BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database

Volume 6(2)

The Booly aliasing resource: A database of grouped biological identifiers

Long Hoang Do[1, 2]* & Ethan Bier[1]

[1]Section of Cell and Developmental Biology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0349, USA; [2]Department of Neurosciences, School of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0649; Long Hoang Do - Email: lhdo@ucsd.edu; *Corresponding author

Received February 15, 2011; Accepted February 17, 2011; Published March 26, 2011

Abstract:
Redundancy among sequence identifiers is a recurring problem in bioinformatics. Here, we present a rapid and efficient method of fingerprinting identifiers to ascertain whether two or more aliases are identical. A number of tools and approaches have been developed to resolve differing names for the same genes and proteins, however, these methods each have their own limitations associated with their various goals. We have taken a different approach to the aliasing problem by simplifying the way aliases are stored and curated with the objective of simultaneously achieving speed and flexibility. Our approach (Booly-hashing) is to link identifiers with their corresponding hash keys derived from unique fingerprints such as gene or protein sequences. This tool has proven invaluable for designing a new data integration platform known as Booly, and has wide applicability to situations in which a dedicated efficient aliasing system is required. Compared with other aliasing techniques, Booly-hashing methodology provides 1) reduced run time complexity, 2) increased flexibility (aliasing of other data types, e.g. pharmaceutical drugs), 3) no required assumptions regarding gene clusters or hierarchies, and 4) simplicity in data addition, updating, and maintenance. The new Booly-hashing aliasing model has been incorporated as a central component of the Booly data integration platform we have recently developed and shoud be broadly applicable to other situations in a which an efficient streamlined aliasing systems is required. This aliasing tool and database, which allows users to quickly group the same genes and proteins together can be accessed at: http://booly.ucsd.edu/alias.

Availability: http://booly.ucsd.edu/alias

Background:
A common problem confronted by bioinformaticians is the need to resolve whether two or more identifiers are identical, i.e., are aliases of each other. A number of aliasing services have attempted to resolve the differing naming conventions created by both computational and manual labelling methods (AliasServer, DAVID, HGNC, SEGUID, MagicMatch, NCBI, ENSEMBL) **[1-7]**. These services differ by their technology and solutions with the general strategy of 1) using either in-house generated unique identifiers (NCBI, DAVID, ENSEMBL), or 2) the generation of unique fingerprints (AliasServer, MagicMatch, SEGUID) by way of cryptographic hashing algorithms which digest large arbitrary blocks of data (e.g., sequence) and returns a fixed-size bit string **[8]**. As each of these systems is designed with a specific goal in mind, none of them are optimized for specifically answering the single root question: are two identifiers the same? **(Figure 1a)**

In the course of designing a comprehensive data warehousing and comparison application called Booly **[9]**, we recognized a need for a dedicated aliasing tool designed to efficiently and flexibly resolve alias identities. One of the main tasks of Booly is to mix and match datasets together using combinations of the Boolean operations. A common usage of such a tool is data aggregation between multiple sources (e.g. the aggregation of Gene Ontology data to that of a home brew spreadsheet table for annotation). When identifiers from both datasets are in the same format (e.g., gene symbol), the process of integrating the data can be performed trivially. However, the process of integrating the data becomes more challenging when converting formats is needed, thus becoming an unwieldy aliasing problem. This aliasing problem is compounded when comparing multiple datasets with differing identifier formats. Furthermore, Booly was created to compare content that extends beyond sequence data (e.g., databases of pharmaceutical drugs, human diseases, or other web-based content). With these requirements in mind, we designed an aliasing system (Booly-hashing) that can quickly resolve heterogeneous identifiers from multiple sources while maintaining flexibility to handle aliases from multiple entities.

Booly-hashing is an aliasing database resource that utilizes a 160-bit Secure Hash Algorithm (SHA) hash key to generate unique fingerprints of sequences and their identifiers represented as a 40 character hexadecimal number **(Figure 1a) [10]**. Our streamlined approach requires the storage of only the hash key and its associated identifier. Current aliasing methods utilizing the hashing technology require the source of the identifiers to be known (AliasServer, SEGUID) **[1, 5]**. This limits the ability to find aliases of identifiers from heterogeneous sources. Our simplified technique is more broadly applicable as it allows for conversion to known hash keys for any identifier regardless of originating source. Another aliasing approach is to convert aliases into known, reference identifiers (e.g. RefSeq, Genebank Gene ID) such that one can then easily ascertain whether two identifiers are the same (DAVID) **[3]**. However, this approach is insufficient as some reference databases are incomplete and lack the overlap required to be inclusive of all known sequences and their

identifiers **(Table 1 see Supplementary material)**. In contrast, our aliasing approach utilizes the sequence hash key as a singular point of conversion. Finally, unlike other sequence-related aliasing technologies, we have developed our Booly-hashing infrastructure to accommodate aliases from other sources such as pharmaceutical drugs or keyword aliases. As an example, in the case of pharmaceutical drugs, the unique fingerprint is the chemical formula that remains intact despite multiple branding names. A comparison table of the differences in features among our approach and other aliasing tools can be found in **Figure 1b**. In aggregate, our aliasing method allows one to efficiently and accurately ascertain whether two or more identifiers are aliases of each other. Furthermore, our streamlined approach is flexible and easy to modify and update. We have incorporated this aliasing model as part of a core component in Booly, our data integration platform designed to aid researchers in making new connections leading to novel discoveries in the laboratory. This generalized aliasing system should be of similar utility for development of other comparative tools that also have the simple requirement of rapidly deciding whether two identifiers are the same. Additionally, we have created an online tool that simply takes as input a list of identifiers and groups them accordingly into similar gene or protein sequence clusters.

