Supplemental Material: Annu. Rev. Cell. Dev. Biol. 2010. 26: 721-744
doi: 10.1146/annurev-cellbio-100109-104122
A Decade of Systems Biology
Chuang, Hofree, and Ideker

Supplemental Methods Section

Article abstracts published from 2000 to 2009 were downloaded from PubMed using a query for the MeSH term “Systems Biology” or MeSH term “Computational Biology” and any of the following words: systems, network, pathway, interaction, model. The publication graph follows the same trend as for the MeSH term “Systems Biology” (Supplementary Figure 1**)**. The Mesh term “Computational Biology” is the superset of the MeSH term “Systems Biology.” Manual inspection of a sample of articles recovered in the above-described way showed few false positives (< 0.05).

Overall, 18,516 article abstracts published from 2000-2009 are included in the analysis.

We use the “tm” R package (Feinerer et al. 2008) to extract word occurrence bag-of-words representations for the article abstracts. Articles are tokenized to words, numbers are filtered, stop-words are filtered, single character words discarded, and a Snowball stemming procedure is applied to convert words into word roots. To limit the dictionary to a manageable size, words occurring less than five times are discarded, leaving a dictionary of 15,721 words. The word document matrix contained a total of 1,415,926 non-zero values.

A Latent Dirichlet Allocations (Blei et al. 2003) topic model was trained over the document set with 50 topics and a uniform alpha prior on the words of 1 and a uniform beta prior on the documents of 0.1.

Topics were manually labeled on the basis of the top 20 most common keywords recovered from each topic. Results are visualized by drawing word sizes proportional to the topic weight in the corpus. Trends of topic frequency across the two sets are calculated by fitting a linear trend line to each topic distribution across the years (Griffiths & Steyvers 2004). The topics are visualized in 2D using an Isomap embedding (Tenenbaum et al. 2000) based on Kullback-Leibler distance as a measure of topic similarity, i.e., topics close in the map share similar words.

Following are the top most common words occurring in each topic. Word size is proportional to the probability of the words occurring in the given topic. Topic assignment is given as the first term when the topic was used in Figure 1*a*. Fifteen of the topics result in word sets that are uninformative as they describe either a methodology (e.g., topic 1) or a broad concept (e.g., topic 5); these are omitted from the final topic map displayed in Figure 1*a*.

(1) – model, compute, experiment, develop, approach, simulate, predict, describe, construct, base,
(2) Protein structure – structure, sequence, protein, align, predict, fold, model, base, similar, homolog,
(3) Animal model organisms – rat, mice, liver, increase, level, change, induce, response, significant, control,

(4) Transcriptional regulation − transcript, regulate, factor, regulatory, promote, bind, motif, site, element, specify,

(5) − research, develop, bioinformatics, inform, science, community, system, project, medic, technology,

(6) Disease association − disease, human, study, association, age, risk, molecular, disorder, factor, understand,

(7) – system, biology, complex, process, level, approach, understand, molecular, engine, organ,

8) Stress response − response, stress, oxidative, mitochondria, induce, level, regulate, involve, adapt, mechanism,

(9) – data, integrate, set, inform, experiment, large, biology, generate, experiment, scale,

(10) Proteomics – protein, proteome, identify, analysis, gel, dimension, express, differential, electrophoresis, spot,

(11) Signal transduction − active, kinas, phosphoryl, signal, receptor, regulate, induce, mediate, factor, apoptosis,

(12) Genomics − DNA, sequence, chromosome, clone, recombine, region, library, length, element, repeat,

(13) Microarrays − array, throughput, base, microarray, system, method, experiment, assay, large, technology,

(14) − membrane, local, transport, protein, lipid, nuclear, assembly, associate, cell, subcellular,

(15) Stem cell – cell, differential, line, cycle, express, growth, culture, stem, cellular, vitro,

(16) Biological annotation − inform, base, ontology, knowledge, term, extract, literature, process, annotate, relate,

(17) − type, mutate, beta, mutant, alpha, result, level, wild, ii, indicate,

(18) Genomics – genome, sequence, organ, complete, annotate, gene, map, compare, species, provide,

(19) Neurobiology − active, receptor, effect, channel, ca, response, depend, inhibit, mediate, induce,

(20) Protein interactions – protein, interact, complex, yeast, cerevisiae, partner, map, saccharomyces, pair,

(21) Developmental biology − develop, organ, pattern, study, drosophila, dure, development, stage, animal, elegans,

(22) Machine learning classification – predict, method, feature, set, base, classify, accuracy, perform, learn, classify,

