Portfolio Analysis for Basic Biomedical Research using NIH Maps: Lessons Learned and Future Possibilities

National Institute of General Medical Sciences

National Institute of Neurological Disorders and Stroke

Lisa Dunbar (NIGMS), Edmund M. Talley (NINDS), Catherine Lewis (NIGMS)

Abstract: NIH is a worldwide leader in supporting fundamental biomedical research, which lays the foundation for disease diagnosis, treatment and prevention. Evaluating this support requires analysis of diverse investments covering the spectrum of biomedical research, to assess balance, trends, outcomes, and relationships to other agencies. In this poster, we show the use of NIH Maps, an online text-mining tool, to analyze research portfolios in NIGMS, a $2.4 billion institute whose mission is devoted to supporting a diverse range of basic research from basic cell biology to pharmacology. The tool was used for discovery and classification of NIGMS research based on grant information including scientific titles and abstracts. We show how this information is complementary to the NIH disease-based research classification (the RCDC system), and how it has provided a comprehensive approach for understanding portfolios, and their changes over time, independent of administrative research programs. This process included expert validation of the automated categories, as well as consideration of the strengths and weaknesses of the approach and the implications for future text mining efforts.

Topic modeling refers to a strategy that identifies informative words in a set of documents and analyzes their co-occurrence to arrive at "topics" or categories that are independent of keywords. Certain groups of words co-occur frequently and are defined as topics; the number of words in each topic and the number of topics set the granularity of the analysis. Note that a particular word can be represented in multiple topics (e.g., microtubules in cell division, cilia or intracellular transport). The topic modeling strategy extracts groups of words from a set of documents and uses Bayesian statistical techniques to infer the underlying distribution of topics. The most relevant topics for a particular document are calculated by the occurrence of the words within each topic and each document can be represented by several topics. Thus, one grant can be represented in several topics categories; each of these is weighted proportionally to represent the fraction of the project devoted to that topic. The dollars that are associated with each of the subtopics within a grant are weighted as well, thereby giving a more accurate distribution of the dollars than would result from counting the entire grant as contributing to only one topic.

Example of Topic Modeling by Latent Drichlet Allocation*

1. I like to eat broccoli and bananas.
2. I ate a banana and spinach smoothie for breakfast.
3. Chinchillas and kittens are cute.
4. My sister adopted a kitten yesterday.
5. Look at this cute hamster munching on a piece of broccoli.

Sentences 1 and 2: 100% Topic A
Sentences 3 and 4: 100% Topic B
Sentence 5: 60% Topic A, 40% Topic B

Topic A: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … (at which point, you could interpret topic A to be about food)
Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals)

Allotting grant dollars to topics

2R01GM026875-35 Biochemical Studies of Mitosis
Highly integrative cellular processes-those involving many components with interactions

Topic Probability	Topic Abbrv
0.1554	actin_cytoskeleton, actin_filaments, actin_binding
0.0960	cell_cycle, cyclin, cell_proliferation, cyclin_dep
0.0876	proteins, purification, tagged, purify, mass_spect
0.0678	zebrafish, embryos, vertebrate, embryonic, develop
0.0621	networks, systems_biology, computational, cellular
0.0480	molecular, cellular, animal, disease, human, mouse
0.0424	membrane, lipid_bilayer, membrane_proteins, vesicl
0.0367	spatial, distribution, map, location, regions, tem
0.0282	mitosis, mitotic_spindle, chromosome_segregation,
0.0282	assessment, quantitative, measurement, measures, c
0.0282	fret, fluorescence, imaging, living_cells, single_
0.0254	shape, structure, size, organization, _3d, archite
0.0226	history, medicine, century, book, pioneer, science
0.0169	arthritis, rheumatoid, arthritis, joint, synovial,
0.0169	myelin, oligodendrocytes, myelination, remyelinati
0.0113	bacterial, cell_division, assembly, spores, e_coli
0.0113	compounds, analogs, potent, derivatives, agents, e
0.0085	pcr, species, dna, sequence, culture, amplificatio
0.0085	egf_receptor, epidermal_growth_factor_receptor_by
0.0056	metastasis, invasion, tumor_cells, cancer_cells, p
0.0056	dosing, administration, administered, oral, pharma
0.0056	researchers, interdisciplinary, science, scientifi

actin_cytoskeleton, actin_filaments, actin_binding
45.8 %
$421,791

cell_cycle, cyclin, cell_proliferation, cyclin_dep
28.3%
$260,743

proteins, purification, tagged, purify, mass_spect
25.8%
$237,737

Total Cost: $920,271

The data presented in the NIH Maps analysis exclude what are considered to be specialized or resource projects, namely pre- and post-doctoral training grants, the IDeA awards transferred from NCRR at the start of FY12, and projects supported through the former Minority Opportunities in Research Division. Topics that were judged to be outside the scientific content of research grants were excluded.

