Reprinted from:

2005 3rd IEEE International Conference on Industrial Informatics (INDIN) Perth, Australia 10-12 August 2005

IEEE Catalog Number ISBN 05EX1057
ISBN 0-7803-9094-6

Ontology-based Knowledge Representation for Protein Data

Amandeep S. Sidhu[1], Member, IEEE, Tharam S. Dillon[2], Fellow, IEEE, Elizabeth Chang[3], Member, IEEE, and Baldev S. Sidhu[4]
[1,2]Faculty of Information Technology, University of Technology Sydney, Australia, e-mail: (asidhu, tharam)@it.uts.edu.au
[3]School of Information Systems, Curtin University of Technology Perth, Australia, e-mail: Elizabeth.Chang@cbs.curtin.edu.au
[4]Punjab State Education Department, India, e-mail: bsidhu@biomap.org

Abstract — The advances in information and communication technologies coupled with increased knowledge about genes and proteins have opened new perspectives for study of protein complexes. There is a growing need to integrate the knowledge about various protein complexes for effective disease prevention mechanisms, individualized medicines and treatments and other accepts of healthcare. In this paper we propose a Protein Ontology that will handle the following computational challenges in the area Proteomics and systems biology in general: (1) It will provide more accurate interpretations and associations as conclusions are based on Data and Semantics. (2) It will make it possible to study relationships among proteins, protein folding, behaviour of protein under various environments, and most importantly cellular function of protein. This Protein Ontology is a unified terminology description integrating various protein database schemas and will provide a easier way to predict and understand proteins.

***Index Terms* — Protein Ontology, Protein Informatics, Biomedical Ontologies, Biomedical Systems, Data Integration**

I. Introduction

Bioinformatics is a study of two important flows Molecular Biology. First is the flow of genetic information, depicted in Molecular Biology by Central Dogma. Second is the flow of experimental information from observed biological phenomena, modelling these phenomena and then test these models on real data [1]. Bioinformatics efforts can be traced back to the early applications of computers to molecular biology for: (1) Graphical Rendering of the Molecular Structures [2], (2) The Molecular Sequence Databases [3], and (3) Three dimensional structure information [4]. Advances in computing power and ease of use have increased the use of information technology methodologies in Life Sciences. The life sciences activities are commonly categorized as computational biology (such as proteomics and genomics) and as database development and exploitation of biological data banks of macromolecules – Proteins, RNA and DNA. Heterogeneity among various information sources is a major issue when extracting value from various distributed biological resources available. Biological Knowledge has to be comprised of multiple sources when answering queries. Information integration from multiple protein databases like PDB [4, 5, 6, 7], SWISS-PROT [8, 9], and PIR [10] needs multi database query formations when answering user queries. Multiple databases may cover same data, but there focus might be different. The SWISS-PROT [8, 9] database provides Protein Sequence Information, PDB [4, 5, 6, 7] database provides Protein Structure Information, and PIR [10] is mainly for cross referencing and linking various protein references. To answer data from these databases the data needs to be combined and represented in consistent fashion. While these data formats are useful for knowledge extraction on per – protein basis, they do not allow for efficient integration of all proteomics data relevant to a particular experiment, and they are certainly not provide all the knowledge needed for protein complexes. It is therefore quite difficult to create self-consistent models, and evaluate the compatibility of individual protein family data sets with these models.

We propose a Protein Ontology, showing the value of structured representations of proteomics data. The creation of a Protein Ontology that provides a comprehensive understanding of Protein Complex Mechanisms will help in the understanding of Cellular Mechanisms. Diverse types of data formats taken from different protein data sources are represented using a set of type definitions within this protein ontology, and these data are linked to each other with numerous connections. Not only does this structured representation allow easier data retrieval to users, but it also facilitates automated data mining by computer programs. In this paper, we describe the design principles behind the proposed Protein Ontology, illustrate how we have represented certain key data types to represent protein data, and describe the resulting Protein Ontology as it is currently publicly available.

