Inauguration de l'Unité Inserm U936

Modélisation des connaissances biomédicales

February 6, 2009

Yes We Can

From biomedical informatics to translational research

De l'informatique biomédicale à la médecine translationelle

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Outline

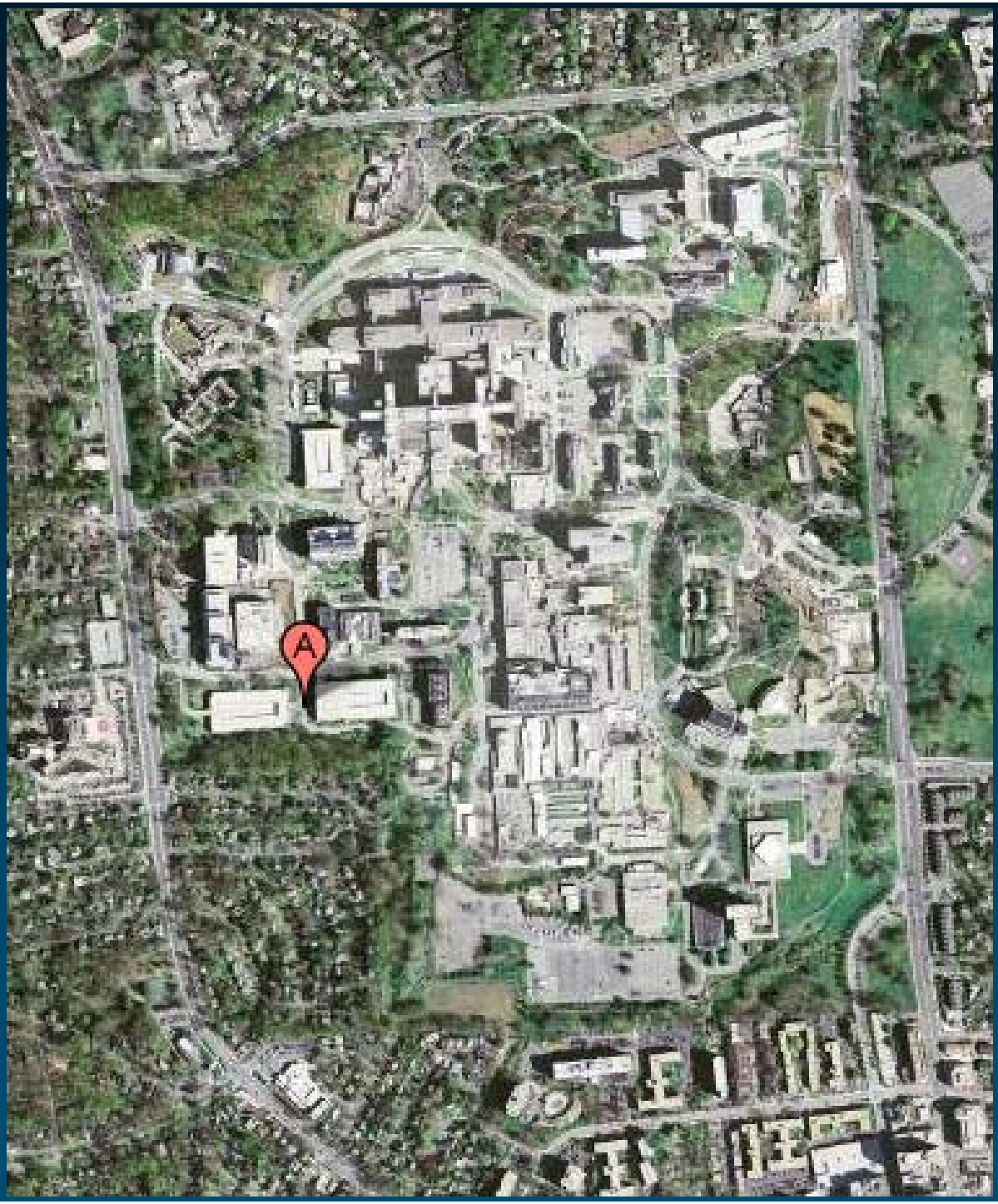
- National Institutes of Health
- Translational research
- Enabling translational research
- Anatomy of a translational research experiment
- Promising results
- Challenging issues

YES
WE
CAN

National Institutes of Health

Vancouver
Seattle
Washington
Portland
Oregon
Montana
North Dakota
Minnesota
South Dakota
Idaho
Wyoming
Minneapolis
Wisconsin
Michigan
Québec
New Brunswick
Ottawa
Montréal
Maine
Nova Scotia
Toronto
Milwaukee
Buffalo
New York
Chicago
Detroit
Cleveland
Toledo
Vermont
New Hampshire
Massachusetts
Rhode Island
Connecticut
New Jersey
Delaware
Maryland
District of Columbia
Virginia
West Virginia
Kentucky
Tennessee
North Carolina
South Carolina
Georgia
Florida
Miami
Jacksonville
Orlando
Atlanta
Alabama
Mississippi
Louisiana
Arkansas
Oklahoma
Texas
Austin
San Antonio
Houston
Dallas
Monterrey
Torreón
Gulf of Mexico
North Atlantic Ocean
Salt Lake City
Nebraska
Lincoln
Omaha
Iowa
Illinois
Indiana
Ohio
Kansas City
St Louis
Cincinnati
Missouri
Kansas
Wichita
Tulsa
Denver
Colorado
Utah
Nevada
Sacramento
San Francisco
Stockton
California
Las Vegas
Bakersfield
Los Angeles
Riverside
San Diego
Arizona
Phoenix
Tucson
Albuquerque
New Mexico
El Paso
A

