Leveraging Biomedical Ontologies and Annotation Services to Organize Microbiome Data from Mammalian Hosts

Indra Neil Sarkar, PhD, MLIS

Center for Clinical & Translational Science, Department of Microbiology & Molecular Genetics, and Department of Computer Science, University of Vermont, Burlington, VT

A better understanding of commensal microbiotic communities ("microbiomes") may provide valuable insights to human health. Towards this goal, an essential step may be the development of approaches to organize data that can enable comparative hypotheses across mammalian microbiomes. The present study explores the feasibility of using existing biomedical informatics resources – especially focusing on those available at the National Center for Biomedical Ontology – to organize microbiome data contained within large sequence repositories, such as GenBank. The results indicate that the Foundational Model of Anatomy and SNOMED CT can be used to organize greater than 90% of the bacterial organisms associated with 10 domesticated mammalian species. The promising findings suggest that the current biomedical informatics infrastructure may be used towards the organizing of microbiome data beyond humans. Furthermore, the results identify key concepts that might be organized into a semantic structure for incorporation into subsequent annotations that could facilitate comparative biomedical hypotheses pertaining to human health.

INTRODUCTION

Conservative estimates suggest that only 1 in 10 cells found in or on the human body are of eukaryotic origin. The vast majority of cells that can be identified are commensal microbes, most of which are of bacterial origin[1,2]. Historically, medical microbiology has focused on the pathogenic effects of microbes relative to human disease[3]. More recently, interest has grown within the medical microbiology community, as well as the greater biomedical research and clinical communities, to better understand the possible positive (and at times, indicators of negative[4]) effect of the immense "microbiome" communities that live among us[5,6].

In response to the need for having a better understanding of the human microbiome, massive government-funded initiatives are leading the way to generate significant catalogues of commensal microbiota associated with humans. Most notable of these are the Human Microbiome Project (HMP[7]), funded as part of the National Institutes of Health Roadmap, and the European Union funded Metagenomics of the Human Intestinal Tract (MetaHIT[8]) projects. The proposed endeavors are comparable in many ways to the Human Genome Project (HGP).

The progress of studies based on the HGP since its completion can be attributed to the volume of comparative information available through the International Nucleotide Sequence Database Consortium (GenBank, EMBL-Bank, and DDBJ)[9] enabled the ability to explore comparative hypotheses across the tree of life[10,11]. For example, one could posit that a set of homologous genes (found across multiple species as a result of evolutionary inheritance) might be associated with a particular phenotype. Perhaps the most classical example of this is the association of homeobox genes associated with development across the entirety of eukaryotic life, from drosophila to humans[12].

The goals of the HMP[7] are to "enable study of the variation in the human microbiome and its influence on disease." An essential aspect to the ultimate measure of the success of the HMP will be the enablement of the development and evaluation of comparative hypotheses based on microbiome data. To date, much of the suggested hypotheses[13-16] for the HMP based data have been comparing condition vs. non-condition states. For example, exploration of the different microbes present in the guts of clinically defined obese individuals versus non-obese individuals[13]. As in classical comparative genomics, there may be benefit to also compare the microbiomes across similar environments across related species. For example, to further understand the impact of the human gut microbiome on Crohn's disease, it may be beneficial to have a better understanding of the bovine gut microbiome (cows can be afflicted with Johne's disease, which has a similar clinical manifestation as Crohn's disease[17]).

As a first step towards enabling comparative microbiome hypotheses, it will be essential to develop mechanisms to identify and organize relevant information from within large repositories. The goal of this study is to explore the ability to leverage existing biomedical ontologies and recently described annotation services to identify potentially relevant data from within GenBank. The present study focuses on identifying relevant bacterial species from microbiomes of common domesticated animals.

MATERIALS AND METHODS

The aims of this study were to (1) identify the volume of microbiome data contained within GenBank; and, (2) determine the anatomical source of the microbiome data using biomedical ontologies. Within the scope of the present study, the microbiomes of ten organisms were targeted: Bovine (Cow), Ovine (Sheep), Murine (Mouse), Equine (Horse), Porcine (Pig), Feline (Cat), Canine (Dog), Hircine (Goat), Asinine (Donkey), and Lapine (Rabbit). These organisms were chosen because of their taxonomic classification (mammal) and that they are all domesticated animals. Because of the latter, they are often co-inhabitants in human environments or live in close proximity with humans and their microbiomes could have direct influence on human microbiomes.