II. Earlier protein Data integration methods

"According to a survey [11] there are at least 335 data sources in 2002 beginning with 6 – 10 sources of similar type for protein, pathway, publication, gene expression data, etc." [12]. Data Integration [13, 14], which is a combination of Artificial Intelligence and Databases, has approaches to provide access to multiple heterogeneous data sources in a uniform fashion [14] such that data model is virtually accessible to the end user. Protein Data Integration approaches at the moment considers data sources as repositories of data, but not as applications; which in turn may embody complex interactions with other sources. Current approaches also do not provide methods both for Generic Protein Mapping Representation, depicting interactions in data it describes and for interfacing existing data. A variety of approaches exist for integrated access to heterogeneous protein data sources. In the link-driven federation approaches [15, 16, 17] the user can switch between sources

using system provided links. These systems do not have any transparency for users. In view integration [18, 19, 20] a virtual global schema in common data model is created using source description. Queries on common data model are then automatically reformulated to source level queries. A variation of view integration is the warehousing approach [19] where instantiation of global schema is created, i.e. all data is locally stored and maintained for integrated access. Both view and warehouse integration provides schema integration, but not the data source transparency. The need for data source transparency leads us to consider semantic integration [21, 22]**.** Karp [23, 24] has identified the several approaches that have been proposed and implemented by bioinformatics researchers and proposed a strategy for data interoperation. For understanding processes like Protein Synthesis usually both data and its biological context determines the complete meaning (or semantics) of the item.

Our Protein Ontology [25, 26, 27, 28, 29, 30] defines a common structured vocabulary for researchers who need to share knowledge in proteomics domain. It includes concepts (type definitions), which are data descriptors for proteomics data and the relations among these concepts. The Key features of Protein Ontology are (1) a hierarchical classification of concepts (classes) from general to specific; (2) a list of attributes for each class; and (3) a set of relations between classes to link concepts in ontology in more complicated ways then implied by underlying hierarchy. The Concepts have instances, which represent concrete examples of more abstract classes found in internal part of the hierarchy. Each attribute of an instance may have a corresponding value, whereas classes only specify that the attribute exists. Ontology & Knowledge Base approaches similar to the proposed approach like Gene Ontology [31, 32, 33] and RiboWEB [34, 35, 36].

III. NEED FOR A PROTEIN ONTOLOGY

The motivations behind proposing a Protein Ontology Model are: (1) Efforts in building consensus on data format using semantics inherent in various protein databases. This can be attained by creation of a data representation standard that defines physiological models at atomic and molecular level. The ability of the protein ontology to define such models for protein molecules and then the ability to model single cells will provide basic data necessary to model entire organs and organisms automatically. (2) Biologists in different specialties tend to use different languages for description of same data. They have their particular theories and models for their own data collection of the domain they are working on. Protein Ontology is a unified data description model that covers all of the working domains. (3) The terms used to describe biomolecular data has different granularity depending on the level at which the abstractions or concepts in the domain and have different scope. Therefore, terms used in different contexts have different meaning. Defining Protein Ontology brings a consistent structured terminology for all biomolecular data. (4) For various Protein Databases there are different data models. It is the interfaces that provide interoperation and data exchange, but there are no interfaces to recognize integration and interactions between various data models and to exchange Data and Meta Data between them in consistent format. Protein Ontology does the Data Integration and Data Exchange between various existing Protein Data Models.

IV. PROTEIN ONTOLOGY CONCEPTS

The Main Class of Protein Ontology is *ProteinOntology*. For each Protein that is entered into the knowledge base of protein ontology, submission information is entered into *ProteinOntology* Class. *ProteinOntologyID* has format like "PO0000000007".

A. Generic Classes

There are six subclasses of ProteinOntology that are used to define complex concepts in other classes of ProteinOntology: *Residues, Chains, Atoms, AtomicBind, Bind, and SiteGroup*. Concepts from these subclasses are referenced in various other Protein Ontology Classes for definition of Class Specific Concepts. Details and Properties of Residues in a Protein Sequence are defined by instances of *Residues* Class. Instances of Chains of Residues are defined in *Chains* Class. All the Three Dimensional Structure Data of Protein Atoms is represented as instances of *Atoms* Class. Defining Chains, Residues and Atoms as individual classes has the benefit that any special properties or changes affecting a particular chain, residue and ATOM can be easily added. Data about binding atoms in Chemical Bonds like Hydrogen Bond, Residue Links, and Salt Bridges is entered into ontology as an instance of *AtomicBind* Class. Similarly the data about binding residues in Chemical Bonds like Disulphide Bonds and CIS Peptides is entered into ontology as an instance of *Bind* Class. All data related to site groups of the active binding sites of Proteins is defined as instances of *SiteGroup* Class.