Pennsylvania
Canton
Pittsburgh
Harrisburg
Allentown
Reading
Clifton
Yonkers
Newark
New York
Edison
Trenton
Lancaster
Philadelphia
Toms River
Morgantown
Cumberland
Frederick
Maryland
New Jersey
Winchester
A
Silver Spring
Baltimore
Gaithersburg
Washington
Delaware
West Virginia
Harrisonburg
Staunton
Charlottesville
Richmond
Lynchburg
Virginia


A

National Institutes of Health

- Research is the heart of the NIH mission
- 27 Institutes and Centers
  - Institutes (Cancer, Heart, Lung & Blood)
  - Centers (NCCAM, NCRR)
- Intramural – 6,000 researchers at NIH Institutes
- Extramural – grants for research, research resources, training

http://www.nih.gov/

Facts about NIH

- $29.5 billion appropriation from Congress
- 80% awarded via competitive grants to more than 325,000 researchers at 3,000 organizations
  - 80,000 new applications received every year
  - Average research grant = $355K, 4 years
- 10% to support the intramural research program
  - 6,000 scientists, mainly on the NIH campus in Bethesda, Maryland

A

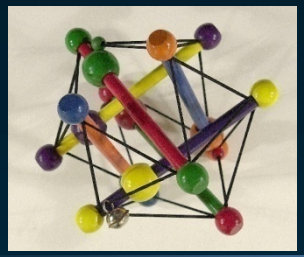
National Library of Medicine

Collaboration NLM-Inserm

Research at NLM

- Two research centers
  - National Center for Biotechnology Information (NCBI)
    - PubMed

    - Other knowledge base in the Entrez system (GenBank, Gene, PubMedCentral, PubChem, GEO, dbGaP,…)
  - Lister Hill National Center for Biomedical Communications
    - Unified Medical Language System (UMLS)
    - ClinicalTrials.gov
    - Genetics Home Reference
    - …



Medical Ontology Research

- Terminology, ontology and knowledge representation in the biomedical domain
  - Basic research
  - Applied research
- Applications
  - Natural language understanding
  - Automatic reasoning
  - Information visualization
  - Information integration

Translational research

(Translational medicine)

Translational medicine/research

- Definition [Butte, JAMIA 2008]
  - Effective transformation of information gained from biomedical research into knowledge that can improve the state of human health and disease
- Goals
  - Turn basic discoveries into clinical applications more rapidly ("bench to bedside")
  - Provide clinical feedback to basic researchers

Combining clinical informatics and bioinformatics

- Associates
  - Clinical informatics
    - Electronic medical records
    - Clinical knowledge bases
  - Common computational resources
    - Biomedical natural language processing
    - Biomedical knowledge engineering
  - Bioinformatics
    - Sequence databases
    - Gene expression
    - Model organism databases

Translational bioinformatics

- *"… the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into proactive, predictive, preventative, and participatory health.*

  Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations.

  The end product of translational bioinformatics is newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders, including biomedical scientists, clinicians, and patients."

AMIA strategic plan
http://www.amia.org/inside/stratplan

Aspects of translational research

- Huge volumes of data
- Publicly available repositories
- Publicly available tools
- Data-driven research

Huge volumes of data

- Affordable, high-throughput technologies
  - DNA sequencing
    - Whole genomes
    - Multiple genomes
  - Single nucleotide polymorphism (SNPs) genotyping
    - Millions of allelic variants between individuals
  - Gene expression data from micro-array experiments
  - Text mining
    - Full-text articles
    - Whole MEDLINE
  - Electronic medical records
  - Genome-wide association studies

Publicly available repositories

- DNA sequences
  - GenBank / EMBL / DDBJ
- Gene Expression data
  - GEO, ArrayExpress
- Biomedical literature
  - MEDLINE, PubMedCentral
- Biomedical knowledge
  - OBO ontologies
- Clinical data (genotype and phenotype)
  - dbGaP

Publicly available tools

- DNA sequences
  - BLAST
- Gene Expression data
  - GenePattern, …
- Biomedical literature
  - Entrez, MetaMap
- Biomedical knowledge
  - Protégé

Culture of sharing encouraged by the funding agencies

- Grants for tools and resource development
- Mandatory sharing plan in large NIH grants
- Mandatory sharing of manuscripts in PMC for NIH-funded research

Data-driven research

- Paradigm shift
  - Hypothesis-driven
    - Start from hypothesis
    - Run a specific experiment
    - Collect and analyze data
    - Validate hypothesis (or not)
  - Data-driven
    - Integrate large amounts of data
    - Identify patterns
    - Generate hypothesis
    - Validate hypothesis (or not) through specific experiments

Biomedical informatics as a supporting discipline for biology and clinical medicine

Biomedical informatics as a discipline in its own right, addressing important questions in medicine

