# Mechanistic hypothesis generation in molecular biology: A grand challenge for knowledge-based reasoning

Position paper

Lawrence E Hunter
University of Colorado School of Medicine
Aurora, CO
larry.hunter@ucdenver.edu

## CCS CONCEPTS

• **Theory of computation → Semantics and reasoning**; • **Computing methodologies → Information extraction**; **Knowledge representation and reasoning**; **Causal reasoning and diagnostics**; *Ontology engineering*; *Reasoning about belief and knowledge*; • **Applied computing → Systems biology**; *Biological networks*; • **Mathematics of computing →** *Causal networks*; *Computing most probable explanation*;

**ACM Reference format:**
Lawrence E Hunter. 2018. Mechanistic hypothesis generation in molecular biology: A grand challenge for knowledge-based reasoning. In *Proceedings of ACM KBCOM conference, Los Angeles, CA, USA, January 2018 (KBCOM),* 3 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 MECHANISMS IN BIOMEDICAL RESEARCH

Biomedical research largely centers around the generation and testing of hypotheses regarding the mechanisms underpinning observed biological phenomena. Over the last 20 years or so, philosophers of science have made significant progress in accounting for the activities of biomedical research as a search for mechanistic accounts, in contrast, for example, with the search for law-like relationships that characterizes physics[7]. This new mechanistic philosophy defines mechanisms as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." In addition to the traditional methods of evaluating scientific theories (e.g. on the basis of their generality, or their predictive accuracy), competing mechanistic explanations are evaluated also on the basis of a variety of potential virtues (and vices) specific to mechanistic accounts, such as superficiality or incompleteness. Table 7.1 in [7] lists a set of questions about the abilities, activities, locations, roles, structures, and temporal features of components and steps in a mechanism that are typically pursued in the development of a mechanistic account of a phenomenon.

Psychological research has also pointed to the importance of forming causal hypotheses in human cognition. Children's frequent "why?" questions serve not just to engage parents in endless conversation, but to help children structure a broad causal account of their world[5]. In Daniel Kahneman's *Thinking: Fast and Slow*, he describes the fast, automatic system 1 as "highly adept at one kind of thinking—it automatically and effortlessly identifies causal connections between events, sometimes even when the connection is spurious." Appreciation of this phenomenon goes back at least as far as Charles Sanders Peirce, who observed in 1903 that "However man may have acquired his faculty of divining the ways of Nature, it has certainly not been by a self-controlled and critical logic. Even now he cannot give any exact reason for his best guesses.... For though it goes wrong oftener than right, yet the relative frequency with which it is right is on the whole the most wonderful thing in our constitution"[16]. He also added that "All the theories of science have been so obtained." Peirce coined the term "abductive inference," to name this task of inference to the best explanation.

However, there are a very limited set of computational tools or theories that either attempt to do this sort of inference or support human beings doing this sort of work. Most computational biology is focused on the analysis of data, not the creation of mechanistic or causal hypotheses to account for the data. Existing computational work mainly appears in the *methods* and *results* sections of scientific publications, not in the *discussion* section where mechanisms are typically elucidated.

There have been several areas of computational research regarding causation and causal inference, but they have a quite different character from what Craver, Kahneman, and Peirce are discussing. Statistical methods, such as structural equation modeling (e.g. [9]), Neyman–Rubin causal modeling[20], or Judea Pearl's work (e.g.[15]) model unobserved data that would be relevant to distinguishing causal direction in correlations, but are not attempts to solve the broader mechanistic inference problem. Causal inference in epidemiology (e.g. Hill's criteria for causation[10]) cannot be straightforwardly translated into an automated method. Historic AI efforts in explanation generation (e.g., [21]) have left little trace in contemporary research. As an example of how important this task is to biomedicine, consider the legal requirements to submit an Investigative New Drug (IND) application to the US Food and Drug Administration[1], which require the documentation of a mechanism of action of the compound. The law requires that this mechanism "must be discussed at various levels, including the cellular, receptor, or membrane level, tissue, the physiologic system level (target organ), and the whole body level, depending on what is known." This concept of mechanism, which involves describing multiple


*KBCOM, January 2018, Los Angeles, CA, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] 21 CFR 201.57(c)(13)(i)(A).

activities and interactions, cannot be satisfied with a statistical correlation, even one where the direction of causality is specified.

## 2 KNOWLEDGE-BASED BIOMEDICAL DATA SCIENCE

Although there is relatively little support for inference of mechanisms thus far in computational biomedicine, there are some current approaches that are relevant to the task.[2] While no knowledge-based computer system has repeatedly generated important biomedical hypotheses *de novo*, promising proof-of-concept systems include systems to generate hypotheses from the literature[22] and those aimed at hypothesis generation or refinement from data (e.g. [17] and [4]), as well as mixed initiative human-computer hypothesis generation[13]. Although it remains aspirational, the synthesis of computational simulation with knowledge-based generation and refinement of hypotheses has received substantial interest from DARPA[24].