B. ProteinComplex Class

The Root Class for definition of a Protein Complex in the Protein Ontology is *ProteinComplex*. There are six main subclasses within *ProteinComplex* class: *Entry*, *Structure*, *StructuralDomains*, *FunctionalDomains*, *ChemicalBonds*, and *Constraints.*

C. Entry Class

Entry specifies the details of a Protein or a Protein Complex that is entered into the knowledge base of protein ontology. Protein Entry Details are entered into *Entry* as instances of *SourceDatabaseID*, *SourceDatabaseName* and *SubmissionDate.* These attributes describe the entry in the original protein data source from where it was taken. *Entry* has three subclasses: *Description*, *Molecule* and *Reference*. *Description* has data about title of the entry, authors of the entry, experiment that produced the entry and keywords describing the entry. The second subclass of *Entry* is *Molecule* which is simply any chemically distinct molecule or compound in a protein complex. *MoleculeID* just uniquely

identifies a Molecule. *MoleculeName* is the Chemical Name of the Molecule. *Molecule Chain* refers the Chain Description. *BiologicalUnit* Instance describes the larger biological unit of which molecule is a part. *Engineered* identifies whether the molecule is engineered using Recombinant Technology or Chemical Synthesis. A specific domain or region of the molecule is defined using *Fragment*. Mutated Molecules of the Protein have *Mutations* Information. Details about various mutations are described in *GeneticDefects* Class. List of Synonyms for Molecule Name are in *Synonyms*. *OtherDetails* describes any other information. *Reference* subclass lists the various literature citations of the protein or protein complex described by the instances of: *CitationReference*, *CitationPublication*, *CitationTitle, CitationAuthors, CitationEditors*, *and CitationReferenceNumbers*.

D. Structure Class

Structure has Protein Sequence and Structure data for a Protein Entry. *Structure* has two subclasses: *ATOMSequence* and *UnitCell*. *ATOMSequence* consists of various chains of residue sequences present in the Protein. Each Chain is a sequence of singular residues. Each Residue or Chain may have distinct properties and functionality. Each Residue has a number of atoms linked to it, that define the three dimensional structure of Protein. Here in Structure, Residue is a sub property of Chain and ATOM is the sub property of Residue. The Containment relationship: *Chain < Residue < ATOM* still represents the hierarchy need for protein sequence and structure data, but also preserves individuality of the components. Data from Protein Crystallography like a, b, c, alpha, beta, gamma, z, and SpaceGroup are entered in *UnitCell*.

E. StructuralDomains Class

Structural Folds and Domains defining Secondary Structures of Proteins are defined in *StructuralDomains*. *SuperFamily* and *Family* Instances of StructuralDomains are used for identifying the Protein Family. The subclasses of StructuralDomains are *Helices*, *Sheets*, and *OtherFolds*. *Helix*, which is a subclass of *Helices*, identifies the helix using *HelixNumber*, *HelixID*, *HelixClass*, and *HelixLength* Instances. *Helix* has a subclass *HelixStructure* gives the detailed composition of the helix in terms of following instances: (1) *Helix Chain:* Chain of Strand (References Chain Details from *Chains* Class), (2) *Helix Initial Residue:* Initial Residue of each Helix (References Residue Details from *Residues* Class), (3) *Helix Initial Residue Sequence Number:* Identifies the Residue Sequence Number of the Intial Residue in the Helix, (4) *Helix End Residue:* End Residue of each Helix (References Residue Details from *Residues* Class) and (5) *Helix End Residue Sequence Number:* Identifies the Residue Sequence Number of the End Residue in the Helix.

