walk through the document space. The computation starts with random weights and re-assigns weights until convergence. (**B**) We show the first five results from a search for proteins that contain the word 'cancer' in their definition, with and without using ranks. Each instance is shown with its context (the entities that are connected to it).

property that as objects defined in different databases overlap, they become more highly connected in the graph therefore volunteering the nature of their true biological context.

Externally, Biozon may appear to resemble services such as Entrez and SRS. They are, however, quite different. Both Entrez and SRS are services that perform searches over indexed text and are perhaps best classified as performing search aggregation as opposed to integration. SRS allows for some advanced capability, such as forming cross-dataset searches that take advantage of hyperlinks or cross-references between data entries. In addition, SRS allows for the creation and manipulation of views for organizing how the data from searches is displayed. Biozon, on the other hand, is not a search aggregator, it is an integrator that maps the external sources into a large data graph where the relationships (edges) are determined by the integration principles employed by Biozon and deduced from the physical attributes of the entities in the source databases. Cross-references, hyperlinks, and the like are ignored unless the exact nature of the link is known.

Our design not only allows to view each biological entity in its extended biological context through its relations to other biological entities, it also allows complex searches on the data graph that specify desired interrelationships between types. Furthermore, with the integration of similarity data one can easily extend these queries to accommodate fuzzy relationships between proteins and extend results to sets based on homology instead of direct reference. The graph structure also lends itself to a spectral analysis that is the basis for a first-of-a-kind biological ranking system which resembles the PageRank method of Google (15).

Furthermore, we created an account management system that enables users to save queries and materialize data from search results to download for further analysis or include as a step in some other complex query. Finally, the database is accompanied with a sophisticated web interface that is accessible online at biozon.org. An online manual is available at biozon.org/help/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation (NSF) under Grant No. 0133311 to Golan Yona and by Sun MicroSystems. Funding to pay the Open Access publication charges for this article was provided by NSF.

Conflict of interest statement. None declared.

REFERENCES

1. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
2. George,D.G., Barker,W.C., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1996) The PIR-international protein sequence database. *Nucleic Acids Res.*, **24**, 17–20.
3. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W. and Weissig,H., Greer,D.S. *et al.* (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
4. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) GenBank. *Nucleic Acids Res.*, **27**, 12–17.
5. Bader,G.D., Donaldson,I., Wolting,C., Ouellette,B.F., Pawson,T. and Hogue,C.W. (2001) BIND-the biomolecular interaction network database. *Nucleic Acids Res.*, **29**, 242–245.
6. Xenarios,I., Fernandez,E., Salwinski,L., Duan,X.J., Thompson,M.J., Marcotte,E.M. and Eisenberg,D. (2001) DIP: the database of interacting proteins: 2001 update. *Nucleic Acids Res.*, **29**, 239–241.
7. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
8. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc database. *Nucleic Acids Res.*, **30**, 56–58.
9. Etzold,T. and Argos,P. (1993) SRS-an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, **9**, 49–57.
10. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–161.
11. Davidson,S.B., Crabtree,J., Brunk,B.P., Schug,J., Tannen,V., Overton,G.C. and Stoeckert,C.J.,Jr (2001) K2/Kleisli and GUS: experiments in integrated access to genomic data sources. *IBM Syst. J.*, **40**, 512–530.
12. Chen,J., Chung,S. and Wong,L. (2003) The Kleisli Query System as a Backbone for Bioinformatics Data Integration and Analysis. In Lacroix,Z. and Critchlow,T. (eds), *Bioinformatics: Managing Scientific Data,* Morgan Kaufmann, Chapter 6, pp. 147–188.
13. Baker,P.G., Brass,A., Bechhofer,S., Goble,C.A., Paton,N.W. and Stevens,R. (1998) TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pp. 25–34.
14. Haas,L., Schwarz,P., Kodali,P., Kotlar,E., Rice,J. and Swope,W. (2001) DiscoveryLink: a system for integrated access to life sciences data sources. *IBM Syst. J.*, **40**, 489–511.
15. Page,L., Brin,S., Motwani,R. and Winograd,T. (1998) The PageRank citation ranking: bringing order to the web. Technical report Stanford Digital Library Technologies Project.
16. Hernandez,T. and Kambhampati,S. (2004) Integration of biological sources: current systems and challenges ahead. *SIGMOD Rec.*, **33**, 51–60.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
19. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
20. Yona,G. and Kedem,K. (2005) The URMS-RMS hybrid algorithm for fast and sensitive local protein structure alignment. *J. Comput. Biol.*, **12**, 12–32.
21. Yona,G., Dirks,W., Rahman,S. and Lin,D. (2005) Effective similarity measures for expression profiles. *Bioinformatics*, in press.
22. Kleinberg,J.M. (1998) In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 668–677.
23. Nagarajan,N. and Yona,G. (2004) Automatic prediction of protein domains from sequence information using a hybrid learning system. *Bioinformatics*, **20**, 1335–1360.