Exploring Biomedical Databases with BioNav

Abhijith Kashyap
CSE Dept
SUNY Buffalo
Buffalo, NY 14260
rk39@cse.buffalo.edu

Vagelis Hristidis
School of Computing and Information Sciences
Florida International University
Miami, FL 33199
vagelis@cis.fiu.edu

Michalis Petropoulos
CSE Dept
SUNY Buffalo
Buffalo, NY 14260
mpetropo@cse.buffalo.edu

Sotiria Tavoulari
Pharmacology Dept
Yale University
New Haven, CT 06520
sotiria.tavoulari@yale.edu

ABSTRACT

We demonstrate the BioNav system, a novel search interface for biomedical databases, such as PubMed. BioNav enables users to navigate large number of query results by categorizing them using MeSH; a comprehensive concept hierarchy used by PubMed. Once the query results are organized into a navigation tree, BioNav reveals only a small subset of the concept nodes at each step, selected such that the expected user navigation cost is minimized. In contrast, previous works expand the hierarchy in a predefined static manner, without navigation cost modeling. BioNav is available at http://db.cse.buffalo.edu/bionav.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process, selection process and information storage and filtering.* H.3.6 [**Information Storage and Retrieval**]: Digital Libraries – *Dissemination, User Issues.* J.3 [**Computer Applications**]: Life and Medical Sciences –*Medical Information Systems.*

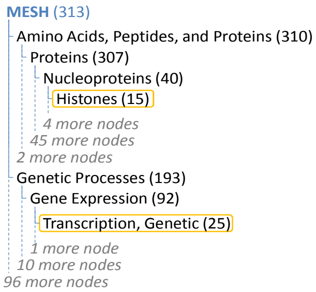
General Terms

Algorithms, Design and Human Factors.

1. INTRODUCTION

The last decade has been marked by unprecedented growth in both the production of biomedical data and the amount of published literature discussing it. The MEDLINE database, on which the PubMed search engine operates, contains over 18 million citations, and the database is growing at the rate of 500,000 new citations each year [7]. Keyword search queries on these databases return a large results set from which only a small portion is relevant for the user. Many solutions have been proposed to address this problem – commonly referred to as *information-overload* [2,3]. These approaches can be broadly classified into two classes: ranking and categorization, which can also be combined.

BioNav belongs primarily to the categorization class, which is ideal for this domain given the rich concept hierarchies available for biomedical data, such as MeSH [5]. Each citation in MEDLINE is associated with several MeSH concepts in two ways: (i) by being explicitly annotated with them, and (ii) by mentioning them in their text. Since these associations are provided by PubMed, a relatively straightforward interface to navigate the query result would first attach the citations to the corresponding MeSH concept nodes and then let the user navigate the concept hierarchy.



MESH (313)
Amino Acids, Peptides, and Proteins (310)
Proteins (307)
Nucleoproteins (40)
Histones (15)
4 more nodes
45 more nodes
2 more nodes
Genetic Processes (193)
Gene Expression (92)
Transcription, Genetic (25)
1 more node
10 more nodes
96 more nodes

Figure 1. Static Navigation on the MeSH Concept Hierarchy

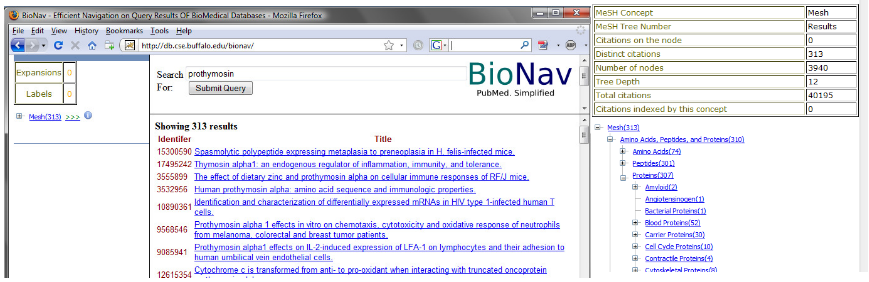
Figure 1 displays a snapshot of such an interface where shown next to each node label is the count of distinct citations in the subtree rooted at that node. For this example, we assume that the user queries MEDLINE for the nucleoprotein “prothymosin” and his personal interests are reflected in the two indicated concepts, corresponding to two independent lines of research related to prothymosin. A typical navigation starts with revealing the children of the root ranked by their citation count, and is continued with expanding one or more of them, revealing their ranked children and so on. Further, the user may click on a concept and inspect the attached citations. A similar interface and navigation method is used by GoPubMed [6] and e-commerce sites, such as Amazon and eBay.