Second Subclass of *StructuralDomains*, *Sheets* contains all the data about sheets present protein using its subclass *Sheet*. *Sheet* identifies individual sheets using *SheetID* and *NumberStrands* which represents the Number of Strands in the Sheet. *Sheet* has subclass called *Strands* that lists strands starting with one edge of the sheet and continuing to the spatial adjacent strand in terms of following: (1) *Strand Number:* Strand Number for each strand within the Sheet, (2) *Strand Chain*: Chain of Strand (References Chain Details from *Chains* Class), (3) *Strand Initial Residue*: Initial Residue of each Strand (References Residue Details from *Residues* Class), (4) *Strand Intial Residue Sequence Number*: Identifies the Residue Sequence Number of the Intial Residue in the Strand, (5) *Strand End Residue*: Initial Residue of each Strand (References Residue Details from *Residues* Class), (6) *Strand End Residue Sequence Number*: Identifies the Residue Sequence Number of the End Residue in the Strand, (7) *Strand Sense*: Sense of Strand with respect to the previous strand in the sheet, (8) *Strand Current ATOM*: ATOM in Current Strand (References Atom Details from Atoms Class), (9) *Strand Current Residue*: Residue in Current Strand (References Residue Details from *Residues* Class), (10) *Strand Current Residue Sequence Number*: Identifies the Residue Sequence Number of the Current Residue in the Strand, (11) *Strand Previous ATOM*: ATOM in Previous Strand (References Atom Details from Atoms Class), (12) *Strand Previous Residue*: Residue in Previous Strand (References Residue Details from *Residues* Class), and (13) *Strand Previous Residue Sequence Number*: Identifies the Residue Sequence Number of the Previous Residue in the Strand.

Third Subclass of *StructuralDomains*, *OtherFolds* consists of loosely coupled folds. One of the most common folds of this category is short loop turns which connect other secondary structure segments, described in *Turn* subclass of *OtherFolds*. A Turn is identified by Instances of *TurnNumber* and *TurnID*. Turn has a subclass TurnStructure that defines the detailed composition of a Turn in terms of following instances: (1) *Turn Chain:* Chain of Turn (References Chain Details from *Chains* Class), (2) *Turn Initial Residue:* Initial Residue of each Turn (References Residue Details from *Residues* Class), (3) *Turn Initial Residue Sequence Number:* Identifies the Residue Sequence Number of the Intial Residue in the Turn, (4) *Turn End Residue:* End Residue of each Turn (References Residue Details from *Residues* Class), and (5) *Turn End Residue Sequence Number:* Identifies the Residue Sequence Number of the End Residue in the Turn.

F. FunctionalDomains Class

Protein Ontology has the first Functional Domain Classification Model defined using *FunctionalDomains* Class using: (1) Data about Cellular and Organism Source in *SourceCell* subclass and (2) Data about Biological Functions of Protein in *BiologicalFunction* subclass and (3) Data about Active Binding Sites in Proteins in *ActiveBindingSites* subclass. *SourceCell* specifies biological or chemical source of each biological molecule (Defined by Molecule Class) in the Protein. *SourceMoleculeID* uniquely identifies each biological molecule. The property is equivalent to *MoleculeID* property in *Molecule* Class. *SourceSynthetic* indicates a chemically-synthesized source. *Sour-*

ceMoleculeFragment specifies a domain or fragment of the biological molecule. *OrganismScientific* and *Organisim-Common* are the Scientific Name and Common Name of the Organism respectively. *Strain* describes the Strain of the Source and *Variant* identifies the variant. *CellLine* Identifies the line of cells used in the experiment. *Organ* defines an organized group of tissues for a specific function. *Tissue* in itself is an organized group of cells with common function. *Cell* identifies a particular cell type and *Organelle* is an organized structure within a cell. *Secretion* identifies the secretion such as saliva or venom, from which molecule was isolated. *CellularLocation* identifies the location inside or outside the cell. *Plasmid* describes the plasmid containing the gene and *Gene* gives detailed description of the gene. *ExpressionSystem* is the system used to express recombinant macromolecules. *SourceOtherDetails* is used to enter any additional data about the source. Biological Functions of the Protein Complex are described in *BiologicalFunction*. BiologicalFunction has two children, *PhysiologicalFunction* and *PathologicalFunction*, and each of these has several children and grand children. The third subclass of FunctionalDomains is *ActiveBindingSites* that has details about active binding sites in the Protein. Active Binding Sites are represented in our ontology as a collection of various Site Groups, defined in *SiteGroup* class. SiteGroup has details about each of the Residues and Chain that form the Binding Site. There can be a maximum of seven Site Groups in the ontology.