Identifying Microbiome Bacteria Entries in GenBank. GenBank queries were developed for each host species using the following general pattern:

```
(1)   16S AND
(2)   ("GENUS SPECIES"[All Fields] OR
(3)   "ADJECTIVE"[All Fields]) AND
(4)   ("Bacteria"[Organism]) NOT
(5)   ("Eukaryota"[Organism]) NOT
(6)   ("Homo sapiens"[Organism] OR
(7)   "Homo sapiens"[All Fields]) NOT
(8)   soil[All Fields]
```

Where, `16S` was used to retrieve sequences associated with the small ribosomal sub-unit and `GENUS SPECIES` (line 2) in addition to `ADJECTIVE` (line 3) were replaced with the appropriate host values. For example, the values "Bos taurus" and "bovine" were respectively used for cow. Two additional lines were used for identifying sheep microbiome entries that accounted for the contained word ("ovine" is within "bovine").

```
(9)   NOT "bovine"
(10)  NOT ("Bos taurus"[Organism])
```

The queries were incorporated into a Ruby script that made use of the BioRuby[18] library interface to Entrez Utilities (E-Utilities[19]). Using the Ruby script, each entry was downloaded in the full GenBank format.

Determining Anatomical Source of GenBank Entries. A series of regular expressions within the aforementioned Ruby script were used to extract relevant fields that may contain anatomical source information for a given downloaded GenBank entry. In particular, two metadata fields were extracted and considered as "direct annotation" of isolation source: (1) `/ISOLATION_SOURCE="[anatomical feature]"`; and, (2) `/NOTE="ISOLATED FROM [anatomical feature]"`. Additionally, publication titles were

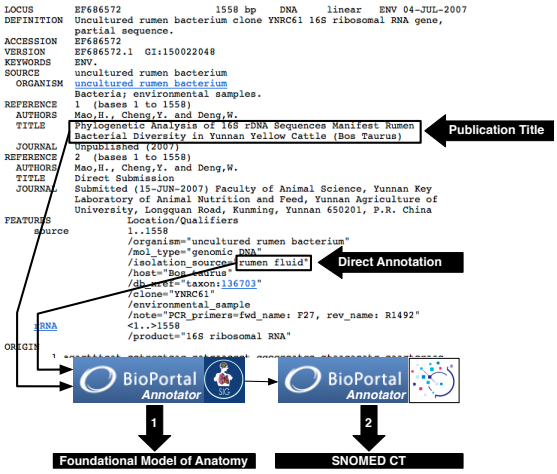

Figure 1: Overview of Annotation Process for the "Direct Annotation" and "Publication Title" Fields of a Given GenBank Entry.

extracted from a given GenBank entry using a regular expression in the Ruby script.

The REpresentational State Transfer (REST) interface to the National Center for Biomedical Ontology Annotator (NCBO Annotator[20]) was then used to extract anatomical terms from the Foundational Model of Anatomy (FMA[21,22]) [NCBO ID: 39966] or SNOMED CT[23,24] [NCBO ID: 40403] from the direct annotations and publication titles. For SNOMED CT, the extracted concepts were limited to the following UMLS Semantic Types:

```
T017: "Anatomical Structure"
T029: "Body Location or Region"
T023: "Body Part, Organ, or Organ Component"
T030: "Body Space or Junction"
T031: "Body Substance"
T022: "Body System"
T018: "Embryonic Structure"
T021: "Fully Formed Anatomical Structure"
T024: "Tissue"
T032: "Organism Attribute"
T020: "Acquired Abnormality"
T190: "Anatomical Abnormality"
T019: "Congenital Abnormality"
```

Additionally, when analyzing direct annotations the following UMLS Semantic Type was also permitted:

```
T047: "Disease or Syndrome"
```

The NCBO Annotator "Whole Word" and "Longest Only" options were set to 'false' when identifying possible anatomical concepts using FMA; conversely, they were set to 'true' when using SNOMED CT. The NCBO Annotator default set of stop words was used regardless of which target ontology was used. An overview of the overall annotation process is shown in Figure 1.

A second Ruby script was developed that tabulated the results from the annotations and organized them according to extracted features. The final results were then manually examined and false positive annotations (e.g., "gene cluster") were flagged and removed from the final result set.