While systems to generate mechanistic hypotheses are rare, a great deal of the infrastructure required for knowledge representation and reasoning in the area already exists. Well developed, community curated ontological resources cover a great deal of the relevant conceptual ground, such as the Gene Ontology[6] or the Ontology of Biomedical Investigations (to describe experiments)[2], as well as the Open Biomedical Ontologies Foundry[23] project to coordinate them. Sophisticated efforts to integrate multiple curated biomedical databases into unified knowledge-bases have made substantial progress, e.g. Bio2RDF[3] and KaBOB[14]. Recently, for example, an effort to do scientifically significant inference based on such a knowledge-base, using a combination of symbolic and neural network approaches, appears to have been successful hypothesizing about factors potentially relevant to generating mechanistic hypotheses, e.g. simultaneously generating hypotheses about protein function, candidate genes of diseases, protein-protein interactions, and drug target relations[1]. Another recent knowledge-based approach generated mechanistically-grounded drug repurposing hypotheses[11]. A long-standing effort to develop a fully automated "robot scientist" recently began using abductive logic programming to generate hypotheses in metabolic network modeling[19].

## 3 THE CHALLENGE

Automated generation of mechanistic explanations for experimentally observed phenomena in contemporary molecular biology would address an increasingly acute need in biomedical research. Genome-scale experiments (not only in genetics, but in proteomics, metabolomics, transcriptomics, etc.) produce data about many thousands of molecules, polymorphisms and other relevant entities involved in many important disease areas (as well as in normal biology). A great deal of information about each of these entities is available from a large and diverse collection of databases; even more is in the biomedical literature. It has been demonstrated repeatedly that biologists bias their interpretations toward the familiar (see, e.g. [18]), possibly causing many potentially important findings to be overlooked. There is a clear desire in the biomedical community to address these acknowledged issues through better computational tools, but, as noted above, few useful ones have thus far been created.

The mechanistic inference problem is inherently knowledge-based. Explanations of biological phenomenon are typically expressed in the scientific literature in symbolic form (in contrast,say, to the mathematical form of explanations found in in physics). The ontological foundations, and perhaps the computationally represented knowledge necessary to generate hypothesized mechanistic accounts of that biomedical data are already available. What is missing is a theoretical account and practical implementation of the inference process itself. Note that although explanations themselves are symbolic, mixed symbolic and nonsymbolic methods (such as [1] and [11]) show great potential, perhaps echoing Kahneman's description of the importance of both system 1 and system 2 in the psychology of causal explanation.

The recent philosophical literature provides descriptions of the process of producing, evaluating and refining mechanistic accounts in science[7], as well as a set of specific examples and a possibly useful diagrammatic representation of them[8]. However, this philosophical description is in places vague and admittedly incomplete. A computational implementation would demonstrate the sufficiency of a specific approach to account for mechanistic theorizing in biomedicine. Automated mechanistic inference for biomedical research would also solve a critical problem in biomedical research itself, which is struggling with data that is too big and too rich to analyze completely in traditional ways.

Knowledge base construction, reasoning, and mining may be a crucial research area in which this research agenda could be advanced. Clearly defining the representations and computations necessary for the task of mechanistic explanation is part of the challenge. While existing biomedical ontologies would seem to be a natural point for grounding biomedical explanations, it is not clear if they are complete for this purpose. Judging by the mechanisms of action on FDA approved drugs, relationships among the entities that participate in an explanation span mathematical, physical, chemical, biological and medical ones that may not all be well described by current ontological resources. Furthermore, evidence suggests that the soundness and quality of an explanation is related not only to how well it accounts for the to-be-explained inputs, but also how it relates to many other "known" explanations. By bringing the tools and techniques of symbolic AI (perhaps in conjunction with other tools and techniques) to address this problem, it may be possible to create novel computational methods with far reaching implications for both biomedical research and our understanding of, in Peirce's words, "the most wonderful thing in our constitution."

[2]This account is condensed from my recent position paper in the IOS journal *Data Science*[12].

## REFERENCES

[1] M. Alshahrani, M. A. Khan, O. Maddouri, A. R. Kinjo, N. Queralt-Rosinach, and R. Hoehndorf. 2017. Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics* 33, 17 (Sep 2017), 2723–2730.