The above *static* navigation method –same for every query result– is problematic when the MeSH hierarchy (or one with similar properties) is used for categorization for the following reasons:

- The massive size of the MeSH hierarchy (with 48,441 concept nodes) makes it challenging for the users to effectively navigate to the desired concepts and browse the associated citations.
- A substantial number of *duplicate* citations are introduced in the navigation tree of Figure 1, since each one of the 313 d*istinct c*itations is associated with several concepts. Specifically, the total count of citations in Figure 1 is 40,195.

BioNav, first proposed in [1], introduces a *dynamic* navigation method that depends on the particular query result at hand. The query results are attached to the corresponding MeSH concept nodes as in Figure 1, but then the navigation proceeds differently. The key action on the interface is the *expansion* of a node that selectively reveals a ranked list of descendant (not necessarily children) concepts, instead of simply showing all its children.

SIGMOD'09, June 29–July 2, 2009, Providence, RI, USA.
Copyright 2009 ACM 1-58113-000-0/00/0004…$5.00.



BioNav - Efficient Navigation on Query Results OF BioMedical Databases - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://db.cse.buffalo.edu/bionav/

Expansions	0
Labels	0

Mesh(313) >>>

Search For: prothymosin

Submit Query

BioNav
PubMed. Simplified

Showing 313 results

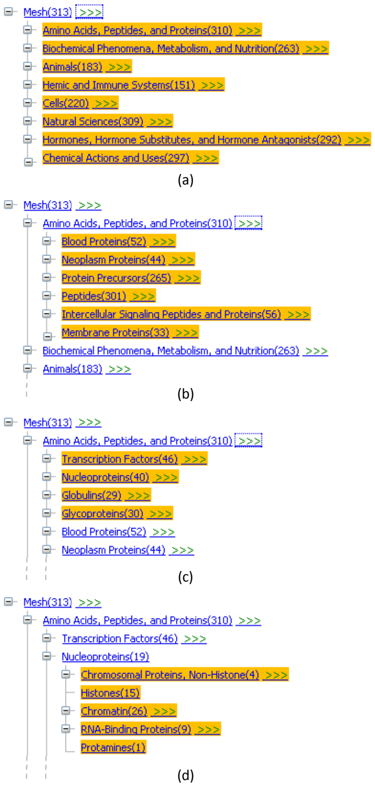
Identifer	Title
15300590	Spasmolytic polypeptide expressing metaplasia to preneoplasia in H. felis-infected mice.
17495242	Thymosin alpha1: an endogenous regulator of inflammation, immunity, and tolerance.
3555899	The effect of dietary zinc and prothymosin alpha on cellular immune responses of RF/J mice.
3532956	Human prothymosin alpha: amino acid sequence and immunologic properties.
10890361	Identification and characterization of differentially expressed mRNAs in HIV type 1-infected human T cells.
9568546	Prothymosin alpha 1 effects in vitro on chemotaxis, cytotoxicity and oxidative response of neutrophils from melanoma, colorectal and breast tumor patients.
9085941	Prothymosin alpha1 effects on IL-2-induced expression of LFA-1 on lymphocytes and their adhesion to human umbilical vein endothelial cells.
12615354	Cytochrome c is transformed from anti- to pro-oxidant when interacting with truncated oncoprotein