RESULTS

A total of 28,959 GenBank entries were identified as potentially being associated with a microbiome organism across all 10 host organisms. The total time to process all the entries, including identification of possible annotation source sites took approximately four days of processing using a standard Internet connection. Based on an analysis of the putative microbiome organisms (as indicated in the ORGANISM GenBank Entry field), the result set reflected 881 bacterial species, of which 248 were explicitly labeled as "uncultured" or "unidentified." The result set contained 130 distinct genera of bacteria.

As summarized in Table 1, anatomical sources could be identified for an average of approximately 91% of the entries (range: 57-100%; overall: 97%). The number of GenBank entries for which anatomical sources could be identified ranged from 31 (Hircine) to 12,857 (Murine). In total, 130 anatomical sources were identified from either direct annotations or from publication titles. Figure 2 depicts the distribution of annotation sources (direct annotation only, publication title, or both). Manual examination of these 130 predicted anatomical sources revealed that direct annotations were associated with 5% error; publication titles had an error rate of 33%. Seventy-eight percent (n=101) of the identified anatomical sources were from FMA; SNOMED CT was associated with 22% (n=29). Of the 101 anatomical sources from FMA, 72% (n=72) were deemed as plausible microbiome collection sites; for SNOMED CT, 93% (n=27) were valid microbiome sites.

Table 1: Number of GenBank Microbiome Entries with Predicted Anatomical Source.

Host (alphabetical)	GenBank Entries	With Predicted Anatomical Source
Asinine	263	263 (100%)
Bovine	11,117	10,810 (97%)
Canine	998	903 (90%)
Equine	773	706 (91%)
Feline	421	412 (98%)
Hircine	54	31 (57%)
Lapine	449	432 (96%)
Murine	12,936	12,857 (99%)
Ovine	1021	952 (93%)
Porcine	1745	1593 (91%)
All	*29,777*	*28,959 (97%)*

DISCUSSION

This study aimed to explore the feasibility to leverage existing biomedical ontologies and available annotation services to identify data that are associated with microbiomes in non-human hosts. Based on the analysis presented here, bovine and murine microbiomes have the most amount of data identifiable within GenBank. This is likely because mice are perhaps the most popular mammalian "model" organism used within the biomedical sciences[25]; cows are also of great importance within the context of both agricultural and environmental studies. Nonetheless, all 10 of the chosen domestic mammal hosts had entries within GenBank that were associated with their respective microbiomes.

With the exception of Hircine host microbiomes, the approach described here was able to suggest anatomical sources for greater than 90% of the GenBank entries examined. Manual examination of the 23 Hircine GenBank entries for which no anatomical source could be predicted revealed that these particular entries were indeed not associated with microbiome samplings. Instead, the sequenced 16S GenBank entries are associated with bacterial infections. Cursory examination of the entries from other hosts reveals that the majority of the bacteria that could not be associated with an anatomical source were those associated with infections, not commensal microbes. This finding suggests that the GenBank query to identify entries associated with microbiomes will possibly need to be made more specific.

The results of this study strongly suggest that both FMA and SNOMED CT are suitable ontological structures to begin classification of anatomical sources that are associated with microbiomes. FMA was chosen because it is generally considered the most complete ontological structure associated with anatomy[26]. However, the anatomy of focus for FMA

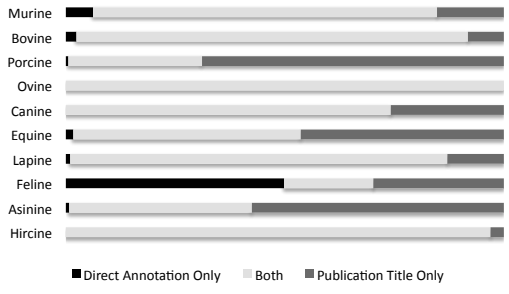

Figure 2: Distribution of Anatomic Sources relative to each Microbiome Host.

is human anatomy, and this is not necessary semantically precise for homologous (and especially for non-homologous) anatomical comparisons. By contrast, SNOMED CT, by historical design includes veterinary concepts[27], contains a great number of anatomical structures that are specifically non-human (e.g., "rumen"). Nonetheless, the Aristotelian organization[21] of structural anatomy in FMA may be leveraged in the development of subsequent semantic structures to categorize anatomical sources associated with microbiomes. Additional terminologies might be considered. To this end, future work may include leveraging the NCBO Ontology Recommender service[28].

