GNOMICS: A one-stop shop for biomedical and genomic data

Charles J. Kronk, B.Sc.[1], Anil Jegga, DVM, M.Res.[1,2,3]
[1]Department of Biomedical Informatics, University of Cincinnati College of Medicine;
[2]Division of Biomedical Informatics, Cincinnati Children's Hospital and Medical Center;
[3]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio, USA

Abstract

The World Wide Web is an indispensable tool for biomedical researchers who are striving to understand the molecular basis of phenotype. However, it presents challenges in the form of proliferation of data resources, with heterogeneity ranging from their content to functionality to interfaces. This often frustrates researchers who must visit multiple sites, become familiar with their interfaces, and learn how to use them to extract knowledge. Even then, one may never feel sure that they have tracked down all needed information. We envision addressing this challenge with GNOMICS *(Genomic Nomenclature Omnibus and Multifaceted Informatics and Computational Suite), a suite with both a programmatic interface and a GUI.* GNOMICS *allows for extensible biomedical functionality, including identifier conversion, pathway enrichment, sequence alignment, and reference gathering, among others. It combines usage of other biological and chemical database application programming interfaces (APIs) to deliver uniform data which can be further manipulated and parsed.*

Introduction

Discussions of "big data" are on the forefront of researchers' minds in almost every discipline. However, in the biomedical sciences, the growth of biological data is unprecedented, with new data being created more quickly than it can be efficiently organized[1]. Automating large-scale, cross-domain heterogeneous data aggregation across many different sites continues to pose a major challenge for bioinformaticians. Several groups, including the U.S. National Center for Biomedical Ontology (NCBO), have aimed to standardize data via the usage of normalized ontologies and nomenclatures[2]. It is important to note that these controlled vocabularies are sometimes difficult to map to or from, due, in part, to looming digital dark ages, referring to difficulties in understanding various file formats and identifier types after their deprecation[3]. Having a "universal" identifier for biological concepts continues to be a challenge. This problem is further compounded as new databases or meta-databases (e.g. comprehensive databases) are constructed. These efforts often end up introducing new identifiers for existing concepts. Further, the status quo as far as advanced biological databases are concerned seems to be locally downloading databases or simply copying them to other servers. This results not only in duplicate data, but also in a data source which can quickly become out-of-date. A potential solution to these concerns, discussed herein, is the usage of application programming interfaces (APIs). These systems abstract database implementations, making input and output data easier to query, consume, and analyze in a standardized manner (Fig. 1).

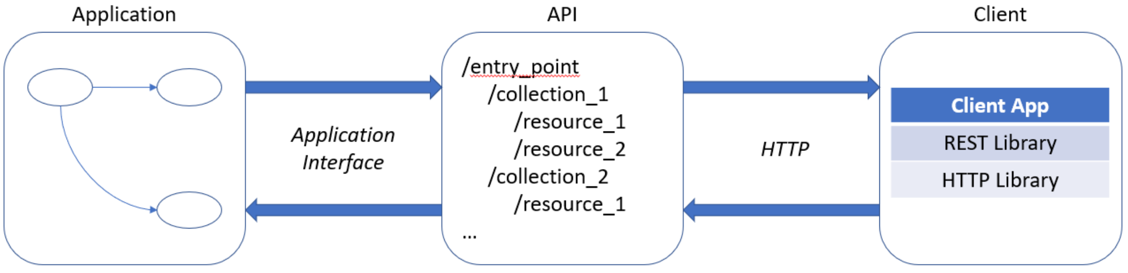

Figure 1. Example outlining data flow to/from a client to an application via an API.

Many biological databases have open APIs, such as Ensembl, UniProt, PubMed, and KEGG, while others are semi-open, requiring registration and some kind of security key (such as UMLS or Open PHACTS)[4,5,6,7]. Since APIs are tied directly to their component databases, they are as up to date as possible. In addition, minimal data are pulled to the local machine when queried, mitigating the copying of large databases.

