DIVERSE DISCIPLINES, ONE COMMUNITY
BiomedicalComputation REVIEW
Published by the Mobilize Center, an NIH Big Data to Knowledge Center of Excellence

Nonprofit Org.
U.S. Postage Paid
Permit No. 28
Palo Alto, CA

Stanford University
318 Campus Drive
Clark Center Room W352
Stanford, CA 94305-5444

SeeingScience

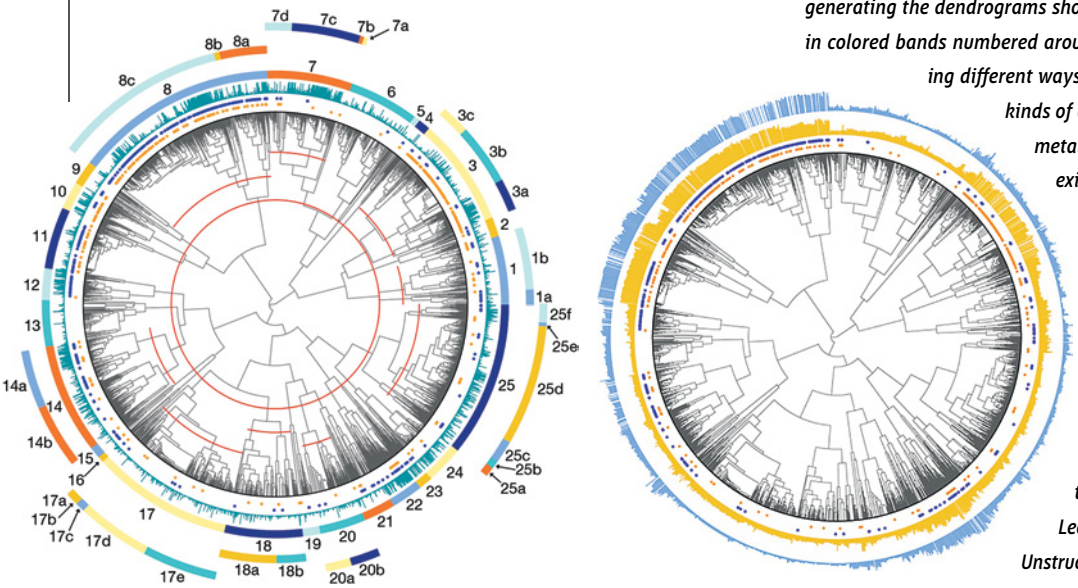
BY KATHARINE MILLER

AUTOMATING LITERATURE SURVEILLANCE

Today, if researchers want to study complex relationships among genes, diseases and drugs, they have to hope that human curators have read the scientific literature, extracted the relevant information, and put it in a database. "It would be a lot more efficient if computers could perform that surveillance of the literature for us," says **Beth Percha**, a graduate student working with **Russ Altman, PhD**, at Stanford University.

In recent work, Percha and Altman made steps toward that goal, effectively extracting drug-gene relationships from the literature and clustering them in ways that proved meaningful (see dendrogram caption). Percha is also applying the same method to other situations such as gene-disease and disease-drug relationships. Ultimately, she'd like to be able predict drug-drug interactions based on drug-gene relationships automatically extracted from the literature.

"The dendrogram is pretty and it's a good sanity check because it reproduces knowledge we already have," Percha says. "But what's exciting is to be able to discover new relationships from the literature quickly, cheaply and without a ton of human effort." □



For 3,514 drug-gene pairs that co-occur at least five times in Medline sentences, each represented as a black dot at the edge of the black circle, Percha and Altman used a novel algorithm that recognizes when two such pairs share a similar relationship. They then used a clustering algorithm to connect drug-gene pairs that act similarly, generating the dendrograms shown here. The clusters revealed 25 "themes"(shown in colored bands numbered around the outside of the circle at far left), representing different ways that drugs interact with genes, such as by various kinds of activation (13-14), inhibition (8, 11) or an effect on metabolism (3). These concurred with information from existing knowledgebases including DrugBank (blue dots) and PharmGKB (orange dots) while also discovering many new relationships that likely should be included in those knowledgebases, as shown in the smaller dendrogram at near left (blue spikes predict drug-target relationships that should be in DrugBank, and orange spikes predict relationships that should be in PharmGKB because mutations in the gene likely impact a person's response to the drug). Reprinted from B Percha, RB Altman, Learning the Structure of Biomedical Relationships from Unstructured Text, PLoS Comp Biol, *11(7):e1004216 (2015).*