[2] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, L. Fan, J. Fostel, G. Fragoso, F. Gibson, A. Gonzalez-Beltran, M. A. Haendel, Y. He, M. Heiskanen, T. Hernandez-Boussard, M. Jensen, Y. Lin, A. L. Lister, P. Lord, J. Malone, E. Manduchi, M. McGee, N. Morrison, J. A. Overton, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, D. Schober, B. Smith, L. N. Soldatova, C. J. Stoeckert, C. F. Taylor, C. Torniai, J. A. Turner, R. Vita, P. L. Whetzel, and J. Zheng. 2016. The Ontology for Biomedical Investigations. *PLoS ONE* 11, 4 (2016), e0154556.

[3] F. Belleau, M. A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. 2008. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 41, 5 (Oct 2008), 706–716.

[4] Alison Callahan, Michel Dumontier, and Nigam H. Shah. 2011. HyQue: evaluating hypotheses using Semantic Web technologies. *Journal of Biomedical Semantics* 2, 2 (17 May 2011), S3. https://doi.org/10.1186/2041-1480-2-S2-S3

[5] Michelle M. Chouinard, P. L. Harris, and Michael P. Maratsos. 2007. Children's Questions: A Mechanism for Cognitive Development. *Monographs of the Society for Research in Child Development* 72, 1 (2007), i–129. http://www.jstor.org/stable/30163594

[6] The Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D1 (Jan 2017), D331–D338.

[7] C.F. Craver and L. Darden. 2013. *In Search of Mechanisms: Discoveries across the Life Sciences.* University of Chicago Press, Chicago, IL. https://books.google.com/books?id=ESI3AAAAQBAJ

[8] Lindley Darden, Lipika R. Pal, Kunal Kundu, and John Moult. 2017. The Product Guides the Process: Discovering Disease Mechanisms. (July 2017). http://philsci-archive.pitt.edu/13176/

[9] J.B. Grace. 2006. *Structural Equation Modeling and Natural Systems.* Cambridge University Press, Cambridge, UK. https://books.google.com/books?id=1suuMOChHWcC

[10] Sir Austin Bradford Hill. 1965. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* 58, 5 (1965), 295–300. https://doi.org/10.1177/003591576505800503 arXiv:https://doi.org/10.1177/003591576505800503 PMID: 14283879.

[11] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* 6, e26726 (2017). https://doi.org/10.7554/eLife.26726

[12] Lawrence Hunter. 2017. Knowledge-based biomedical data science. *Data Science* 1 (2017), 1–7. https://doi.org/10.3233/DS-170001

[13] S. M. Leach, H. Tipney, W. Feng, W. A. Baumgartner, P. Kasliwal, R. P. Schuyler, T. Williams, R. A. Spritz, and L. Hunter. 2009. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput. Biol.* 5, 3 (Mar 2009), e1000215.

[14] K. M. Livingston, M. Bada, W. A. Baumgartner, and L. E. Hunter. 2015. KaBOB: ontology-based semantic integration of biomedical databases. *BMC Bioinformatics* 16 (Apr 2015), 126.

[15] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statist. Surv.* 3 (2009), 96–146. https://doi.org/10.1214/09-SS057

[16] C.S. Peirce and P.A. Turrisi. 1997. *Pragmatism as a Principle and Method of Right Thinking: The 1903 Harvard Lectures on Pragmatism.* State University of New York Press, New York, NY. https://books.google.com/books?id=_TUqldjTO80C

[17] S. A. Racunas, N. H. Shah, I. Albert, and N. V. Fedoroff. 2004. HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics* 20 Suppl 1 (Aug 2004), i257–264.

[18] R. Rodriguez-Esteban and X. Jiang. 2017. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Med Genomics* 10, 1 (Oct 2017), 59.

[19] Robert Rozanski, Stefano Bragaglia, Oliver Ray, and Ross King. 2015. *Automating the Development of Metabolic Network Models.* Springer International Publishing, Cham, 145–156. https://doi.org/10.1007/978-3-319-23401-4_13

[20] Jasjeet S. Sekhon. 2008. The Neyman-Rubin Model of Causal Inference and Estimation Via Matching Methods. In *The Oxford Handbook of Political Methodology.* Oxford University Press, Oxford, UK. http://sekhon.berkeley.edu/papers/SekhonOxfordHandbook.pdf

[21] Roger C. Schank. 1986. *Explanation Patterns: Understanding Mechanical and Creatively.* L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

[22] N. R. Smalheiser, V. I. Torvik, and W. Zhou. 2009. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Comput Methods Programs Biomed* 94, 2 (May 2009), 190–197.

[23] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 11 (Nov 2007), 1251–1255.

[24] J. You. 2015. Artificial intelligence. DARPA sets out to automate research. *Science* 347, 6221 (Jan 2015), 465.