Our system, labeled *GNOMICS* (Genomic Nomenclature Omnibus and Multifaceted Informatics and Computational Suite) is a grouping of API-based Python programs organized into objects which can be extended or scaled as needed. In addition to the command-line interface, a beta front-end is also available, having been built primarily using Electron and AngularJS.

Methods

The *GNOMICS* system was built upon user-friendly and understandable objects, of which there are currently 25 (Table 1). While these modules and others are still under development, they allow for a diverse array of functions, including *arterial supply* for anatomical structures, *effective time* for drugs, *metabolic rate* for taxa, and *effective rotor count* for compounds, among others.

Additionally, a twenty-sixth User object is available. This object stores information about the eleven API systems that require a separate user API key for usage (ChemSpider, DPLA, Elsevier, EOL, FDA, ISBNdb, NCBO, OMIM, OpenPHACTS, Springer, and UMLS)[4,5,6,7,8,9]. Currently 35 API systems are available, ranging from UniProt to Ensembl to Wikipedia. Lastly, 51 Python 3 packages were adapted for *GNOMICS* use, including BeautifulSoup, Bio, BioPython, Bioservices, the ChEMBL Webresource Client, ChemSpiPy, GEOparse, Intermine, LibChEBIpy, MyGene, MyVariant, NLTK, PubChemPy, and PyTaxize[10,11,12,13,14,15,16]. The inclusion of these packages prevented unnecessary code duplication.

Table 1. *GNOMICS* objects breakdown. Showcased are the approximate number of available records, number of identifiers currently available, and the number of properties included in the object. Please note that not all sources have published exact numbers of available objects.

Object	# Records	# Identifiers	# Properties
Adverse Event	30,126	7	0
Anatomical Structure	526,706	39	8
Assay	1,267,241	4	14
Biological Process	26,678	1	0
Cell Line	125,369	4	0
Cellular Component	4,150	1	0
Clinical Trial	619,701	5	24
Compound	332,518,372	23	55
Disease	382,830	28	11
Drug	1,541,646	34	21
Gene	45,569,934	21	16
Genotype	1,069	1	0
Molecular Function	11,134	2	0
Patent	4,000,000	3	0
Pathway	2,010,623	5	10
People	26,759,399	1	0
Phenotype	358,575	19	0
Procedure	397,186	8	0
Protein	586,444,112	15	9
Reference	477,357,444	20	13
Symptom	3,129	3	0
Taxa	6,739,136	11	98
Tissue	17,520	6	0
Transcript	202,338	2	0
Variation	801,208,836	5	0
Total:	2,288,123,254	268	279

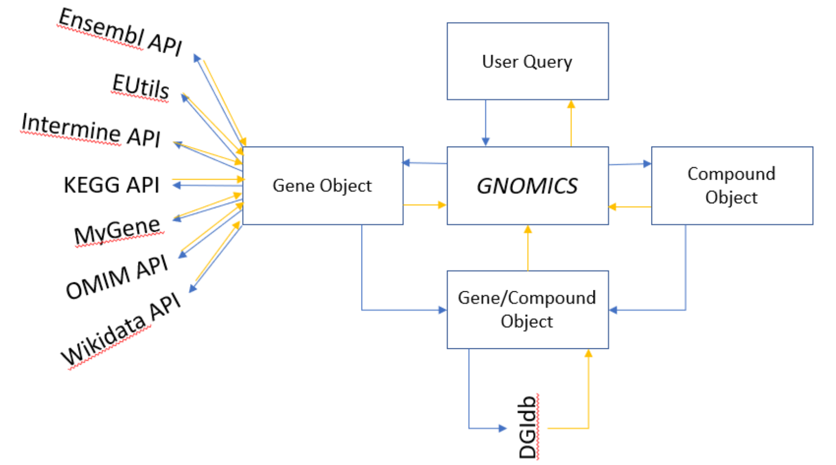

Figure 2. Example workflow for user query for a gene/compound interaction. Blue arrows signify data calls, whereas yellow arrows signify the returned data pipeline. Only individual gene API calls are expanded for emphasis.