MeSH Concept	Mesh
MeSH Tree Number	Results
Citations on the node	0
Distinct citations	313
Number of nodes	3940
Tree Depth	12
Total citations	40195
Citations indexed by this concept	0

- Mesh(313)
  - Amino Acids, Peptides, and Proteins(310)
    - Amino Acids(74)
    - Peptides(301)
    - Proteins(307)
      - Amyloid(2)
      - Angiotensinogen(1)
      - Bacterial Proteins(1)
      - Blood Proteins(52)
      - Carrier Proteins(30)
      - Cell Cycle Proteins(10)
      - Contractile Proteins(4)
      - Cytoskeletal Proteins(8)

Figure 2. BioNav Interface after Querying for "prothymosin" and its Associated Subtree Information Window

BioNav Interface. Figure 2 shows the state of the BioNav interface after querying for "prothymosin". The root of the MeSH tree can be seen on the left pane. The right pane shows the results under the current node of the navigation tree of the left pane. The user can also view more information about a subtree rooted at a given concept node by clicking on the ⓘ icons that appear next to each concept label. The table of the pop-up window in Figure 2 shows various characteristics of the current subtree, including the fact that the 313 citations in the query result are spread over 3940 concept nodes.



- Mesh(313) >>>
  - Amino Acids, Peptides, and Proteins(310) >>>
  - Biochemical Phenomena, Metabolism, and Nutrition(263) >>>
  - Animals(183) >>>
  - Hemic and Immune Systems(151) >>>
  - Cells(220) >>>
  - Natural Sciences(309) >>>
  - Hormones, Hormone Substitutes, and Hormone Antagonists(292) >>>
  - Chemical Actions and Uses(297) >>>

(a)

- Mesh(313) >>>
  - Amino Acids, Peptides, and Proteins(310) >>>
    - Blood Proteins(52) >>>
    - Neoplasm Proteins(44) >>>
    - Protein Precursors(265) >>>
    - Peptides(301) >>>
    - Intercellular Signaling Peptides and Proteins(56) >>>
    - Membrane Proteins(33) >>>
  - Biochemical Phenomena, Metabolism, and Nutrition(263) >>>
  - Animals(183) >>>

(b)

- Mesh(313) >>>
  - Amino Acids, Peptides, and Proteins(310) >>>
    - Transcription Factors(46) >>>
    - Nucleoproteins(40) >>>
    - Globulins(29) >>>
    - Glycoproteins(30) >>>
    - Blood Proteins(52) >>>
    - Neoplasm Proteins(44) >>>

(c)

- Mesh(313) >>>
  - Amino Acids, Peptides, and Proteins(310) >>>
    - Transcription Factors(46) >>>
    - Nucleoproteins(19)
      - Chromosomal Proteins, Non-Histone(4) >>>
      - Histones(15)
      - Chromatin(26) >>>
      - RNA-Binding Proteins(9) >>>
      - Protamines(1)

(d)

Figure 3 BioNav Navigations

BioNav Navigation. Figure 3a shows the initial expansion of the root node where only 8 (highlighted) descendants are revealed compared to 98 children shown in Figure 1. The concepts are ranked by their relevance to the user query and the number of them revealed depends on the characteristics of the query results. Next, assuming the user is interested in the "Amino Acids..." node and judging that the 310 attached citations is still a big number, she expands it by clicking on the ">>>" hyperlink next to it in Figure 3b. The user inspects the 6 concepts revealed and decides that she is not interested in any of them. Hence, she expands the "Amino Acids..." node one more time in Figure 3c, revealing 4 additional concepts. Note that "Nucleoproteins" is an example of a descendant node being revealed, since its parent node "Proteins" (shown in Figure 1) is not revealed in Figure 3c. In Figure 3d, the user expands the "Nucleoproteins" node and reveals "Histones", one of the two key concepts for the query. Note that to reach "Histones" using the BioNav navigation method only 23 concepts are revealed, after 4 node expansions, compared to 152 concepts, also after 4 expansions, with the static navigation method of Figure 1.