G. ChemicalBonds Class

Chemical Bonds in a Protein are defined using *ChemicalBonds* class. Various Chemical Bonds defined in ontology by respective subclasses are: *DisulphideBond, CISPeptide, HydrogenBond, ResidueLink, and SaltBridge*. As said earlier, the binding atoms in Chemical Bonds like Hydrogen Bond, Residue Links, and Salt Bridges is entered into ontology as an instance of *AtomicBind* Class. Similarly the data about binding residues in Chemical Bonds like Disulphide Bonds and CIS Peptides is entered into ontology as an instance of *Bind* Class. The respective classes defining specific chemical bonds use *Bind* to define participating binding *Residues* and *AtomicBind* to define participating binding *Atoms*.

H. Constraints Class

Last subclass of Protein Complex describes the constraints that affect final protein conformation. The constraints described in Protein Ontology at the moment are: (1) Monogenetic and Polygenetic defects present in genes that are present in molecules making proteins in *GeneDefects* subclass, (2) Hydrophobicity properties in *Hydrophobicity* Class, and (3) Modification in Residue Sequences due to Chemical Environment and Mutations are entered in *ModifiedResidue* Class. Data in *GeneDefects* class is entered as instances of GeneDefects Class and is normally taken from OMIM database [37] or literature.

V. IMPLEMENTATION

The Ontology is at: **http://www.proteinontology.info/**. The Class Diagram and UML Diagrams for Protein Ontology are available at the website. The Ontology Currently contains 91 *concepts* or classes, 248 *attributes* or properties and 99 instances. Protein Ontology describes the concepts of interest in protein complex mechanisms and proteomics process. The protein data source attributes are mapped to these defined concepts.

VI. SUMMARY

The Protein Ontology is an ontology based integration of heterogenous protein and biological data sources. Protein Ontology converts the enormous amounts of data collected by geneticists and molecular biologists into information that scientists, physicians and other health care professionals and researchers can use to easily understand the mapping of relationships inside protein molecules, interaction between two protein molecules and interactions between protein and other macromolecules at cellular level. Protein Ontology also helps to codify proteomics data for analysis by researchers. Protein Ontology contains templates for all kinds of protein data that is need to understand proteins, their functionality and the cellular processes. Previously there is not such integrated and structured data representation format available. Most of the values for many attributes unlike earlier methods are not simply text strings, but has been entered into the ontology as instances of other concepts, defined by Generic Classes.

In future, we will provide more instances to validate the Protein Ontology. In long term, we would like to create data input software that can be used to transfer data from Protein Databases into Protein Ontology Knowledge Base.

VII. REFERENCES

[1] Altman, R. B. (1998) "Bioinformatics in support of molecular medicine." AMIA Annual Symposium, Orlando, FL.

[2] Langridge, R. (1974). "Interactive three-dimensional computer graphics in molecular biology." Fed Proc 1974 33(12): 2332-5.

[3] Smith, T. F. (1990). "The history of genetic sequence databases." Genomics 6(4): 701-7.

[4] Bernstein, F. C., T. F. Koetzle, et al. (1977). "The Protein Data Bank: a computer-based archival file for macromolecular structures." Journal of Molecular Biology 112(3): 535-42.

[5] Weissiga, H. and P. E. Bourne (2002). "Protein structure resources." Biological Crystallography D58: 908-915.

[6] Westbrook, J., Z. Feng, et al. (2002). "The Protein Data Bank: unifying the archive." Nucleic Acid Research 30(1): 245-248.

[7] Bhat, T. N., P. E. Bourne, et al. (2001). "The PDB data uniformity project." Nucleic Acid Research 29(1): 214-218.

[8] Bairoch, A. and R. Apweiler (1997). "The SWISS-PROT protein sequence data bank and its supplement TrEMBL." Nucleic Acids Research 25(1): 31–36.

[9] Bairoch, A., P. Bucher, et al. (1997). "The PROSITE database, its status in 1997." Nucleic Acid Research 25(1): 217–221.

[10] George, D. G., R. J. Dodson, et al. (1997). "The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database." Nucleic Acids Research 25(1): 24–27.

[11] Baxevanis, A. (2002). "The Molecular Biology Database Collection: 2002 update." Nucleic Acids Research 30(1).

[12] Srivastava, B. (2002). "Data Integration Approaches in Bioinformatics - A Tutorial Survey." Delhi, IBM Research Lab: 18.