Each component object has several subfolders whose functions relate to parsing specific materials returned from component APIs for command-line usage, allowing for simpler functions such as identifier conversion. *Interaction objects*, which allow for relating an object to another object or series of objects, are paramount to the success of *GNOMICS* (Fig. 2). For example, a gene can be related to several other genes via orthology; a disease can be related to a particular anatomical structure if it occurs there; or a drug can be related to adverse events (AEs) through drug-compound interactions. 63 such interaction objects are available or under development.

In addition, there are areas for locally downloading data to facilitate faster querying time (as well as allowing for integration of databases without programmatic access) and an area for documentation. Allowing for different import and export formats is crucial as well, with CSV, TSV, XML, HTML, JSON, JSONP, and YAML appearing as the most commonly parsed formats, alongside biological formatting standards such as BED and FASTA. Accessory are preformatted reports summarizing data in PDF, Microsoft Word, Microsoft Excel, CSV, and TSV files.

Results

268 object-specific identifiers were included in *GNOMICS* with various mappings originating from the Ontology Lookup Service (OLS) at EMBL-EBI, the National Center for Biomedical Ontology (NCBO) BioPortal, the UMLS Metathesaurus, Wikidata, and others[2,4,17]. These mappings may be one-to-one or one-to-many. The Human Disease Ontology (DOID) for heart disease, for example, maps to 6 SNOMED-CT U.S. IDs in OLS and NCBO, but only to one UMLS CUI in both sources[2,4,17]. Non-object-specific identifiers (such as MedDRA, SNOMED-CT, or MeSH) must be mapped to an object-specific identifier before inclusion with that specific ID[2,4,17,19,20]. Including non-object-specific identifiers, 533 identifier types are available among the 25 object types. Many identifiers appear in multiple sources allowing for mapping if a single source is under maintenance or briefly unavailable for other reasons (on average, each identifier is available from three sources).

Included with identifiers are terms and labels provided by the various sources, as well as accepted synonyms and translations. 76 languages are available due to integration of translatable terms from MeSH, MedDRA, CPT, LOINC, and Wikipedia.

First, a few test searches were run for all objects with available search functionality ("breast cancer," "SLC4A1," "ibuprofen", etc.). Each test was run five times across the 17 such objects, with all component databases and ontologies returning results in an average 0.160864599 seconds (s = 0.01142). Next, we timed 45 randomly chosen interaction object search functions using the same methodology. These returned results in an average 0.007318 seconds (s = 0.00525). Artificially crippling single APIs with repeated data (such as