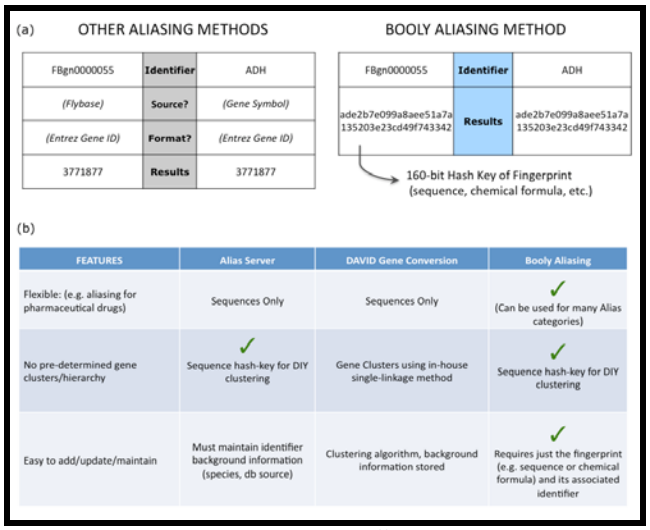

Figure 1: Booly aliasing resource. **(a)** Difference between other aliasing approaches and the Booly-hashing method. The single question we wish to answer efficiently is, whether two identifiers (e.g., FBgn0000055 and *ADH*) are one and the same? Booly-hashing utilizes a 160-bit SHA-1 hash key to generate unique fingerprints of sequences and their identifiers represented as a 40 character hexadecimal number. Identifiers with the same hash-keys are considered as aliases of each other. Other approaches require knowledge of the source of the original identifier or knowledge of a conversion format requiring additional steps that increase complexity and programming **(b)** Comparison of two commonly used aliasing tools in bioinformatics (AliasServer and DAVID Gene Conversion Tool) against the Booly-hashing resource.

Summary:
The process of determining whether two or more identifiers are aliases of each other is a common recurring problem in bioinformatics. To this end, we have created a streamlined aliasing method that utilizes a fingerprint such as a sequence or chemical formula for the purpose of creating unique hash-key identifiers. Our approach affords us a number of advantages over existing aliasing solutions, including a reduction in run time complexity, increased flexibility, flexible alias clusters, and simplicity in addition of new data, updating, and maintenance. In addition to performing well for Booly, these advantages should allow better integration of data containing heterogeneous identifiers leading to new connections and novel discoveries within many fields of science.

Author Contributions:
EB advised on the study and helped write the manuscript. LHD conceived of the study, was responsible for its design and coordination, implemented the Booly aliasing resource, and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements:
We thank Francisco Esteves and Vipul Bhargava for insightful discussions on this manuscript. Funding: NIH RO1 AI070654 and NS29870

References:
[1] Babnigg G & Giometti CS. *Proteomics* 2006 **6**: 4514 [PMID: 16858731]
[2] Eyre TA *et al. Nucleic Acids Res.* 2006 **34**: D319 [PMID: 16381876]
[3] Huang da W *et al. Bioinformation* 2008 **2**: 428 [PMID: 18841237]
[4] Hubbard TJ *et al. Nucleic Acids Res.* 2009 **37**: D690 [PMID: 19033362]
[5] Iragne F *et al. Bioinformatics* 2004 **20**: 2331 [PMID: 15059813]
[6] Smith M *et al. Bioinformatics* 2005 **21**: 3429 [PMID: 15961438]
[7] Wheeler DL *et al. Nucleic Acids Res.* 2008 **36**: D13 [PMID: 18045790]
[8] RL Rivest. *CRYPTO '90.* 1991 **537**: 303
[9] Do LH *et al. BMC Bioinformatics.* 2010 **11**: 513 [PMID: 20942966]
[10] http://www.nist.gov/manuscript-publication-search.cfm?pub_id=902326

Edited by P Kangueane
Citation: Do & Bier. Bioinformation 6(2): 83-85 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Redundancy of common reference databases. The DAVID Gene Conversion tool creates clusters of gene groups analogous to Entrez Gene. Gene clusters from DAVID are labeled with numerical identifiers that are reused and recycled after each update, thus not optimal for use as a reference identifier of a gene. A common approach is to convert aliases into a single source database (REFSEQ, Entrez, etc.) identifier for comparison. The table shows the lack of complete redundancy across multiple reference databases. Only 29% (37081/127749) of gene clusters identified by DAVID (v6.7) are found to be present in all five reference databases from the three organisms.

Unique DAVID Id's Mapped	Fruitfly	Mouse	Human	Total
DAVID-Refseq mRNA	13978	16262	16060	46300
DAVID-Entrez Gene ID	21227	58530	40959	120716
DAVID-Ensembl ID	13945	18307	18370	50622
DAVID-Genbank GI	14253	48480	37281	100014
DAVID-Gene Symbol	20571	44859	32100	97530
Total DAVID ID Overlap	11525	13231	12325	37081
Total Unique DAVID IDs	23569	59881	44299	127749