Translational bioinformatics as a discipline

- *"The availability of substantial public data enables bioinformaticians' roles to change. Instead of just facilitating the questions of biologists, the bioinformatician, adequately prepared in both clinical science and bioinformatics, can ask new and interesting questions that could never have been asked before. […] There is a role for the translational bioinformatician as question-asker, not just as infrastructure-builder or assistant to a biologist."*

[Butte, JAMIA 2008]

YES
WE
CAN

Enabling translational research

Clinical Translational Research Awards (CTSA)

Translational research NIH Roadmap

Clinical and Translational Science Awards

- *The purpose of the CTSA Program is to assist institutions to forge a uniquely transformative, novel, and integrative academic home for Clinical and Translational Science that has the consolidated resources to:*
  - *1) captivate, advance, and nurture a cadre of well-trained multi- and inter-disciplinary investigators and research teams;*
  - *2) create an incubator for innovative research tools and information technologies; and*
  - *3) synergize multi-disciplinary and inter-disciplinary clinical and translational research and researchers to catalyze the application of new knowledge and techniques to clinical practice at the front lines of patient care.*

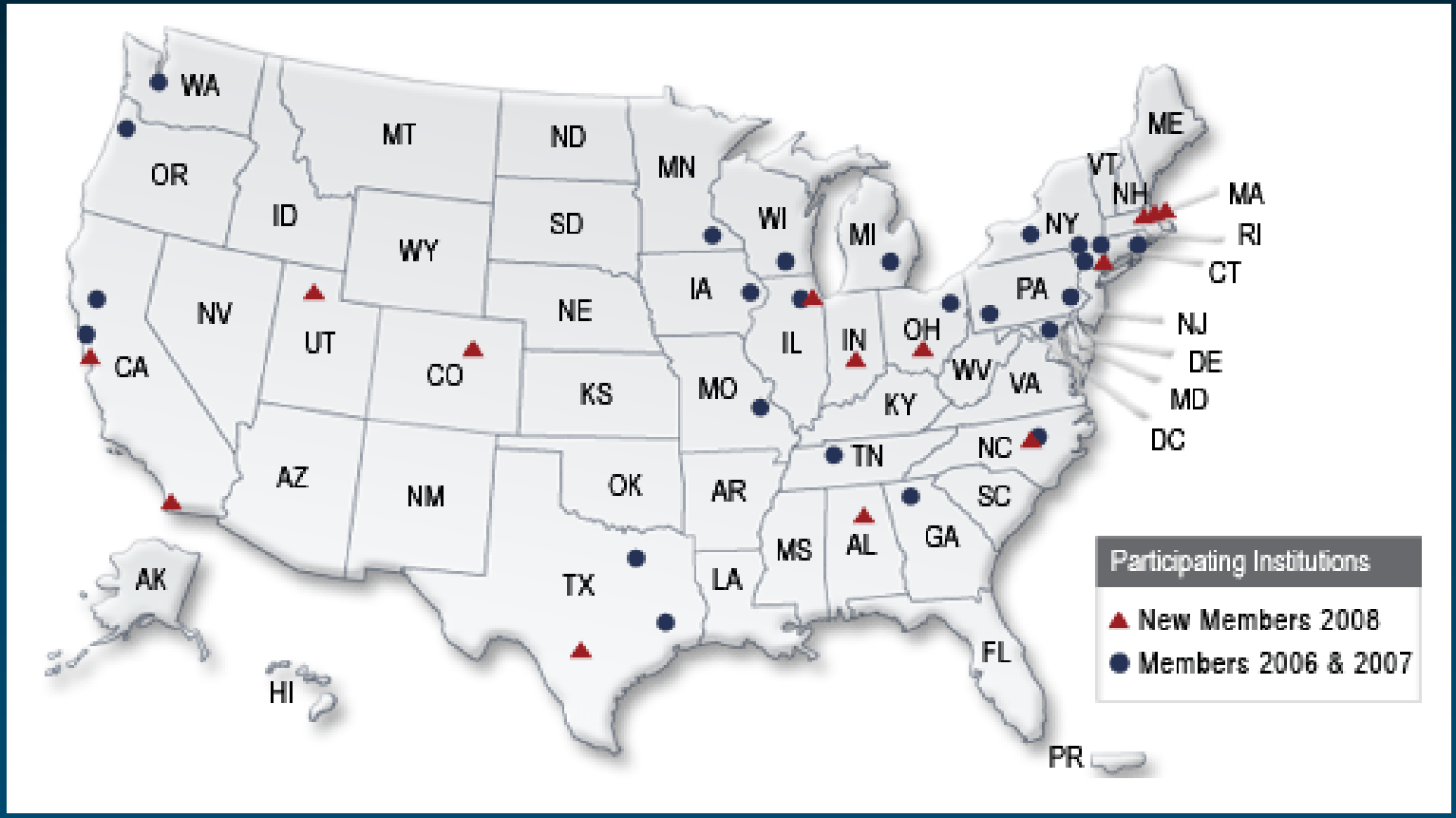
http://nihroadmap.nih.gov/

CTSA program (NCRR)