Validation and Clustering Topics into Scientific Categories

Topics

actin dynamics & processes
rho, actin, signaling, rhoa, cytoskeletal, actin_cytoskeleton, cdc42, rac, gtpase, rac1, cytoskeleto
myosin, actin, muscle, force, contractile, troponin, contraction, filaments, thin_filament,
actin_cytoskeleton, actin_filaments, actin_binding, cytoskeleton, dynamics, actin_polymerizati
cell_migration, cell_motility, movement, chemotaxis, adhesion, cellular, edge, directional,
mechanical, forces, cell, mechanics, mechanotransduction, stress, stretch, physical, loading,
barrier, tight_junction, cadherin, cell_cell, adherens, integrity, adhesion, junctional, epithelial,
integrin, cell_adhesion, focal_adhesion_kinase, signaling, extracellular_matrix, binding, ecm,

While the topic modeling process provides valuable granularity by using over 700 topics to describe NIH research, there was a need for broader categories to carry out higher level analyses. For NIGMS, ~300 topics were found to account for 90% of NIGMS research spending; these 'GM topics' and the total costs of each award were assigned as a percentage to the top three topics in each project. The dollars associated with each topic were weighted accordingly. These topics were then clustered using an algorithm based on co-occurrence, which resulted in 124 categories. These categories were further reduced to 50 based on review by NIGMS staff.

4 NIGMS Spending in Major Categories as a Percentage of NIH-wide Spending

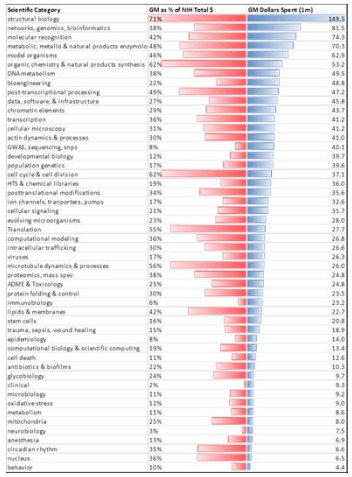

Figure 4 shows the clusters accounting for top 90% of the NIGMS investment listed in descending order by NIGMS dollars. 'Structural biology' is at the top of the list, with a total cost investment by NIGMS of $149.5 million as seen in blue on the right. Other significant investments are 'networks, genomics, bioinformatics', 'molecular recognition', and 'metabolic, metallo & natural products enzymology'.
This tabulation also shows what percentage of the NIH total investment for each topic is funded with NIGMS dollars (red bars in Figure III). For 'structural biology', the NIGMS investment represents 71% of the NIH total. Other scientific areas where NIGMS has made significant investments are: 'organic chemistry & natural product synthesis' at 62%, 'cell cycle & cell division' at 62%, 'translation' at 55%, and 'microtubule dynamics & processes' at 56%.