Beyond the veterinary terms that are explicitly included in SNOMED CT, this study leveraged the disease semantic types (UMLS Semantic Type T047). This was particularly useful when direct annotations referred to a disease condition, not a particular anatomical feature. For example, in GenBank entry 229890767 the direct annotation is "young host with arthritis." While arthritis is not an anatomical location, the disease references in the direct annotation can be used to infer the anatomical source based on the disease condition. Therefore, identification of "arthritis" (SNOMED CT ID: 3723001) can be used to lead to an ultimate classification to "joint." Future work will leverage semantic relationships that will enable such mappings to be automatically generated based on disease conditions. However, the allowance for such mappings may be best limited to only direct annotations, since more confidence can be attributed to such inferences.

The high error rate of concepts associated with FMA is likely an artifact of the inability to leverage UMLS Semantic Types for FMA using the NCBO Annotator. Almost all of the incorrect suggested anatomical features from FMA were from the "biological macromolecule" hierarchy (e.g., "oligosaccharide"), and could therefore be explicitly excluded in future applications of the approach described here. There have been studies done that suggest that the concept overlap between SNOMED CT and FMA is significant[29], which might imply that SNOMED CT could potentially be used as the primary source for anatomical sources associated with microbiomes. This might be especially well suited because of the veterinary concepts that are contained with SNOMED CT. A future study will thus be to assess the ability to annotate the same set of GenBank entries exclusively with SNOMED CT.

Examination of the top ten anatomical sources associated with the microbiome data demonstrates the volume of entries associated with each of the mammalian hosts (shown in Table 2). As expected, the vast majority of entries associated with rumens (associated with bovine, ovine, or hircine hosts), which is perhaps the most studied anatomical sources of non-human mammals in general. Similarly, the majority of the top ten consists of anatomical sources associated with the gastrointestinal tract. This finding is consistent with the current sampling of the human microbiome projects (32% of the US-led HMP is from the gastrointestinal tract; the entirety of the EU-led MetaHIT project is devoted to the intestinal microbiome).

In addition to the potential scientific value of the findings presented here, this study demonstrates the ability for the existing biomedical informatics infrastructure to meet needs beyond human-centric biomedical inquiries. In particular, this study did not require any local installation of software or databases – the searches to GenBank were done using the E-Utilities; the annotations were done using REST Web services to the NCBO Annotator. The only data stored locally were the resultant annotations. This demonstration further underscores the possibility of distributed, "cloud" computing environments within the context of biomedical research. This is not to imply that further work is not needed in natural language processing or ontology development. Instead, this suggests that the fruits of significant natural language processing and ontology endeavors

Table 2: Host Distribution of Top Ten Anatomical Sources Identified From Within GenBank.

	Asinine	**Murine**	**Equine**	**Ovine**	**Bovine**	**Hircine**	**Canine**	**Porcine**	**Lapine**	**Feline**	**Total**
Rumen	-	-	-	1554	8283	30	-	-	-	-	9867 (18%)
Gut	255	5266	22	7	27	-	-	1113	219	-	6909 (12%)
Feces	255	418	591	-	2753	-	405	-	219	191	4832 (9%)
Intestine	-	4178	446	-	24	-	5	169	-	-	4822 (9%)
Small intestine	-	4172	-	-	-	-	-	8	-	-	4180 (8%)
Epidermis	-	2583	-	4	5	-	-	-	-	-	2592 (5%)
Ileum	-	1483	-	-	-	-	68	580	-	27	2158 (4%)
Bladder	-	-	-	-	2005	-	-	-	-	-	2005 (4%)
Cecum	-	1808	-	-	-	-	-	16	70	-	1894 (4%)
Foot	-	-	-	2	829	-	-	-	-	-	831 (2%)

are now capable of being consumed almost immediately by consumers, regardless of their specific area of focus. This is in contrast to how earlier studies of the kind described here would require complete installation of software tools, knowledge management constructs, and database management systems.

Based on the volume of annotatable microbiome results across the host organisms, a significant next step will be to perform sequence analysis comparisons between the commensal bacteria across organisms. Notably missing from this study was human commensal bacteria, which are generally annotated within the context of the HMP and MetaHIT projects. It is planned to add the human microbiome data within the context of future studies. A possible future study might be to compare the constituency of gastrointestinal bacteria found across organisms and compare it to their diet types (e.g., strict carnivore *vs.* strict herbivore *vs.* omnivore).