- 38 academic health centers in 23 states
  - 14 centers added in 2008
  - 60 centers upon completion
- Funding provided for 5 years
- Total annual cost: $500 M
- Annual funding per center: $4-23 M
  - Depending on previous funding

http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/

Clinical and Translational Science Awards

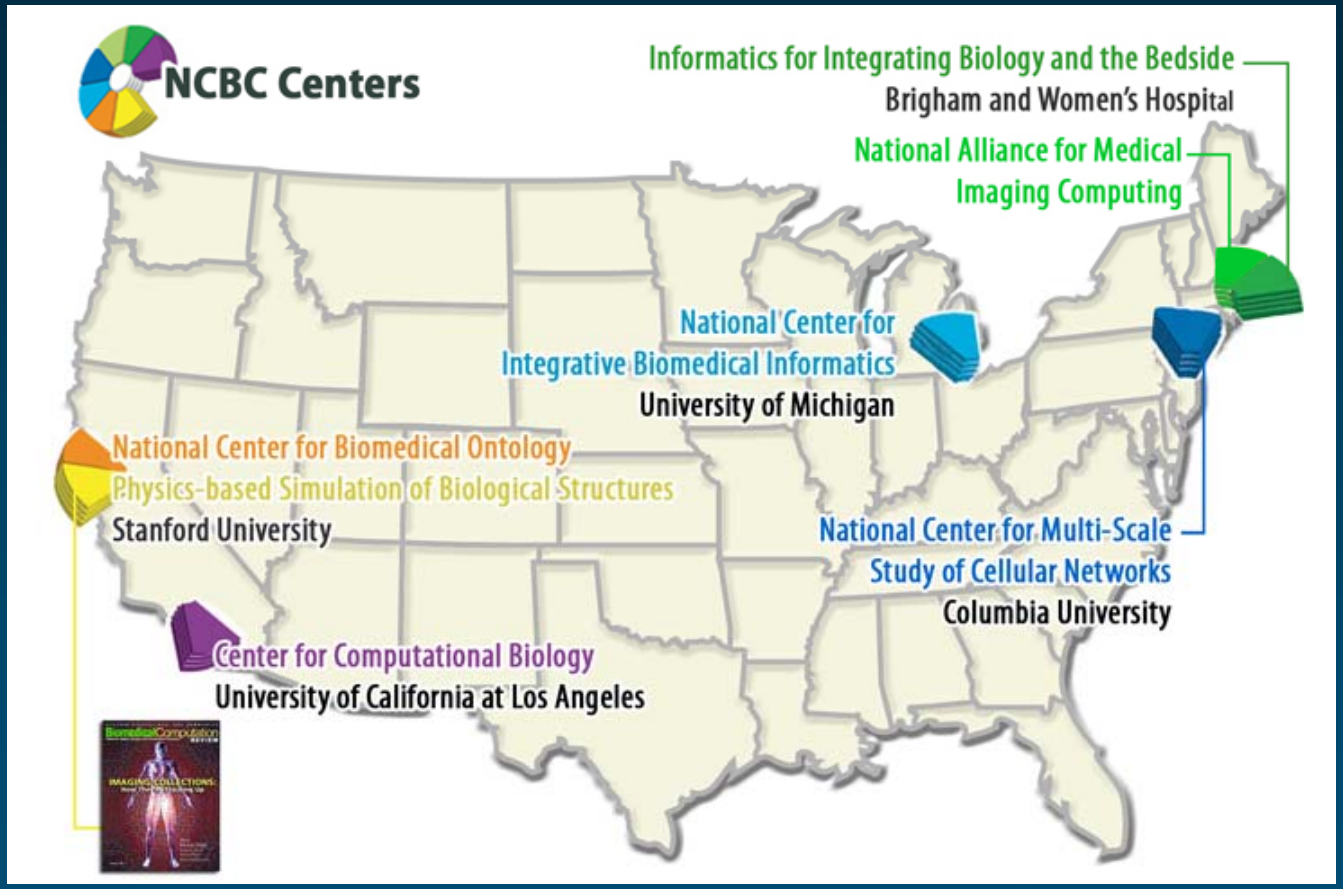


http://www.ctsaweb.org/

Other related programs

- National Centers for Biomedical Computing

"networked national effort to build the computational infrastructure for biomedical computing in the nation"



http://www.ncbcs.org/

Other related programs

- Cancer Biomedical Informatics Grid (caBIG)

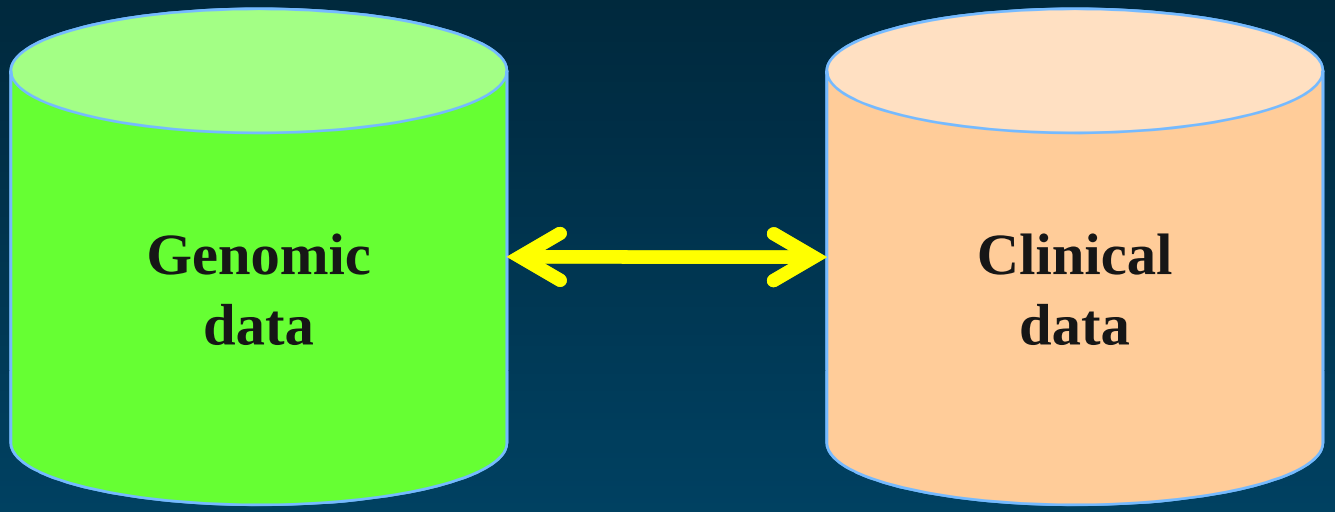
"an information network enabling all constituencies in the cancer community – researchers, physicians, and patients – to share data and knowledge."