2. SYSTEM OVERVIEW

The MeSH concept hierarchy is a labeled tree [5], where the label of a child concept node is more specific than the one of its parent. Once the user issues a keyword query, PubMed–BioNav uses the Entrez Programming Utilities (eUtils) [4] –returns a list of citations, each associated with several MeSH concepts. BioNav constructs a *navigation tree* by attaching to each concept node of the MeSH concept hierarchy a list of its associated citations and removing all nodes with no citations, while preserving the ancestor-descendent relationship. The *navigation tree* $T(V,E,r)$ is the maximum embedding of an initial navigation tree $T_I(V_I,E_I,r)$ such that no node $n \in V$ is labeled with an empty results list $L(n)$, excluding the root (in order to maintain the tree structure and avoid the creation of a forest).

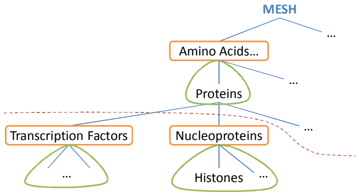


Figure 4. Navigation Tree, EdgeCut and Component Subtrees

We model a node expansion at a given navigation step as an *EdgeCut* in the navigation tree. In Figure 4, the dashed line illustrates the EdgeCut corresponding to the expansion of the node "Amino Acids…". This expansion reveals the highlighted concepts of Figure 4, which include a subset of the highlighted concepts in Figure 3c. The EdgeCut consists of the edges ("Proteins", "Transcription Factors") and ("Proteins", "Nucleoproteins"). A valid EdgeCut of a tree $T(V,E,r)$ is an EdgeCut $C \subseteq E$ such that no two edges in C appear in a path from the root to a leaf node. We only consider valid EdgeCuts, because invalid EdgeCuts lead to unintuitive navigations.

Component Subtrees. An EdgeCut causes the creation of two types of *component subtrees*, a single *upper* and possibly multiple *lower*. Figure 4 shows two lower component subtrees, rooted at "Transcription Factors", and "Nucleoproteins", and an upper component subtree comprising of the node being expanded "Amino Acids…" and all nodes not in any of the lower component subtrees.

2.1 Navigation and Cost Model

BioNav initiates a navigation by constructing the initial results tree and displaying its root to the user. Subsequently, the user navigates the tree by performing one of the following actions on a given component subtree *I(n)* rooted at concept node *n*:

1. **EXPAND** *I(n)*: The user clicks on the ">>>" hyperlink next to node *n* and causes an *EdgeCut(I(n))* operation to be performed on it, thus revealing a new set of concept nodes from the set *I(n)*.
2. **SHOWRESULTS** *I(n)*: By performing this action, the user sees the results list *L(I(n))* of citations attached to the component subtree *I(n)*.
3. **IGNORE** *I(n)*: The user examines the label of concept node *n*, ignores it as unimportant and moves on to the next revealed concept.

This navigation process continues until the user finds *all* the citations she is interested in. The cost of a navigation is computed as follows: We assign (i) cost of 1 to each newly revealed concept node that the user examines after an EXPAND action, (ii) a cost of *B* (determined empirically) to each EXPAND action the user executes, and (iii) cost of 1 to each citation displayed after a SHOWRESULTS action. BioNav estimates the navigation cost by taking in to account the probability that the user will execute an EXPLORE or SHOWRESULTS action at each step of the navigation. The EXPLORE probability is proportional to the number of unique results in the corresponding component subtree, whereas *normalized* entropy of the component subtree is used as the SHOWRESULTS probability.

2.2 System Architecture

The BioNav system architecture is shown in Figure 5 and consists of two parts. The off-line components populate the BioNav database with the MeSH concept hierarchy and the associations of the MEDLINE citations with MeSH concepts to decrease the on-line response time. The on-line components support BioNav's web interface and the EXPAND/SHOWRESULTS user actions.

Off-Line Pre-Processing The BioNav database is first populated with the MeSH hierarchy. Next, the associations of MEDLINE citations and MeSH concepts are populated by issuing a query on PubMed for each concept *c*. For each citation t_i returned by the query, we add the association $<c, t_i>$ in our database.