5 Distribution of Categories Across NIGMS Scientific Divisions

Scientific Category	B	C	G	P
structural biology	15.32%	72.58%	5.12%	6.99%
networks, genomics, bioinformatics	43.39%	31.99%	17.53%	7.09%
molecular recognition	3.87%	58.65%	13.60%	23.88%
metabolic, metallo & natural products enzymology	0.92%	17.37%	8.91%	72.90%
model organisms	2.63%	17.66%	73.54%	6.18%
organic chemistry & natural products synthesis	0.54%	3.99%	1.48%	93.99%
DNA metabolism	0.13%	10.51%	78.70%	10.66%
data, software, & infrastructure	68.44%	16.39%	8.60%	6.57%
post-transcriptional processing	8.25%	20.55%	67.55%	5.65%
bioengineering	42.93%	26.86%	7.14%	23.07%
chromatin elements	1.20%	11.34%	78.89%	8.57%
GWAS, sequencing, snps	7.94%	13.72%	57.34%	21.00%
actin dynamics & processes	5.53%	71.00%	16.90%	6.56%
population genetics	11.53%	8.16%	70.55%	11.76%
transcription	1.40%	20.47%	69.70%	8.43%
developmental biology	3.33%	16.71%	70.27%	9.68%
cellular microscopy	26.48%	61.26%	5.60%	6.67%
cell cycle & cell division	1.71%	18.63%	57.03%	2.63%
posttranslational modifications	6.54%	23.82%	34.18%	35.47%
cellular signaling	4.02%	30.45%	21.64%	43.89%
HTS & chemical libraries	5.77%	11.99%	4.77%	57.48%
ion channels, transporters, pumps	6.93%	63.95%	6.60%	22.52%
computational modeling	70.74%	11.00%	13.60%	4.66%
evolving microorganisms	8.00%	29.32%	51.72%	10.96%
ADME & Toxicology	2.58%	3.69%	2.64%	91.09%
translation	0.00%	26.02%	63.55%	10.43%
intracellular trafficking	6.69%	81.63%	9.39%	8.29%
viruses	7.50%	83.26%	3.95%	5.29%
protein folding & control	2.09%	72.67%	9.23%	16.01%
microtubule dynamics & processes	1.31%	90.24%	7.46%	1.00%
trauma, sepsis, wound healing	1.10%	2.99%	1.00%	94.92%
lipids & membranes	5.91%	72.54%	3.39%	18.17%
proteomics, mass spec	54.14%	33.47%	5.44%	6.95%
immunobiology	2.05%	14.94%	11.47%	71.55%
stem cells	6.02%	7.61%	80.58%	5.79%
computational biology & scientific computing	70.28%	11.61%	14.75%	3.36%
epidemiology	63.26%	9.46%	16.94%	9.86%
cell death	22.01%	8.78%	40.31%	28.90%
antibiotics & biofilms	16.26%	14.06%	23.32%	46.36%
clinical	16.34%	22.49%	17.89%	43.27%
glycobiology	13.85%	13.26%	1.07%	71.82%
metabolism	7.19%	35.19%	22.06%	35.55%
microbiology	12.20%	4.78%	65.26%	17.78%
oxidative stress	0.00%	19.09%	11.59%	69.31%
anesthesia	0.00%	3.44%	6.68%	89.89%
mitochondria	0.00%	35.00%	14.15%	50.85%
nucleus	2.35%	59.29%	29.61%	8.74%
circadian rhythm	0.00%	7.06%	86.47%	6.47%
neurobiology	7.35%	34.12%	46.43%	12.10%
behavior	8.37%	1.90%	75.96%	13.78%
Relative Distribution	**13.68%**	**33.69%**	**29.02%**	**23.61%**

Figure 5 shows the same list of top NIGMS investments, but as they are represented in each division. B: Biomedical Technology, Bioinformatics, and Computational Biology C: Cell Biology and Biophysics G: Genetics and Developmental Biology P: Pharmacology, Physiology and Biological Chemistry. One can see that 72% of 'structural biology' is funded by CBB, but that the other 28% is spread across the other divisions. This was not at all apparent from looking at the program areas and PCC codes alone. From the traditional description, one may have concluded that NIGMS invested only $70 - 80 million in structural biology, mostly through programs in CBB. For other topics such as posttranslational modifications, the investment is spread somewhat evenly across three out of four divisions.

6 Topics within a Category Across NIGMS Scientific Divisions

	B	C	G	P
⊟structural biology	15.32%	72.58%	5.12%	6.99%
crystallization, crystals, x_ray_crystallography, protein, structure_determination, structural, m	17.20%	81.61%	0.05%	0.94%
structural, complexes, biochemical, molecular, structure, x_ray_crystallography, proteins, dom	6.01%	69.28%	11.95%	12.77%
computational, simulations, molecular_dynamics, protein, structures, molecular, structural, m	19.70%	69.20%	2.49%	8.61%
nuclear_magnetic_resonance, spin, structure, nmr_spectroscopy, labeled, _15n, _13c, solution	0.60%	90.05%	1.80%	8.15%
shape, structure, size, organization, _3d, architecture, assembly, shape, arrangement, organize	18.50%	56.51%	22.50%	2.48%
nuclear_magnetic_resonance, spectrometer, instrument, mhz, electron_paramagnetic_resonanc	82.28%	17.22%	0.00%	0.50%
spectroscopy, raman_spectroscopy, spectral, sers, optical, infrared, scattering, laser, vibration	56.52%	34.91%	0.00%	8.57%

Figure 6 shows more detailed information on the differences among the divisions within the structural biology category. For example, Biomedical Technology, Bioinformatics, and Computational Biology (B) division focuses on NMR instrumentation, while projects that use NMR to determine protein structure are concentrated in the Cell Biology and Biophysics division (C).