  - Key elements
    - Bioinformatics and Biomedical Informatics
    - Community
    - Standards for Semantic Interoperability
    - Grid Computing
  - 1000 participants from 200 organizations
  - Funding: $60 M in the first 3 years (pilot)

https://cabig.nci.nih.gov/

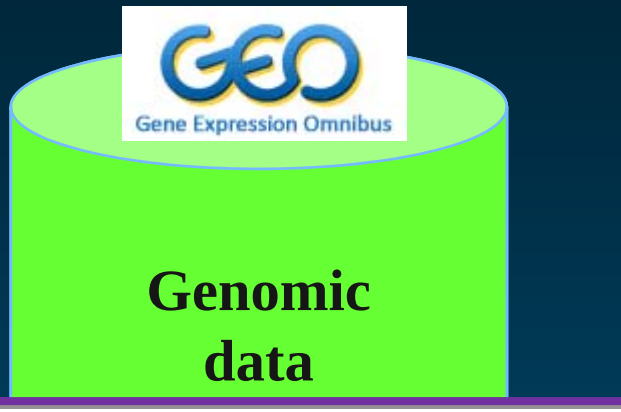
Anatomy of a translational research experiment

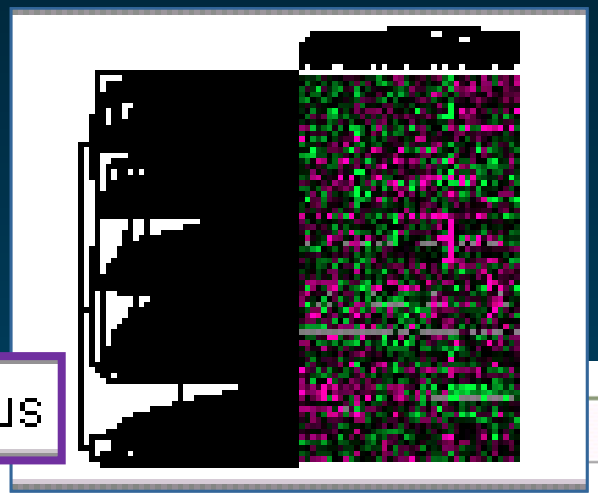
Integrating genomic and clinical data



- No genomic data available for most patients
- No precise clinical data available associated with most genomic data (GWAS excepted)

Integrating genomic and clinical data





Aging and cognitive impairment: hippocampus

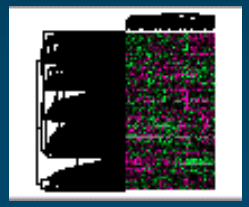
Accession:	GDS2639 ☛ View Expression (GEO profiles)
Title:	Aging and cognitive impairment: hippocampus
DataSet type:	gene expression array-based (RNA / in situ oligonucleotide)
Summary:	Analysis of hippocampi from aged learning-impaired animals on the last day of training in the Morris water maze (MWZ) or 21 days post-training. The MWZ task is a dorsal hippocampal-dependent task. Results provide insight into the molecular basis of aging-related cognitive impairment.
Platform:	GPL341: Affymetrix GeneChip Rat Expression Set 230 Array RAE230A
Citations:	Rowe WB, Blalock EM, Chen KC, Kadish I et al. Hippocampal expression analyses reveal selective association of immediate-early, neuroenergetic, and myelinogenic p… rats. *J Neurosci* 2007 Mar 21;27(12):3098-110. PMID: 17376971

Sample organism:	Rattus norvegicus	**Platform organism:**	
Feature count:	15923	**Value type:**	
Series:	GSE5666	**Series published:**	
Last GDS update:	04/27/2007		

MeSH Terms:

- Age Factors
- Animals
- Cognition Disorders/genetics

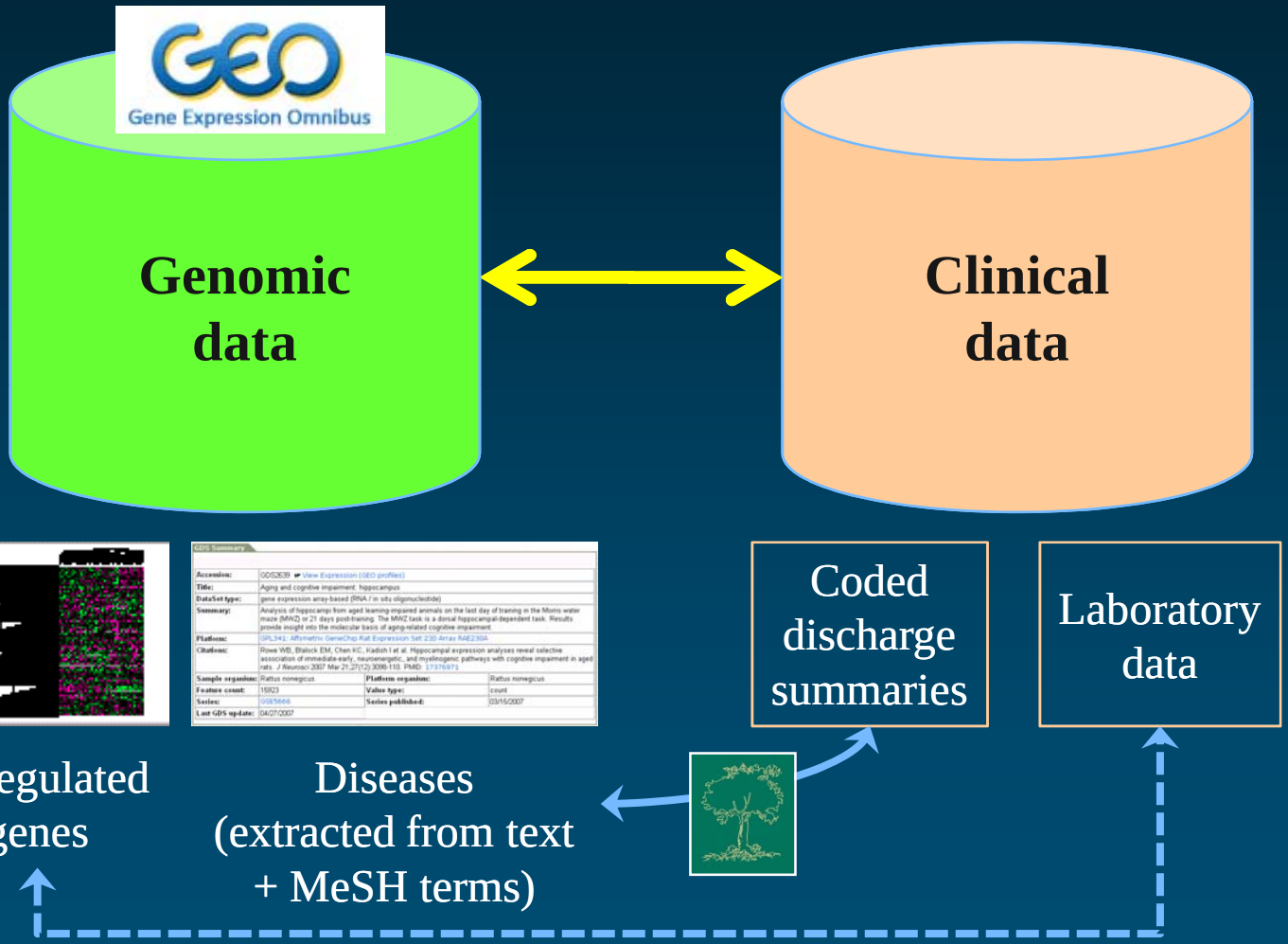
Integrating genomic and clinical data



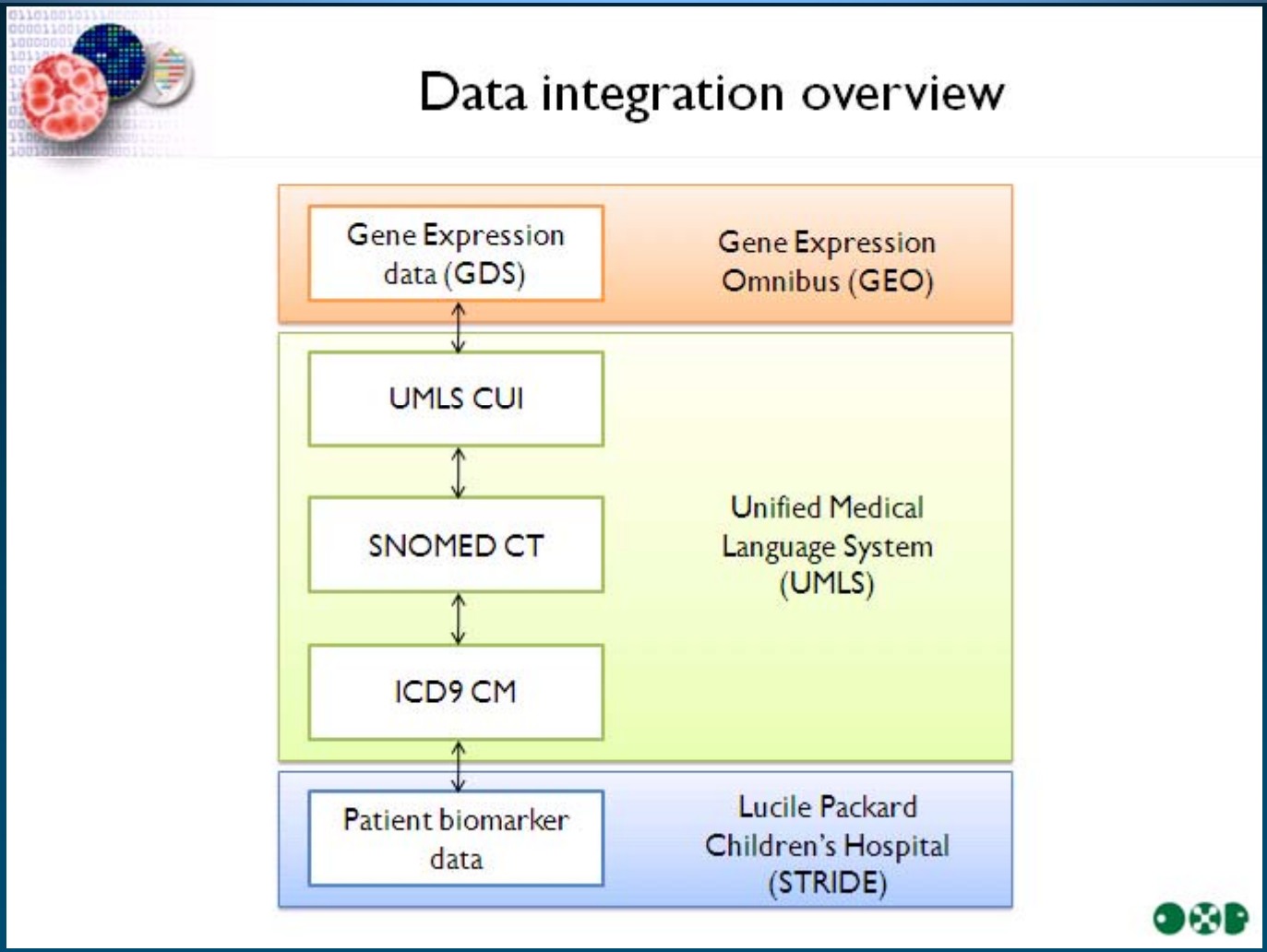
Upregulated genes

Diseases (extracted from text + MeSH terms)

Integrating genomic and clinical data



The Butte approach Methods



Courtesy of David Chen, Butte Lab

The Butte approach Results

Results

- 737 GEO Data Sets that were related to human disease
- 238 disease concepts were associated with GDS subsets
- 29,541 microarray samples were coded with SNOMED CT identifiers
- Note, we only included GDS that compared disease state to normal state
- 13,452 patients (of 49,414) mapped to 211 (of 238) of the disease concepts

Courtesy of David Chen, Butte Lab

The Butte approach

- Extremely rough methods
  - No pairing between genomic and clinical data
  - Text mining
  - Mapping between SNOMED CT and ICD 9-CM through UMLS
  - Reuse of ICD 9-CM codes assigned for billing purposes
- Extremely preliminary results
  - Rediscovery more than discovery
- Extremely promising nonetheless

The Butte approach References

- Dudley J, Butte AJ "Enabling integrative genomic analysis of high-impact human diseases through text mining." *Pac Symp Biocomput* 2008; 580-91
- Chen DP, Weber SC, Constantinou PS, Ferris TA, Lowe HJ, Butte AJ "Novel integration of hospital electronic medical records and gene expression measurements to identify genetic markers of maturation." *Pac Symp Biocomput* 2008; 243-54
- Butte AJ, "Medicine. The ultimate model organism." *Science* 2008; 320: 5874: 325-7

Promising results

Pharmacogenomics of warfarin

- Narrow therapeutic range
- Large interindividual variations in dose requirements
- Polymorphism involving two genes
  - CYP2C9
  - VKORC1
- Genetic test available
- Development of models integrating variants of CYP2C9 and VKORC1 for predicting initial dose requirements (ongoing RCTs)
- Step towards personalized medicine

Integration of existing studies/datasets

- 49 experiments in the domain of obesity
  - Rediscovery of known genes
  - Identification of potential new genes

[English, Bioinformatics 2007]