On-Line Operation Upon receiving a keyword query from the user, BioNav executes the same query against the MEDLINE database and retrieves only the IDs (PubMed Identifiers) of the citations in the query result using the ESearch utility [4] and constructs the navigation tree by retrieving the MeSH concepts associated with each citation in the query result. Initially, the root of this navigation tree is shown to the user. Subsequently, when she requests an EXPAND action on the root, the Navigation Subsystem executes a heuristic algorithm to compute the best *EdgeCut* and the roots of the resulting component subtrees are visualized on the web-interface.

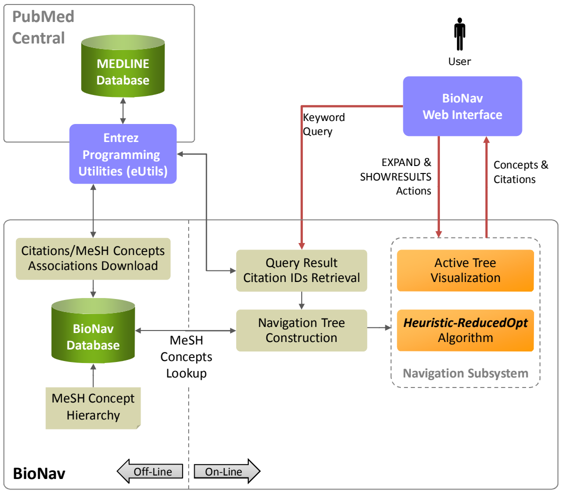

PubMed Central
MEDLINE Database
Entrez Programming Utilities (eUtils)
User
BioNav Web Interface
Keyword Query
EXPAND & SHOWRESULTS Actions
Concepts & Citations
Citations/MeSH Concepts Associations Download
Query Result Citation IDs Retrieval
Active Tree Visualization
BioNav Database
MeSH Concepts Lookup
Navigation Tree Construction
Heuristic-ReducedOpt Algorithm
Navigation Subsystem
MeSH Concept Hierarchy
BioNav
Off-Line
On-Line

Figure 5. BioNav System Architecture

3. DEMONSTRATION PLAN

In this demonstration we will show the effectiveness of BioNav in reducing the overall navigation cost as compared to *static* navigation. Specifically, we show that BioNav, (a) shows significantly fewer number of concept labels as compared to static navigation, (b) selects *relevant* descendents, as opposed to children, to be revealed at each expansion step and (c) number of EXPAND actions is comparable to that of static navigation. The BioNav system is available at http://db.cse.buffalo.edu/bionav and the static navigation is available through the information pop up window (right side of Figure 2) for the users to compare in parallel. Further, we provide a set of sample queries at BioNav website for users with limited biomedical background. For performance reasons we disallow queries that return more than 3000 results and the user is asked to refine such a query.

Acknowledgements: Vagelis Hristidis was partially supported by NSF grants IIS-0811922 and IIS-0534530.

REFERENCES

[1] A. Kashyap, V. Hristidis, M. Petropoulos and S. Tavoulari: *BioNav: Effective Navigation on Query Results of Biomedical Databases*. ICDE 2009 (to appear).

[2] K. Chakrabarti, S. Chaudhuri and S.W. Hwang: *Automatic Categorization of Query Results. SIG*MOD Conference 2004: 755-766.

[3] Z. Chen and T. Li: *Addressing Diverse User Preferences in SQL-Query-Result Navigation*. SIGMOD Conference 2007: 641-652.

[4] Entrez Programming Utilities. http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

[5] Medical Subject Headings(MeSH). http://nlm.nih.gov/mesh/

[6] Transinsight GmbH – GoPubMed. http://gopubmed.org

[7] J.A. Mitchell, A.R. Aronson and J.G. Mork: *Gene Indexing: Characterization and Analysis of NLM's GeneRIFs*. In Proceedings of the AMIA Symposium, 8th–12th November, Washington, DC, pp. 460–464, 2003.