CONCLUSION

The results of this study suggest that identification of anatomical features associated with microbiome studies, that do not currently have explicit anatomical annotation, can be done with an average accuracy of 97%. Furthermore, this study demonstrates how the existing biomedical informatics ontology (FMA and SNOMED CT) and natural language processing (NCBO Annotator) infrastructure can be used for the identification of anatomical sources of commensal microbiome hosts.

ACKNOWLEDGEMENTS

Gratitude is given to the NCBO team (especially Dr. Nigam Shah) for the availability, description of, and assistance with the NCBO Annotator Web Service. INS is funded by NIH and NSF grants, R01-LM009725 and IIS-0241229, respectfully.

REFERENCES

1. Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell. 2006 Feb 24;124(4):837-48.
2. Tsai F, Coyle WJ. The microbiome and obesity: is obesity linked to our gut flora? Curr Gastroenterol Rep. 2009 Aug;11(4):307-13.
3. Isenberg HD. Clinical microbiology: past, present, and future. J Clin Microbiol. 2003 Mar;41(3):917-8.
4. Mans JJ, von Lackum K, Dorsey C, Willis S, Wallet SM, Baker HV, et al. The degree of microbiome complexity influences the epithelial response to infection. BMC Genomics. 2009;10:380.
5. Kinross JM, von Roon AC, Holmes E, Darzi A, Nicholson JK. The human gut microbiome: implications for future health care. Curr Gastroenterol Rep. 2008 Aug;10(4):396-403.
6. Tuohy K. The human microbiome--a therapeutic target for prevention and treatment of chronic disease. Curr Pharm Des. 2009;15(13):1401-2.
7. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, et al. The NIH Human Microbiome Project. Genome Res. 2009 Dec;19(12):2317-23.
8. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. Mar 4;464(7285):59-65.
9. http://insdc.org/.
10. Cai X, Hu H, Li X. A new measurement of sequence conservation. BMC Genomics. 2009;10:623.
11. Noonan JP. Regulatory DNAs and the evolution of human development. Curr Opin Genet Dev. 2009 Dec;19(6):557-64.
12. Kappen C, Ruddle FH. Evolution of a regulatory gene family: HOM/HOX genes. Curr Opin Genet Dev. 1993 Dec;3(6):931-8.
13. Turnbaugh PJ, Gordon JI. The core gut microbiome, energy balance and obesity. J Physiol. 2009 Sep 1;587(Pt 17):4153-8.
14. Proal AD, Albert PJ, Marshall T. Autoimmune disease in the era of the metagenome. Autoimmun Rev. 2009 Jul;8(8):677-81.
15. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. Metagenomic pyrosequencing and microbial identification. Clin Chem. 2009 May;55(5):856-66.
16. Hattori M, Taylor TD. The human intestinal microbiome: a new frontier of human biology. DNA Res. 2009 Feb;16(1):1-12.
17. Greenstein RJ. Is Crohn's disease caused by a mycobacterium? Comparisons with leprosy, tuberculosis, and Johne's disease. Lancet Infect Dis. 2003 Aug;3(8):507-14.
18. http://www.bioruby.org/.
19. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. Jan;38(Database issue):D5-16.
20. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA. Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics. 2009;10 Suppl 2:S1.
21. Rosse C, Mejino JL, Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003 Dec;36(6):478-500.
22. http://sig.biostr.washington.edu/projects/fm/.
23. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform. 2006;121:279-90.
24. http://www.ihtsdo.org/snomed-ct/.
25. Brown SD, Hancock JM. The mouse genome. Genome Dyn. 2006;2:33-45.
26. Zhang S, Bodenreider O. Experience in Aligning Anatomical Ontologies. Int J Semant Web Inf Syst. 2007;3(2):1-26.
27. Zimmerman KL, Wilcke JR, Robertson JL, Feldman BF, Kaur T, Rees LR, et al. SNOMED representation of explanatory knowledge in veterinary clinical pathology. Vet Clin Pathol. 2005;34(1):7-16.
28. Jonquet C, Shah NH, Musen MA, editors. Prototyping a Biomedical Ontology Recommender Service. 2009 Bio-Ontologies SIG at ISMB/ECCB 2009; 2009; Stockholm, Sweden.
29. Bodenreider O, Zhang S. Comparing the representation of anatomy in the FMA and SNOMED CT. AMIA Annu Symp Proc. 2006:46-50.