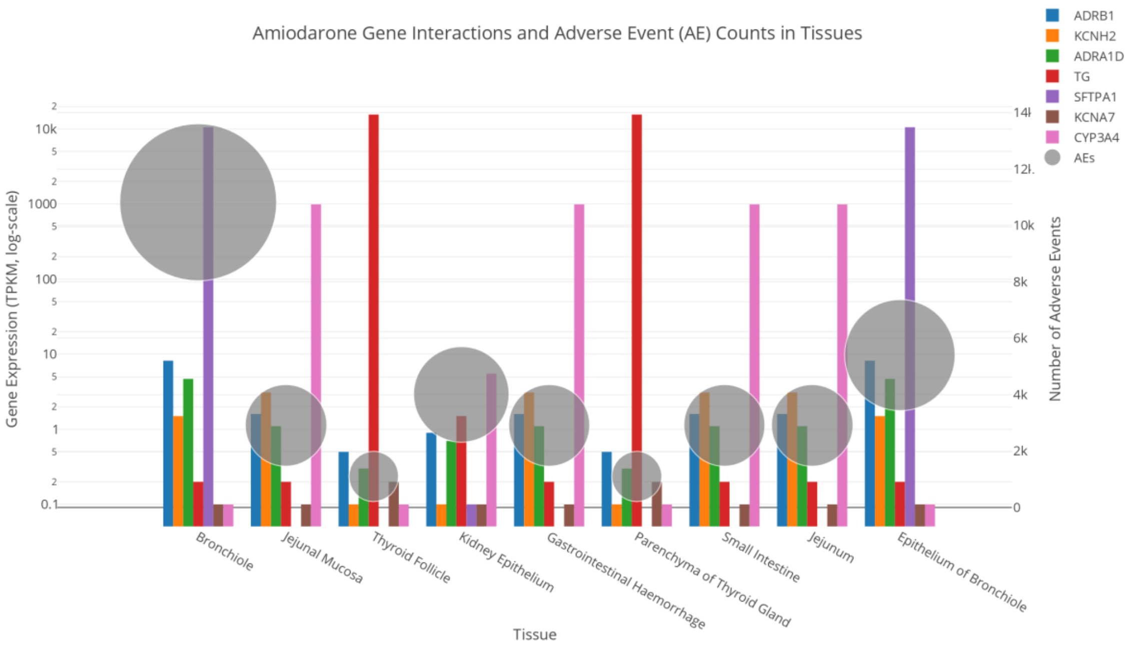

Figure 3. Example output showcasing amiodarone's observed AEs mapped to tissues, alongside normal gene expression for known drug-gene interactions in amiodarone. Note the large number of AEs mapped to bronchiole and the relatively high expression of *SFTPA1* in bronchiole. Further, *SFTPA1* polymorphisms have been noted as influencing susceptibility to pulmonary fibrosis and pulmonary toxicity is one of the most serious AEs related to amiodarone usage[24,25].

disallowing access to the OLS domain) was met with no significant change in speed, as most identifiers and search results could be found on multiple platforms. This, however, is not currently the case for PubMed or Ensembl, among others.

Two more additional complex tests looked at synthesizing phenotypes related to adverse events caused by a drug and phenotypes related to the disease that drug is used to treat, as well as normal gene expression for genes known to interact with a given drug in certain tissues alongside AE counts endemic to those same tissues[8,21,22,23]. The first test found 200 phenotypes associated with amiodarone usage and 20 phenotypes observed in the adverse event (AE) cardiac arrhythmia. There were 14 overlapping phenotypes discovered, including bradycardia, congestive heart failure, cardiac arrest, and ventricular tachycardia; all of which are common side effects of amiodarone. The second test involved looking at amiodarone AE counts and mapping those events to tissues; next we found drug/gene interactions for amiodarone and found normal expression levels for those genes in the same tissues (Figure 3).

Following this, extensibility was tested by adapting PubTator and NOBLE Coder into *GNOMICS*, allowing for programmatic annotation of free text. The NLTK package was added to supplement these programs as well[26,27].

Discussion

Cross-domain integration and comparison of large volumes of information across clinical and translational research applications continues to be a major challenge in the biomedical domain. With both a programmatic and visual interface, as well as a focus on terminology translation, we anticipate that *GNOMICS* and its continued development could accelerate the process of bringing together the two disparate fields. Additionally, *GNOMICS* has an intuitive object-based usage that can be understood quickly, keeping the relative startup cost smaller than that of a dedicated database download. The system manages to deliver uniform data which can be easily manipulated and parsed as the user finds necessary. *GNOMICS* further remains continually up-to-date given its connectedness with its component sources and continues to be a portable and scalable solution in the biomedical data landscape. However, *GNOMICS* has certain limitations. Because it is based on web services, *GNOMICS* requires a dedicated internet connection and more substantial API limits. While several biomedical and genomics databases have open API systems, others require registration and complex layers of security tokens and request tickets. While *GNOMICS* does some of the heavy lifting in these situations, users still have to navigate some of these processes manually. Further, for what *GNOMICS* gains in portability, it loses in speed, as certain databases may take longer to respond to queries. However, it has been shown that given increases in processing power, this tradeoff is often negligible to the end-user.