[13] Langridge, R. (1974). "Interactive three-dimensional computer graphics in molecular biology." Fed Proc 1974 33(12): 2332-5.

[14] Levy, A. (1998). "Combining Artificial Intelligence and Databases for Data Integration."
[15] Etzold, T. and P. Argos (1993). "SRS: An Indexing and Retrieval Tool for Flat File Data Libraries." Computer Application of Biosciences 9: 49-57.
[16] Fujibuchi, W., S. Goto, et al. (1998). "DBGET/LinkDB: an Integrated Database Retrieval System." Pacific Symposis for Biocomputing.
[17] Rebhan, M., V. Chalifa-Caspi, et al. (1997). "GeneCards: encyclopedia for Genes, Proteins, and Diseases." Rehovot, Israel, Weizmann Institute of Science, Bioinformatics Unit and Genome Center.
[18] Buneman, P., S. Davidson, et al. (1995). "A Data Transformation System for Biological Data Sources. VLDB."
[19] Davidson, S., J. Crabtree, et al. (2001). "K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources." IBM Systems Journal.
[20] Hass, L., P. Schwarz, et al. (2001). "DiscoveryLink: A System for integrated access to life sciences data sources." IBM Systems Journal 40(2).
[21] Adak, S., V. Batra, et al. (2002). "Bioinformatics for Microarrays."
[22] Goble, C., R. Stevens, et al. (2001). "A Transparent Access to Multiple Bioinformatics Information Sources." IBM Systems Journal 40(2): 532-551.
[23] Karp, R. M. (2003). Keynote Address: The Role of Algorithmic Research in Computational Genomics. IEEE Computational Systems Bioinformatics (CSB'03).
[24] Karp, P. D. (2000). "An Ontology for Biological Function Based on Molecular Interactions." Bioinformatics 16(2).
[25] Sidhu, A.S., T.S. Dillon et al. (2005). "Protein Ontology: Vocabulary for Protein Data." IEEE ICITA 2005, Sydney, Australia.
[26] Sidhu, A. S., T. S. Dillon, et al. (2005). "The Protein Ontology Project: Structured Vocabularies for Proteins." Data Mining 2005, Greece, Wessex Institute of Technology (WIT), UK.
[27] Sidhu, A. S., T. S. Dillon, et al. (2004). Making of Protein Ontology. 2nd Australian and Medical Research Congress 2004 (Invited Speaker), Sydney, National Heath and Medical Research Council.
[28] Sidhu, A. S., T. S. Dillon, et al. (2004). Protein Knowledge Base: Making of Protein Ontology. HUPO 3rd Annual World Congress 2004, Beijing, China, American Society for Biochemistry and Molecular Biology.
[29] Sidhu, A. S., T. S. Dillon, et al. (2004). A Unified Representation of Protein Structure Databases. Bioconvergence 2004 (Invited Paper), Punjab, India, 145.
[30] Sidhu, A. S., T. S. Dillon, et al. (2004). An XML based semantic protein map. Data Mining 2004, Malaga, Spain, WIT Press.
[31] Harris, M. A., J. Clark, et al. (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Research 32(Database issue): 258-261.
[32] Yeh, I., P. D. Karp, et al. (2003). "Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO)." Bioinformatics 19(2): 241-248.
[33] Ashburner, M., C. A. Ball, et al. (2001). "Creating the Gene Ontology Resource: Design and Implementation." Genome Research 11: 1425–1433.
[34] Altman, R. B., M. Bada, et al. (1999). "RiboWeb: An Ontology-Based System for Collaborative Molecular Biology." IEEE Intelligent Systems (September/October 1999): 68-76.
[35] Bada, M. A. and R. B. Altman (1999). Computational Modeling of Structured Experimental Data. Stanford, CA, Stanford Medical Informatics SMI-1999-0764.
[36] Abernethy, N. F., J. J. Wu, et al. (1999). "Sophia: A Flexible, Web-Based Knowledge Server." IEEE Intelligent Systems (JULY/AUGUST 1999).
[37] McKusick, V. A. (2000). Online Mendelian Inheritance in Man, OMIM. Baltimore, MD, Johns Hopkins University, National Center for Biotechnology Information, and National Library of Medicine.