

WORKSHOP ON TOPOLOGY AND ABSTRACT ALGEBRA FOR BIOMEDICINE

ERIC K. NEUMANN

Foundation Medicine, Cambridge, MA 02139, USA

Email: enemann@foundationmedicine.com

SVETLANA LOCKWOOD

School of Electrical Engineering and Computer Science,

Washington State University, Pullman, Washington, USA

Email: svetlana.lockwood@email.wsu.edu

BALA KRISHNAMOORTHY

School of Electrical Engineering and Computer Science,

Washington State University, Pullman, Washington, USA

Email: bkrishna@math.wsu.edu

DAVID SPIVAK

Department of Mathematics,

MIT, Cambridge, MA 02139, USA

Email: dspivak@math.mit.edu

The use of large-scale data analytics, aka Big Data, is becoming prevalent in most information technology discussions, especially for the life and health sciences. Frameworks such as MapReduce/Hadoop are offered as “Swiss-army knives” for extracting insights out of the terabyte-sized data. Beyond the sheer volume of the data, the complexity of the data structure associated with such data sets is another issue, and may not be so readily mined using only these technological solutions. Rather, the issues around data structure and data complexity suggest new representations and approaches may be required. The LinkedData standard (W3C, Semantic Web) has been promoted by some communities to address complex and aggregatable data, though it focuses primarily on querying the data and performing logical inferences on it, and its use in deep mining application is still in the early stages. In summary, there appears to be a gap between how we access structured data, and the deeper analyses we want to perform on it that preserve representation.

Over the last few years, an increasing number of examples from life science research have appeared that apply topological and algebraic forms to genomic and complex data problems (Isomap[†], PLEX[‡], Ayasdi, FQL[§], BioHaskell^{**}). The relevance of finding structure in rich data has been underscored by the increasing efforts to combine

[†] <http://isomap.stanford.edu>

[‡] <http://www.math.colostate.edu/~adams/jplex/index.html>

[§] <http://categoricaldata.net/fql.html>

^{**} <http://biohaskell.org>

clinical data with genomic analyses. Although much attention has been placed on Big Data and *batching* computational algorithms (e.g., MapReduce), understanding the structure of the data to better analyze, extract, and infer insights from it are also critical. These areas, however, are currently not supported sufficiently for the health and life sciences communities, and many possible applications are only recently being proposed^{1,4,6}.

Coming from a very different perspective, abstract algebra and algebraic topology (AAAT) may provide new powerful insights to biomedical data sciences. Historically, these algebra forms have been very successful in the study of many profound topics, yielding an understanding of rich mathematical and logic structures, as well as their relations to one another. Several key advances in the computer sciences over the last few decades (relational algebras (SQL), monadic structures (javascript), description logic (OWL), homotopy theory), have emerged from these fields of study. Yet, due to their mathematical generalities, other facets of abstract algebras have not easily been applied to domain-specific applications such as biomedical research. The potential is just beginning to emerge from limited cross-pollination as the landscape shifts to greater use of large, diverse data sets. What is yet lacking is a set of lucid, yet powerful examples from AAAT to biomedical applications that will help establish a bridge between these diverse disciplines.

Life science data is a mix of conceptual relations (aka knowledge, e.g., *proteins encoded by genes*) based on our current understanding of biology, and the data measurements gathered from applying large-scale chemical and genomic profiling technologies. The latter set is often assumed to “rest” on top of the conceptual entities (genes, proteins, mRNA, cellular structures), which have specific relations with each other (e.g., protein -> gene deterministic mappings). The logic associated with conceptual models that house data could be extended with additional AAAT theorems in order to enable a much deeper analysis of the data.

As an initial example, consider some concepts around topologies, which can be used to describe different “molecular spaces”, including a sequence topology based on what makes one sequence similar or different from another. Here one element represents an entire genome for a given individual of a species, and the adjacent elements (neighborhood) are genomes from other individuals that differ in only a few bases. No scalar metric may exist in this space, but the overlap of subsets containing similar elements, and subsets that are only related by many different subset coverings provides a very discrete topology^{††}. In addition to the elements, edges between the elements may be included that represent the incremental mutational transitions, unequal rearrangements, and reciprocal recombinations that may occur. Given a starting set of elements (genomes with only a few alleles), multiple applications of recombination to the elements will define a limited space of “accessible” genomes, known as a *closure*. A corollary from this is the *Founder Effect*

^{††} It is enormous, since every 1000-base string has $3 * 1000$ one-step neighbors and 9,000,000 (3000^2) two-step neighbors, and so on.

and H-W equilibrium for a limited starting population cut-off from larger allelic set. Only additional mutations can free genomes from these closures.

Topologies can obviously be applied to protein sequences as well, but proteins also offer additional relations including interactions to other proteins. One also realizes that such interactions depend (in complex ways) to their underlying sequence, so that the genome topology space captures the interaction graph of the proteins “fibered” above each coding region of the genome. One can continue to build upon these objects, to yield dynamic networks that can affect states and synthesis/degradation of all biomolecules. Eventually a topological mapping between genome space and phenotype spaces can begin to be formally represented², and to some extent, be possibly projected from the underlying genomic information.