Conclusion

GNOMICS is under continuous development, including addition of more object types, functions, and identifier mappings. Currently, the *GNOMICS* system is biased towards well-established and/or heavily trafficked databases, especially those with APIs. Additional multi-system testing is ongoing, with the front-end interface still in active development. This includes taking into account portability- and scalability-related issues alongside those components of the command-line interface. To address these points and others, a public GitHub source (https://github.com/Superraptor/gnomics) is available with a README and Wiki. Documentation is provided which offers insight into identifier types available, database coverage, and language coverage, among other things. Additions to and recommendations for the system are highly encouraged, with plugins written in Python or other programming languages easily adaptable into the existing framework.

References

1. Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomed. Inform. Insights. 2016 Jan;8:1-10.
2. Martínez-Romero, M., Jonquest, C., O'Connor, MJ., Graybeal, J., Pazos, A., Musen, MA. NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation.
3. U.S. National Academy of Sciences, U.S. National Academy of Engineering, U.S. Institute of Medicine. Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. Washington, D.C. (USA): U.S. National Academies Press, 2009.
4. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Res. 2004 Jan;32(Database Issue):D267-D270.

5. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. KEGG as a Reference Resource for Gene and Protein Annotation. Nucleic Acids Res. 2016 Jan;44(Database Issue):D457-D462.
6. Waagmeester, A., Kutmon, M., Riutta, A., Miller, R., Willinghagen, EL., Evelo, CT., Pico, AR. Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. PLoS Comput. Biol. 2016 Jun;12(6):e1004989.
7. Ruffier, M., Kähäri, A., Komorowska, M., Keenan, S., Laird, M., Longden, I., Proctor, G., Searle, S., Staines, D., Taylor, K., Vullo, A., Yates, A., Zerbino, D., Flicek, P. Ensembl Core Software Resources: Storage and Programmatic Access for DNA Sequence and Genome Annotation. Database (Oxford). 2017;2017(2017):bax020.
8. Kass-Hout, TA., Xu, Z., Mohebbi, M., Nelsen, H., Baker, A., Levine, J., Johanson, E., Bright, RA. OpenFDA: An Innovative Platform Providing Access to a Wealth of FDA's Publicly Available Data. J. Am. Med. Inform. Assoc. 2016 May;23(3):596-600.
9. Amberger, JS., Bocchini, CA., Schiettecatte, F., Scott, AF., Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders. Nucleic Acids Res. 2015 Jan;43(Database Issue):D789-D798.
10. de Matos, P., Alcantara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., Steinbeck, C. Chemical Entities of Biological Interest: An Update. Nucleic Acids Res. 2010 Jan;38(Database Issue):D249-D254.
11. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., Steinbeck, C. The ChEBI Reference Database and Ontology for Biologically Relevant Chemistry: Enhancements for 2013. Nucleic Acids Res. 2013 Jan;41(Database Issue):D456-D463.
12. Bento, AP., Gaulton, A., Hersey, A., Bellis, LJ., Chambers, J., Davis, M., Krüger, FA., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., Overington, JP. The ChEMBL Bioactivity Database: An Update. Nucleic Acids Res. 2014 Jan;42(D1):D1083-D1090.
13. Kim, S., Thiessen, PA., Bolton, EE., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, BA., Wang, J., Yu, B., Zhang, J., Bryant, SH. PubChem Substance and Compound Databases. Nucleic Acids Res. 2016 Jan;44(D1):D1202-1213.
14. Pence, HE., Williams, A. ChemSpider: An Online Chemical Information Resource. J. Chem. Educ. 2010 Aug;87(11):1123-1124.
15. Blake JA., Eppig, JT., Kadin, JA., Richardson, JE., Smith, CL., Bult, CJ., Mouse Genome Database Group. Mouse Genome Database (MGD)-2017: Community Knowledge Resource for the Laboratory Mouse. Nucleic Acids Res. 2017 Jan;45(D1):D723-D729.
16. Wu, C., MacLeod, I., Su, AI. BioGPS and MyGene.info: Organizing Online, Gene-centric Information. Nucleic Acids Res. 2013 Jan;41(Database Issue):D561-D565.
17. Côté, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, JA., Hermjakob, H. The Ontology Lookup Service: Bigger and Better. Nucleic Acids Res. 2010 Jul;38(Web Server Issue):W155-W160.
18. Kibbe, WA., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, CJ., Binder JX., Malone, J., Vasant, D., Parkinson, H., Schriml, LM. Disease Ontology 2015 Update: An Expanded and Updated Database of Human Diseases for Linking Biomedical Knowledge through Disease Data. Nucleic Acids Res. 2015 Jan;43(Database Issue):D1071-D1078.
19. Brown, EG., Wood, L., Wood, S. The Medical Dictionary for Regulatory Activities (MedDRA). Drug Saf. 1999 Feb;20(2):109-117.
20. Sarntivijai, S., Zhang, S., Jagannathan, DG., Zaman, S., Burkhart, KK., Omenn, GS., He, Y., Athey, BD., Abernethy, DR. Linking MedDRA(®)-Coded Clinical Phenotypes to Biological Mechanisms by the Ontology of Adverse Events: A Pilot Study on Tyrosine Kinase Inhibitors. Drug Saf. 2016 Jul;39(7):697-707.
21. Köhler, S., Vasilevsky, NA., Engelstad, M., Foster, E., McMurry, J., Aymé, S., Baynam, G., Bello, SM., Boerkoel, CF., Boycott, KM., Brudno, M., Buske, OJ., Chinnery, PF., Cirpriani, V., Connell, LE., Dawkins, HJS., DeMare, LE., Devereau, AD., de Vries, BBA., Firth, HV., Freson, K., Greene, D., Hamosh, A., Helbig, I., Hum, C., Jähn, JA., James, R., Krause, R., Laulederkind, SJF., Lochmüller, H., Lyon, GJ., Ogishima, S., Olry, A., Ouwehand, WH., Pontikos, N. Rath, A., Schaefer, F., Scott, RH., Segal, M., Sergouniotis, PI., Sever, R., Smith, CL., Straub, V., Thompson, R., Turner, C., Turro, E., Veltman, MWM., Vulliamy, T., Yu, J., von Ziegenweidt, J., Zankl, A., Züchner, S., Zemoijtel, T., Jacobsen, JOB., Groza, T., Smedley, D., Mungall, CJ., Haendel, M., Robinson, PN. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017;45(Database Issue):D865-D876.