(23) − correlation, select, posit, rate, relate, result, observe, pair, pattern, found,

(24) Networks biology – network, module, regulatory, connect, biology, topology, scale, infer, graph, reconstruct,

(25) − recent, review, discuss, technology, develop, application, research, advance, field, current,

(26) Comparative genomics − human, mouse, gene, identify, conserve, region, splice, mammalian, alternative, genome,

(27) Protein binding − bind, domain, site, acid, amino, ligand, peptide, protein, residue, receptor,

(28) Transcriptomics − RNA, target, translate, mRNA, identify, miRNA, specify, function, novel, post,

(29) Clustering methods − algorithm, compute, cluster, optimal, method, base, efficiency, result, program, set,

(30) Pathway biology – pathway, signal, cellular, process, component, response, involve, mechanism, regulate, molecular,

(31) − increase, muscle, blood, flow, insulin, heart, body, diabetes, significant, dure,

(32) − brain, neuron, study, relate, al, et, suggest, associate, neural, central,

(33) Protein docking − structure, energy, conform, molecular, calculate, chain, dock, free, atom, bond,

(34) Dynamic modeling − dynamic, time, control, change, mechanism, robust, behavior, conditional, phase, transit,

(35) Evolutionary biology − evolution, species, evolutionary, family, phylogenetic, tree, duplicate, divers, eukaryote, conserve,

(36) Mass spectrometry − ms, mass, peptide, spectrometry, identify, proteome, sample, separate, label, chromatography,

(37) Gene expression – gene, express, microarray, analysis, identify, profile, cluster, level, differential, transcript,

(38) Plant biology – plant, Arabidopsis, rice, root, thaliana, develop, seed, light, relate, species,

(39) Drug design – drug, target, chemic, discovery, develop, compound, screen, potential, novel, approach,

(40) – function, import, study, mani, provide, specify, role, character, essential, however,

(41) – active, enzyme, substrate, modify, inhibitor, protease, methyl, specify, involve, catalyst,

(42) Microscopy – image, surface, fluoresce, single, link, cross, detect, direct, technique, resolute,

(43) – analysis, study, profile, approach, metabolism, combine, analysis, metabolite, tool, provide,

(44) Cancer research – cancer, tumor, patient, biomarker, clinic, specify, tissue, breast, identify, normal,

(45) Dynamic modeling – model, parameter, simulate, dynamic, time, kinetic, reaction, system, biochemist, equate,

(46) Association studies – genet, phenotype, population, associate, study, variant, map, genotype, single, polymorph,

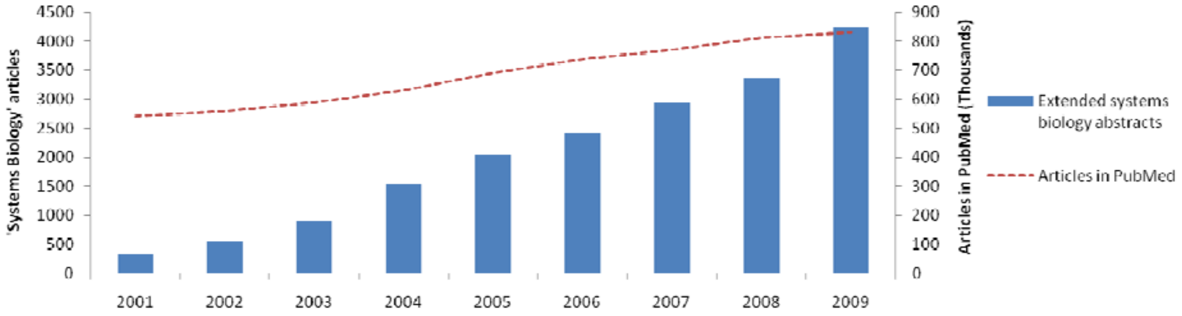
47) Virology – host, infect, pathogen, immune, strain, bacteria, virus, bacteria, resist, response,

(48) Metabolic pathways – metabolic, product, pathway, acid, coli, enzyme, flux, growth, reaction, metabolite,

(49) – method, statist, estimate, data, result, measure, test, error, propos, value,

(50) Online tools – database, avail, tool, http, web, software, data, user, access, inform,

Supplemental Figure 1 – PubMed articles according to MeSH term "Computational Biology" and word filter



References

Blei D, Ng A, Jordan M. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993-1022

Feinerer I, Hornik K, Meyer D. 2008. Text mining infrastructure in R. *J. Stat. Software* 25:1-54

Griffiths TL, Steyvers M. 2004. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* 101:5228

Tenenbaum JB, Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319