7 Detecting Emerging Topics and Clusters with NIH Maps Data

Clusters with >59% invested in either Projects in years 1-10 or Projects in years >10 are highlighted

Cluster Name	>10 Yrs	< 10 Yrs
metabolic, metallo & natural products enzymology	62%	38%
DNA metabolism	66%	34%
transcription	64%	36%
translation	60%	40%
intracellular trafficking	66%	34%
protein folding & control	66%	34%
microtubule dynamics & processes	60%	40%
lipids & membranes	61%	39%
oxidative stress	63%	37%
circadian rhythm	65%	35%

Cluster Name	>10 Yrs	< 10 Yrs
networks, genomics, bioinformatics	35%	65%
bioengineering	39%	61%
population genetics	35%	65%
computational modeling	39%	61%
stem cells	15%	85%
computational biology & scientific computing	34%	66%
epidemiology	21%	79%
clinical	27%	73%
microbiology	21%	79%
behavior	18%	82%

post-transcriptional processing	**49.8%**	**50.2%**
rna_processing, rna_binding, processing, dead_box, rna, mrna, rna, assembly, proteins, mase, expor	61.9%	38.1%
rna_splicing, alternative_splicing, pre_mrna, rna, exon, intron, splice, mrna, transcripts, proteins	55.7%	44.3%
noncoding_rna, ncrnas, transcripts, antisense, noncoding, mrna, protein_coding, small_rna, gene_expr	32.7%	67.3%
rna, rna_interference, small_rna, silencing, sirna, gene_silencing, dicer, mirna, dsrna, editing, ar	43.3%	56.7%
microrna, mirna, mir, microrna_mirnas, mrna, targets, genes, noncoding_rna, gene_expression, _3_utr,	37.7%	62.3%

stem cells	**15.0%**	**85.0%**
pluripotent, stem_cells, differentiation, ips_cells, cells, human, reprogramming, ips, es_cells, plu	7.0%	93.0%
regeneration, stem_cells, regenerative, tissue_regeneration, regenerative_medicine, repair, adult, p	33.9%	66.1%
differentiation, stem_cell, tissue_engineering, mesenchymal_stem_cells, _3d, mscs, tissue, cells, mi	11.3%	88.7%
stem_cells, progenitor_cells, differentiation, lineage, population, markers, stem, self_renewal, adu	73.8%	26.2%

In Figure 7 we looked to see if newer projects might be indicative of scientific categories that are growing/emerging. Projects are grouped into "newer" or "older" categories based on the project year date of <10 years. Categories with a significant history of NIGMS investment such as enzymology, DNA metabolism, transcription, and translation have a greater number of established projects relative to other categories such as networks, genomics, and informatics, population genetics, and microbiology. A higher percentage of newer projects within a category may also reflect investments from targeted RFAs and/or PARs. Within categories particular topics show a higher percentage of newer grants, which appears to reflect the current state of the field. For example, newer projects emphasize topics on non-coding RNAs and microRNAs in post-transcriptional processing, while there is an emphasis on induced pluripotent stem cells and differentiation for newer grants in the stem cell field.

Conclusions:

- NIH Maps is based on a quantitative approach and is viewer independent, transparent, adjustable, and reproducible.
- NIH Maps topics generally represented the scientific content of NIGMS research grants quite well.
- NIH Maps NIH Maps data provides the ability to proportionally allot dollars, which is a major advantage over other analysis tools.
- <700 topics provides appropriate granularity
- Determining how to combine topics into a reasonable number of scientific categories relevant to the IC for reporting purposes requires significant input from subject matter experts

References:
*http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation

Talley, E. M. et al (2011). Database of NIH grants using machine-learned categories and graphical clustering. Nature Methods, 8, 443-444.