22. Mungall, CJ., McMurry, JA., Köhler, S., Balhoff, JP., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, JP., Jacobsen, JOB., Keith, D., Laraway, B., Lewis, SE., NguyenXuan, J., Shefchek, K., Vasilevsky, N., Yuan, Z., Washington, N., Hocheiser, H., Groza, T., Smedley, D., Robinson, PN., Haendel, MA. The Monarch Initiative: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species. Nucleic Acids Res. 2017 Jan;45(Database Issue);D712-D722.
23. Nelson, SJ., Zeng, K., Kilbourne, J., Powell, T., Moore, R. Normalized Names for Clinical Drugs: RxNorm at 6 Years. J. Am. Med. Inform. Assoc. 2011 Jul-Aug;18(4):441-448.
24. Cottin, V., Wijsenbeek, M., Bonella, F., Vancheri, C. Slowing Progression of Idiopathic Pulmonary Fibrosis with Pirfenidone: From Clinical Trials to Real-life Experience. Clin. Invest. 2014;4(4):317-330.
25. Wolkove, N., Baltzan, M. Amiodarone Pulmonary Toxicity. Can. Respir. J. 2009 Feb;16(2):43-48.
26. Wei, C., Kao, H., Lu, Z. PubTator: A Web-based Text Mining Tool for Assisting Biocuration. Nucleic Acids Res. 2013 Jul;41(Web Server Issue);W518-W522.
27. Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., Jacobson, RS. NOBLE - Flexible Concept Recognition for Large-scale Biomedical Natural Language Processing. BMC Bioinformatics. 2016 Jan;17:32.