Categorical Theory^{3,4} (CT) is major device that originates from abstract algebras, and has several powerful features for organizing concepts and inherent system logic. Categories are composed of objects, morphisms (relations), and the ability to compose morphisms into transitive maps. Here *objects* are equivalent to what most of us call *classes*, and the morphisms define the relations between objects. The universal properties that come along with these entities allow combining objects, determining uniqueness, and establishing equivalencies between the objects and some fundamental morphisms, e.g., maps from an unique initial object to any other object defines exactly one relation per object called an *element*). These can be populated with a set of genes of interest and the relations they yield, including *commuting* paths. For example, for every protein p , the map (i) from p to its transcript, r , can be composed with the map (j) from r to the gene g it is expressed from, to yield the composition $(k) = (j) \circ (i)^{\#\#}$. Not only does (k) map a protein to a gene, but it is guaranteed to always have the same results as $(j) \circ (i)$, even though multiple proteins can map to the same transcript, and multiple transcripts can map to the same gene. We say that these relational structures *commute*.

One important feature of CT is the definition and use of *functors*³, which not only transform objects to other objects (within or between categories), but also their morphisms to other morphisms. They become very useful when taking data structures of one model (e.g., a genome topology) to a more advantageous form for a different problem (e.g., a graph of molecular interactions). Since the relations (morphisms) come along as well, both the data and their semantics can be effectively transformed together. As will be described below, this applies to databases as well as analytical manipulations.

Another important area of topology is the representation of simplicial complexes, which are the compositions of ordered relations of entities for different dimensional objects: points, edges, faces, volumes, etc. Each n -dimensional object, or n -simplex, is composed of $n+1$ $(n-1)$ -dimensional objects: a 3d tetrahedron has 4 2d-triangular

^{\#\#} The notation for composition is always read right-to-left, since they are operators.

faces, each with 3 1d edges, consisting of 2 0d points. If the edges between any 2 points are less than a distance ϵ , they can be chained into complexes. Furthermore, if any form a cycle of 3 edges, a face object is induced and identified with directionality (clockwise or counter-clockwise); the same method is applied to successfully generate and chain higher structures, such as tetrahedrons (3-simplex), and beyond. When applied to complex data that have some form of distance metric, they form clusters of multidimensional simplicial complexes chains. The analysis of such complexes yields understanding of the general structure, or homotopy, of the data field. The use of *barcode analysis*, or *persistent homology*, by current researchers^{5,6,7} is one path of analysis that is helping identify complex relations with biomedical data.

Altogether, the above areas support data mining of complex “high-feature” data while also aligning it to established and hypothetical concepts/relations and the logic they induce on the data elements. Often data analytics is tied to data representations within a database or other kind of repositories. As stated before, AAAT has helped shape current tools and methodologies supporting schemas and ontologies. This can be further enhanced by recent work on Category Theory as applied to databases⁸. These can address important issues on data migration, schema changes, data integrity and normalization, and intelligent query strategies. Most database operations are some combination of three fundamental operations: project, join, union^{8,9}.

Combining the logical manipulations of biomolecular relations along with the data captured for these under select conditions, new data constructs can be produced or perhaps even automatically generated to address complex analytics. Time series and tissue-specific data (e.g., with gene expression) can be formally encoded as (Cartesian) products of simpler objects within a category, and inherit their logical relations directly from original set of morphisms by applying universal properties (e.g., limits) and *functors*. Analytic tools that understand such composite structures, as well as the multitude of properties linked to each object (gene, patient, tumor, study, etc.), can then perform deep analytics intelligently using decomposition rules on the data subsets (sigma algebras). The vision would be that biomedical data becomes less pre-structured by computer science, and more emergent in structure based on the rules for combining data, on analyzing data, and inferring hypotheses from data. What is most important now is to come together as a community and discuss what directions we should consider exploring, and to identify a few relevant exemplar cases to work on in order to validate these ideas.

References

1. I. C. Baianu. “A Category Theory And Higher Dimensional Algebra Approach To Complex Systems Biology, Meta-Systems And Ontological Theory Of Levels: Emergence of Life, Society, Human Consciousness and Artificial Intelligence”, *Mathematics and Bioinformatics* 30(12): Special Issue, 09/2012.

2. F. Mynard and G. J Seal, “Phenotype spaces”, *Journal of Mathematical Biology*, 60(2):247-66, 2009.
3. Saunders Mac Lane, “Categories for the Working Mathematician (Graduate Texts in Mathematics)”, Springer-Verlag, 1998.
4. David Spivak, “Category Theory for the Sciences”, MIT Press, 2014.
5. G. Carlsson and A. Zomorodian, “Computing persistent homology”, *Journal of Discrete and Computational Geometry*, 2004.
6. G. Carlsson, M. Nicolau and A. Levine, “Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival”, *Proceedings of the National Academy of Sciences*, , April 26, 2011.
7. Vin de Silva, “Topological and Symmetrical Structures in Data Analysis”, http://www.samsi.info/sites/default/files/deSilva_Lecture5_august2013.pdf
8. D. I. Spivak, R. Wisnesky Relational Foundations For Functorial Data Migration, *CoRR* vol. abs/1212.5303, 2012.
9. D. J. Abadi , A. Marcus , S. R. Madden , K. Hollenbach, Scalable semantic web data management using vertical partitioning, *Proceedings of the 33rd international conference on Very large data bases*, September 23-27, 2007, Vienna, Austria