Mining MEDLINE with HAPI (UCSD) & Arrowsmith (UIC)

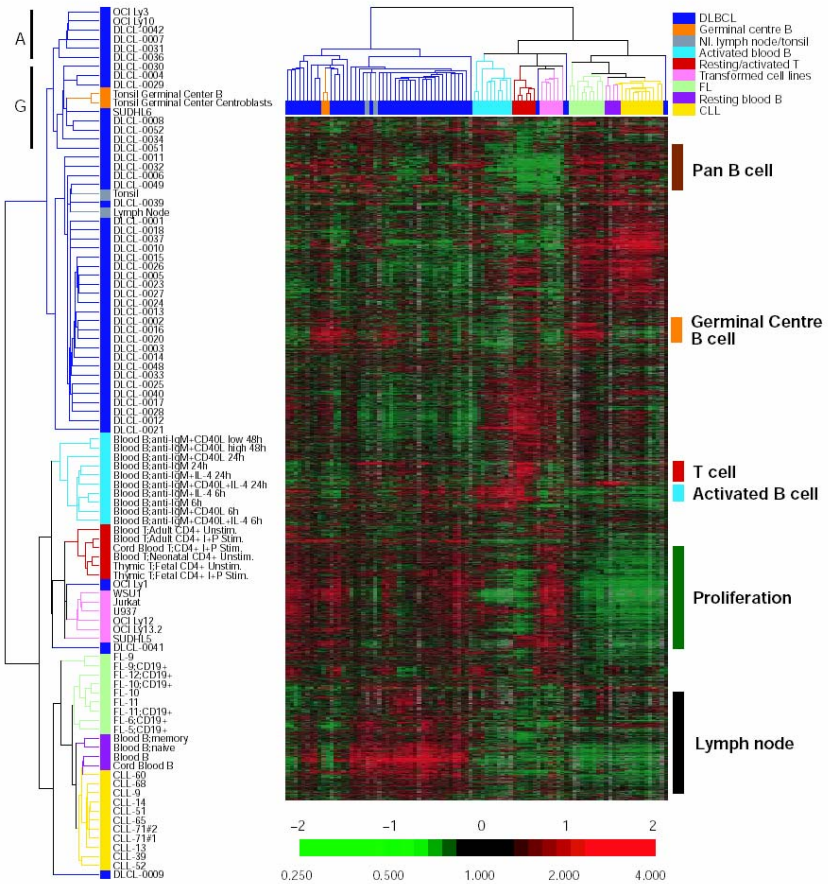
Ramin Homayouni, Ph.D.
Department of Neurology
University of Tennessee Health Science Center

Genomics and Bioinformatics
Center of Excellence

CITR

Center *for* Neurobiology *of* Brain Diseases

Gene Expression Profiling



Alizadeh, et al., (2000) Nature 403:503.

Now What?

Products of the National Library of Medicine

- Databases
  - GenBank, UniGene, LocusLink
  - **MEDLINE**
  - OMIM
- Services
  - HealthSTAR
  - Health Services Research Projects in Progress
  - HSTAT
- Vocabulary
  - **Medical Subject Headings (MeSH)**
  - NLM Classification
  - Unified Medical language Systems

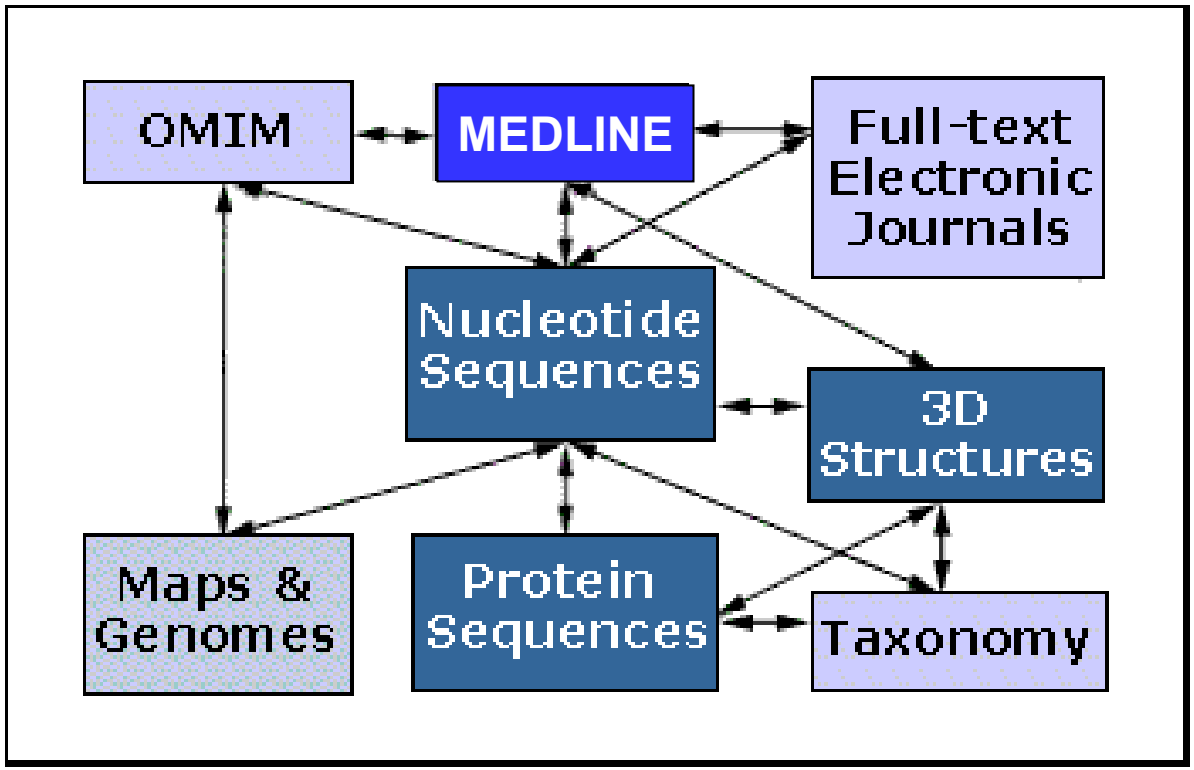
MEDLINE

Pubmed: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

- Medical literature, analysis and retrieval system online
- Over 12 million citations from over 4,600 international journals (89% are in English)
- Covers basic biomedical research and clinical sciences dated back to 1966.
- Citations have defined structure. (Example)
- Can be searched through PubMED® using MeSH terms, author names, title words, journal names, phrase, or any combination of these.

Pubmed (entrez):

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed



Medical Subject Heading (MeSH)

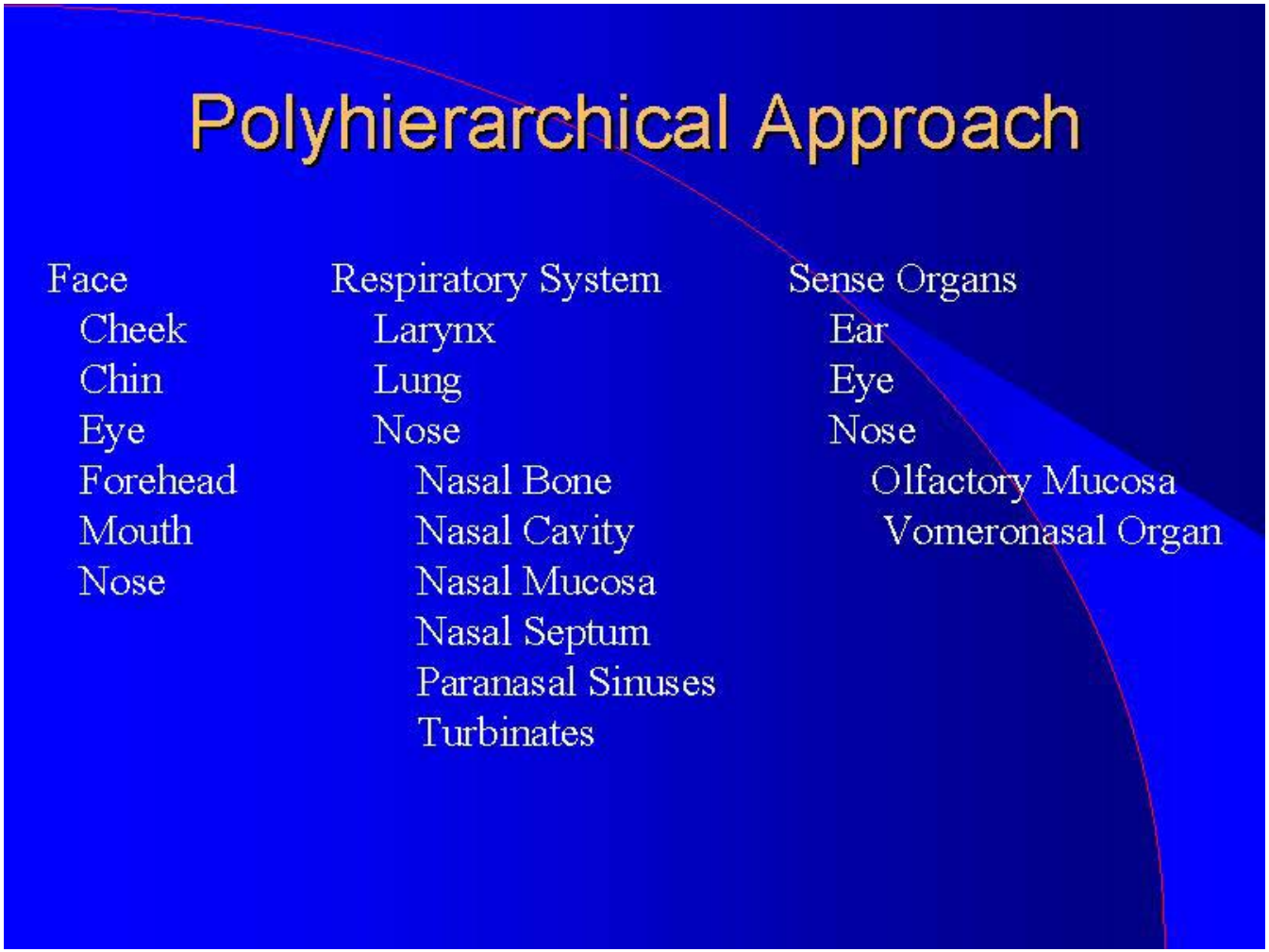
- http://www.nlm.nih.gov/mesh/meshhome.html

- First edition 1966
- Controlled vocabulary – A Thesaurus
- Used for indexing MEDLINE and Index Medicus
- 21,973 descriptors in hierarchical and alphabetical structure

MeSH Keywords are organized in 16 Concept Hierarchies

- Anatomy
- Organisms
- Diseases
- Chemicals and Drugs
- Analytical, Diagnostic and Therapeutic Techniques and Equipment
- Psychiatry and Psychology
- Biological Sciences
- Physical Sciences
- Anthropology, Education, Sociology and Social Phenomena
- Technology and Food and Beverages· Information Science
- Humanities
- Persons
- Health Care
- Geographic Locations

MeSH


Polyhierarchical Approach
Face
Cheek
Chin
Eye
Forehead
Mouth
Nose
Respiratory System
Larynx
Lung
Nose
Nasal Bone
Nasal Cavity
Nasal Mucosa
Nasal Septum
Paranasal Sinuses
Turbinates
Sense Organs
Ear
Eye
Nose
Olfactory Mucosa
Vomeronasal Organ

High Density Array Interpreter (HAPI)

http://array.ucsd.edu/hapi/

The **H**igh-density **A**rray **P**attern **I**nterpreter (HAPI) provides a novel method for interpreting the conceptual similarities of a cluster or group of genes that have been identified by a statistical methods .

Data linkages for Affy U95a

Probe sets:	17,768
GenBank Acc#:	13,488
Citations:	9,061
Unique Citations:	7,679
MeSH Terms:	107,493

Fraction of array with 1 or more matching citations: 78.8

Masys et al. (2001) Bioinformatics 17:319-326..

Predictive Genes Associated with ALL and AML

Genes predictive of Acute Lymphocytic Leukemia (ALL)	Genes predictive of Acute Myelogenous Leukemia (AML)
U22376 c-myb X59417 Proteasome iota PROS-27 U05259 MB-1 M92287 cyclin D3 M31211 Myosin light chain X74262 RbAp48 retinoblastoma binding protein D26156 Transcriptional activator hSNF2b S50223 HKR-T1=Kruppel-like zinc finger protein M31523 E2A transcription factor L47738 Inducible protein U32944 Dynein light chain 1 Z15115 TOP2 DNA topoisomerase II) X15949 IRF2 Interferon regulatory factor X63469 TFIIE beta transcription factor M91432 MCAD medium-chain acyl-CoA dehydrogenase U29175 BRG1 Transcriptional activator Z69881 Ca2+ ATPase U20998 SRP 9 Signal recognition particle subunit 9 D38073 MCM3 hRlf beta subunit (p102 protein) U26266 Deoxyhypusine synthase M31303 Op18 Oncoprotein 18 Y08612 Rabaptin Nup88 protein U35451 Heterochromatin protein M29696 IL-7 Interleukin-7 receptor M13792 ADA Adenosine deaminase (ADA)	M5150 Fumarylacetoacetate hydrolase X95735 Zyxin 2 U50136 LTC4S Leukotriene C4 synthase M16038 LYN tyrosine kinase U82759 HoxA9 Homeodomain protein M23197 CD33 Human differentiation antigen M84526 Adipsin/complement factor D Y12670 Leptin receptor M27891 CST3 cystatin C X17042 Hematopoetic proteoglycan core protein Y00787 MDNCF monocyte-derived neutrophil chemotactic factor M96326 Azurocidin U46751 p62 for the Lck SH2 domain M80254 hCyP3 Cyclophilin isoform L08246 MCL1 Myeloid cell differentiation protein M62762 Vacuolar H+ ATPase proton channel subunit M28130 Interleukin 8 (IL8) M63138 Cathepsin D (catD) gene M57710 Epsilon-BP IgE-binding protein M69043 MAD-3 mRNA encoding IkB-like activity M81695 Leukocyte adhesion glycoprotein p150,95 X85116 Epb72 M19045 Lysozyme mRNA M83652 Properdin X04085 Catalase

Golub et al., Science. 1999 Oct 15;286(5439):531-7.

Arrowsmith

http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

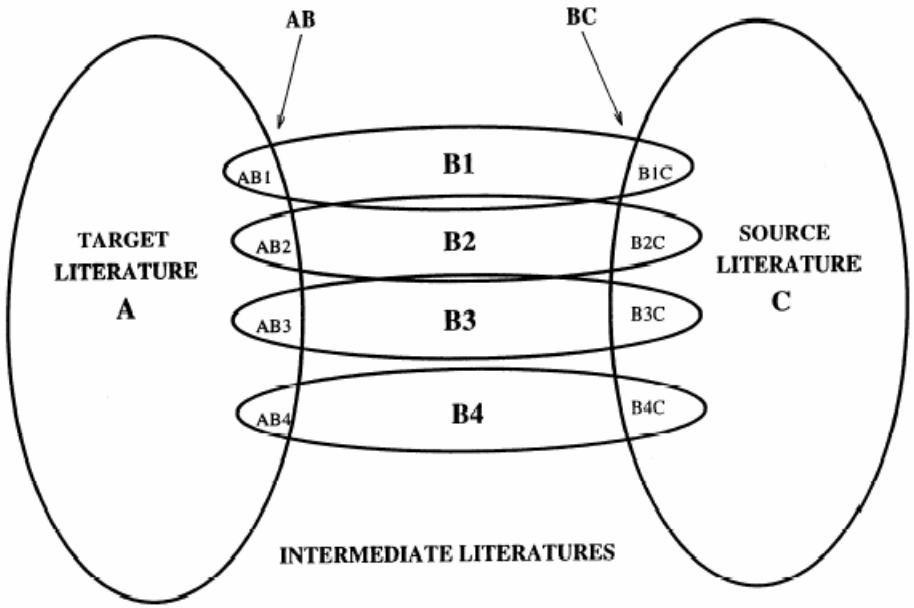


Fig. 1. A Venn diagram that represents sets of articles, or literatures, containing the words A and C in their titles. Sets A and C are linked through intermediate sets B_i ($i = 1, 2, 3\ldots$) which contain the word B_i in their titles and which overlap both A and C. By examining the articles in the pairs of intersections AB_i and B_iC, useful information may be inferred regarding possible biological linkages among A, B and C. (A and C are shown here as having no articles in common. When there is overlap between sets A and C, the articles in the direct intersection should first be identified and evaluated prior to carrying out an ARROWSMITH search.) Modified from [9] with permission.

Useful Links

- **MeSH:** http://www.nlm.nih.gov/mesh/meshhome.html
- **HAPI:** http://array.ucsd.edu/hapi/
- **Arrowsmith:** http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html