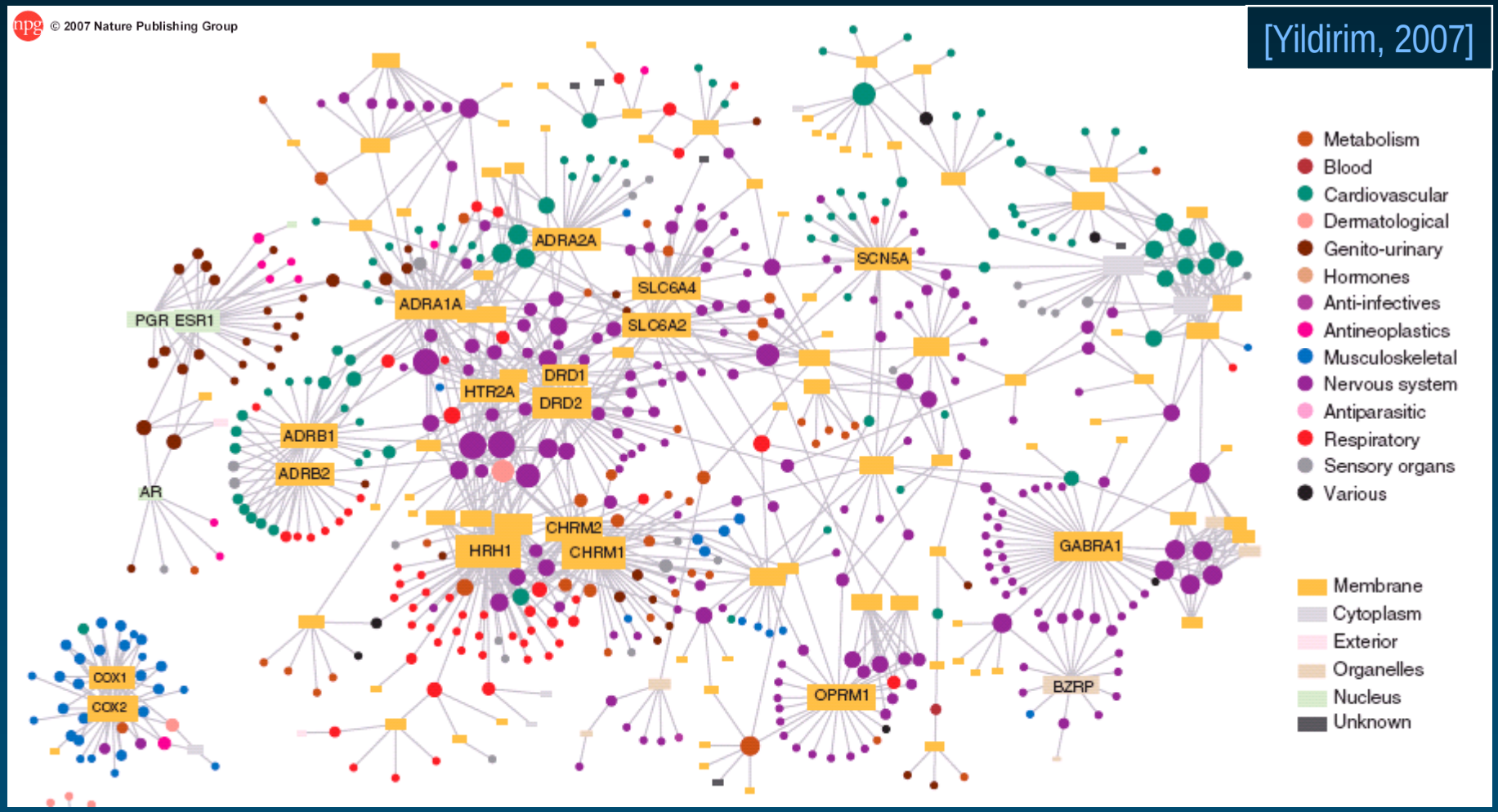
- Analysis of genes potentially associated with nicotine dependence
  - Rediscovery of known findings

[Sahoo, JBI 2008]

- Identification of networks of genes associated with type II diabetes mellitus

[Liu, PLoS 2007; Rasche, MBC Gen. 2008]

Drug-target networks



Challenging issues

Challenging issues

- Datasets
- Ontologies
- Tools
- Other issues

Challenging issues Datasets

- Lack of annotated datasets
  - Largely text-based (need for text mining)
- Limited availability of clinical data (EHRs, PHRs)
  - Need for deidentification
  - Largely text-based (need for text mining)
- Heterogeneous formats
  - Need for conversion
- Lack of metadata
  - Limited discoverability, limited reuse

Challenging issues Ontologies

- Lack of universal identifiers for biomedical entities
  - Need for normalization through terminology integration systems (e.g., UMLS)
- Lack of standard for identifiers
  - Need for bridging across formats
- Lack of universal formalism
  - Need for conversion between formalisms
- Limited availability of some ontologies
- Delay in adopting standards
  - e.g., SNOMED CT

Challenging issues Tools

- Lack of semantic interoperability
  - Difficult to combine tools/services
- Limited scalability of automatic reasoners
  - Difficult to process large datasets

Other challenging issues

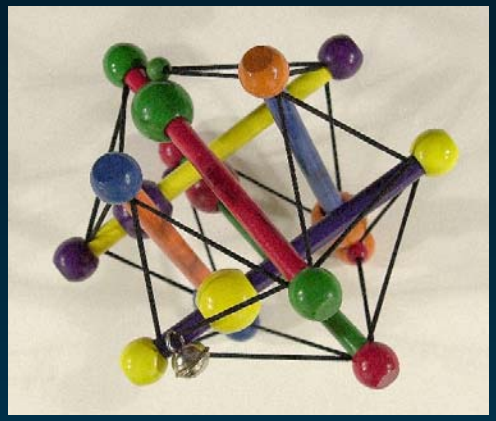
- Limited number of researchers "*adequately prepared in both clinical science and bioinformatics*"
- Need for validation of potential *in silico* discoveries through specific experiments
  - Collaboration with (wet lab) biologists
  - Must be factored in in grants

Conclusions

Conclusions

- Translational medicine is an emerging discipline
  - We live in partially unchartered territory
- Biomedical informatics is at the core of translational medicine
  - Strong informatics component to translational medicine
- We live in exciting times
  - New possibilities for biomedical informaticians
  - From service providers…
    …to biomedical researchers

YES
WE
CAN



Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA