National Library of Medicine
Informatics Training Conference

June 18-19, 2013
University of Utah

NATIONAL LIBRARY OF MEDICINE

In Memorium

Homer Warner, MD, PhD, was one of the earliest pioneers in the field of Biomedical Informatics. He published his first paper, "*A mathematical approach to medical diagnosis: Application to congenital heart disease*" in JAMA in 1961. He founded the Utah informatics program in the early 1960s and served as its chair until the mid-1990s. Homer passed away on November 30, 2012. Truly, he was one of the giants on whose shoulders we all stand.

Photo credit: **The New York Times**

TABLE OF CONTENTS

I. Agenda....5

II. Presenters Listed Alphabetically....13
Podium Presenters....13
Poster Presenters....15
Open Mic Session Presenters....17

III. **Tuesday, June 18, 2013**....5

Plenary Paper Session #1
Le, Washington....18
Weiss, Wisconsin-Madison....19
Smith, Vanderbilt....20
Hebert, Ohio....21
Culbertson, Regenstrief/Indiana....22

Poster Session/Coffee Break: Day 1 Group
Topic 1 – Health Care and Public Health
Woollen, Columbia....23
McCarthy, Washington....24
Zirkle, VA....25
Fillmore, Utah....26

Topic 2 – Clinical/Translational
Yim, Washington....27
Ross, UC San Diego....28
Szabo, Stanford....29

Topic 3 – Translational and Bioinformatics
Taylor, Rice/UTexas....30
Vincent, Wisconsin-Madison....31
Wu, Yale....32
McGinnis, Harvard....33
Vuong, Vanderbilt....34
Laderas, OHSU....35

Parallel Paper Focus Session A
Focus Session A1
Stokes, Pittsburgh....36
Pei, Yale....37
Hebbring, Wisconsin-Madison....38

Focus Session A2
Kendall, Washington....39
Chase, OHSU....40

Open-Mic Sessions – Table of Presenters....17

Plenary Paper Session #2
Hillenmeyer, Stanford 41
Allen, Colorado 42
Gibbs, OHSU 43
Liu, Rice 44
Levy, UC San Diego 45

Poster Session/Afternoon Break: Day 1 Group
Topic 1 – Health Care and Public Health
Woollen, Columbia....23
McCarthy, Washington 24
Zirkle, VA 25
Fillmore, Utah 26

Topic 2 – Clinical/Translational
Yim, Washington 27
Ross, UC San Diego 28
Szabo, Stanford 29

Topic 3 – Translational and Bioinformatics
Taylor, Rice/UTexas 30
Vincent, Wisconsin-Madison 31
Wu, Yale 32
McGinnis, Harvard 33
Vuong, Vanderbilt 34
Laderas, OHSU 35

Parallel Paper Focus Session B
Focus Session B1
Lyon, VA 46
Salmasian, Columbia 47
Gimenez, Stanford 48
Gupta, Harvard 49

Focus Session B2
Garcia-Hernandez, Utah 50
Shenvi, UC San Diego 51
VanHouten, Vanderbilt 52
Myers, Rice/UTexas 53

IV. **Wednesday, June 19, 2013 .....10**

Plenary Paper Session #3
Sarmiento, NLM 58
Engelhard, Virginia 59
Davis, Utah 60
Zimolzak, Harvard 61

Poster Session/Coffee Break: Day 2 Group
Topic 1 – Health Care and Public Health
Schulman, VA 62
Dunlea, Utah 63
Zhang, UC San Diego 64
Whaley, Rice/UTexas 65
Naumann, Harvard 66

Topic 2 – Clinical/Translational
Mortensen, Stanford 67
Romagnoli, Pittsburgh.....68
Hruby, Columbia 69

Topic 3 – Translational and Bioinformatics
Fillmore, Wisconsin-Madison 70
Gupta, Yale 71
Ogoe, Pittsburgh 72
Lazar, OHSU 73
Harrell, Vanderbilt 74
Grinter, Missouri-Columbia 75

Parallel Paper Focus Session C
Focus Session C1
Carr, VA 76
Rance, NLM 77
Weiskopf, Columbia 78

Focus Session C2
McDade, Pittsburgh 79
Stanton, Yale 80
Eickholt, Missouri-Columbia 81

Poster Session/Afternoon Break: Day 2 Group
Topic 1 – Health Care and Public Health
Schulman, VA 62
Dunlea, Utah 63
Zhang, UC San Diego 64
Whaley, Rice/UTexas 65
Naumann, Harvard 66

Topic 2 – Clinical/Translational
Mortensen, Stanford 67
Romagnoli, Pittsburgh 68
Hruby, Columbia 69

Topic 3 – Translational and Bioinformatics
Fillmore, Wisconsin-Madison 70
Gupta, Yale 71
Ogoe, Pittsburgh 72
Lazar, OHSU 73
Harrell, Vanderbilt 74
Grinter, Missouri-Columbia 75

V. **Map/Transportation**
Campus Map 78
Transportation Information 79

VI. **CONFERENCE EVALUATION SURVEY** 80

Agenda at a Glance ..... **Back Cover**

NLM Informatics Training Conference 2013

University of Utah, Salt Lake City, Utah
June 18-19, 2013

Agenda

TUESDAY, JUNE 18, 2013

7:00 – 8:00 AM Breakfast
Students - *Heritage Center*
Faculty - *Guest House*
NLM – *Marriott*

8:00 - 8:45 AM Welcome remarks: Dr. John Hurdle, University of Utah
Skaggs Hall NLM Director Remarks: Dr. Donald Lindberg
Introduction of Training Directors and Trainees: Dr. Valerie Florance

8:45 – 10:00 AM **Plenary Session #1 - Five 15 minute talks**
Skaggs Hall **Moderator: Alexa McCray, Harvard**

- Stakeholders' Evaluation of Visualization Approaches of Wellness
  (Thai Le, U of Washington)
- Multiplicative-Forest Continuous-Time Disease Prediction from EHRs
  (Jeremy Weiss, Wisconsin-Madison)
- Adverse Drug Effect Discovery from Data Mining of Clinical Notes
  (Joshua Smith, Vanderbilt)
- Administrative vs. Problem List Data for Comorbidity Detection: Re-Use Implications
  (Courtney Hebert, Ohio State U)
- SemRep to Extract Meaning from FDA Black Box Warnings: a Feasibility Study
  (Adam Culbertson, Regenstrief/Indiana)

10:00 – 11:00 AM **Poster Break -- Day 1 Group, 13 posters**
HSEB Atrium

***Health Care and Public Health*:**

- Engaging Hospitalized Patients with an Inpatient Personal Health Record
  (**Janet Woollen, Columbia**)
- A Mobile App for Hypothesis testing Symptom Triggers in Tourette Syndrome
  (**Ted McCarthy, U of Washington**)
- Developing Manually Annotated Corpus of VA EMR Notes for Symptoms and Treatments of PTSD
  (**Maryan Zirkle, Veteran's Admin**)
- Systematic Review of CDS with Potential for Inpatient Cost Reduction
  (**Christopher Fillmore, U of Utah**)

***Clinical/Translational*:**

- Real-Time Identification of Pneumonia Using Clinical Data
  (**Wen-wai Yim, U of Washington**)
- UMLS-Based Ontology of Phenotypes in the Database of Genotypes and Phenotypes (dbGaP)
  (**Mindy Ross, UC San Diego**)
- Identification of Novel Biomarkers for Early Detection of Ovarian Cancer
  (**Linda Szabo, Stanford**)

***Translational and Bioinformatics**:*

- Ovarian Cancer and the Hematopoietic Household
  (**Morgan Taylor, Rice/MD Anderson Cancer Center)**
- Novel Amino Acid Counting and Machine Learning Product Ion Identification Algorithms Improve Mass Spectrometry-Based Protein Identification
  (**Catherine Vincent, U of Wisconsin-Madison**)
- Understanding Interferon-Stimulated Genes Dynamic by Mathematical Modeling of Liver Cell Transcriptional Response to Interferon Alpha and Beta Stimuli
  (**Jialiang Wu, Yale**)
- DNA Methylation Modifications and Role in Fetal Growth Restriction
  (**Denise McGinnis, Harvard**)
- Commonality of Frequently Mutated Functional Domains Across Cancers
  (**Huy Vuong, Vanderbilt)**
- RPPAth: Integrate, Visualize, and Model Time-Series Data in Cancer Cell Lines
  (**Ted Laderas, OHSU**)

11:00 – 11:45 AM **Parallel Focus Sessions A, three 10 minute talks plus discussion**
HSEB 1700

***Focus Session A1; Moderator:* Perry Miller, Yale**

- The Application of Network Label Propagation for Biomarker Ranking in Genome-Wide Data
  (**Matthew Stokes, U of Pittsburgh**)
- Identification of Transcribed Pseudogenes in Human Genome
  (**Baikang Pei, Yale**)
- PheWAS – a Novel Method that Combines Genetics and Medical Informatics
  (**Scott Hebbring, U of Wisconsin-Madison**)

***Focus Session A2; Moderator:* Cindy Gadd, Vanderbilt**

- Physician Handoffs and Cross-Cover Chart Biopsy: A Descriptive Approach
- (**Logan Kendall, U of Washington**)
- Technology and Communications Between Providers – a Frontline Report
  (**Dian Chase, OHSU**)

11:45 AM – 1:00 PM **Executive Session of Training Directors**
HSEB 2100 (Session Chair: Dr. Donald A.B. Lindberg; **one** program faculty member per program

Heritage Center Birds of a Feather Lunch for students, remaining faculty, and staff

1:00 – 2:00 PM **Open Mic Session (14 talks, 4 minutes each)**
Skaggs Hall **Moderator: Bill Hersh, OHSU**

- Emotions in Medical Decision Making
  (**Margo Bergman, U of Washington**)
- Predicting Adverse Events and Patient Non-Compliance for Smarter Behavioral Health Screening
  (**Nima Behkami, Harvard**)
- Participatory Knowledge Acquisition for Tailored Diabetes Patient Decision Support
  (**Heather Cole-Lewis, Columbia**)
- Trial-Trak: Development of a User-Oriented Smart System for Identification of Eligible Clinical Trials
  (**Hua Fan-Minogue, Stanford**)
- Systematic Review of Usability Evaluation in EHRs
  (**Julie Doberne, OHSU**)
- Meta-unsupervised Molecular Subtyping of Breast Cancer Patients from Multiple Clinical Trials
  (**Katie Planey, Stanford**)
- Social Media for Health
  (**Jina Huh, U of Washington**)
- GPU Accelerated Protein Structure Prediction using Small Angle X-ray Scattering Experimental Restraints
  (**Daniel Putnam, Vanderbilt**)
- Real-world Application of Natural Language Processing for Surveillance of Healthcare
  (**Valmeek Kudesia, Harvard**)
- LRP6 May Cause a Novel Human Developmental Disease
  (**Jacques Zaneveld, Rice**)
- Clinician-focused Pharmacogenomics Information Resource: An Initial Prototype
  (**Katrina Romagnoli, Pittsburgh**)
- Computational Approach to Identify Motifs of Major Histocompatibility Complex (MHC) Class I Binding Peptides
  (**Xueheng Zhao, Stanford**)
- Towards an Ontology for Hospital Readmissions
  (**Colin Walsh, Columbia**)
- A Corpus-Based Study of Temporal Relations in Clinical Text
  Natalya Panteleyeva, Colorado)

2:00 – 3:15 PM **Plenary Session #2, five 15 minute talks;**
Skaggs Hall **Moderator: Mark Craven, U Wisconsin**

- Identification of Gene-Level Associations in Exome Data: Application to Warfarin
  (**Sara Hillenmeyer, Stanford University**)
- The p53 Transcriptome

(**Mary Allen, U of Colorado**)

- Multi-Omic Co-Expression Network Signatures of Disease
  (**David Gibbs, OHSU**)
- Pleiotropy and Polygenes in Adaptive Hybridized Gene Clusters in Mice
  (**Kevin Liu, Rice University**)
- Using Distributed Data to Enhance Predictive Models
  (**Eric Levy, UC San Diego**)

3:15 – 4:15 PM **Poster Break -- Day 1 Group, 13 posters**
HSEB Atrium ***Health Care and Public Health*:**

- Engaging Hospitalized Patients with an Inpatient Personal Health Record
  (**Janet Woollen, Columbia**)
- A Mobile App for Hypothesis testing Symptom Triggers in Tourette Syndrome
  (**Ted McCarthy, U of Washington**)
- Developing Manually Annotated Corpus of VA EMR Notes for Symptoms and Treatments of PTSD
  (**Maryan Zirkle, Veteran's Admin**)
- Systematic Review of CDS with Potential for Inpatient Cost Reduction
  (**Christopher Fillmore, U of Utah**)

***Clinical/Translational*:**

- Real-Time Identification of Pneumonia Using Clinical Data
  (**Wen-wai Yim, U of Washington**)
- UMLS-Based Ontology of Phenotypes in the Database of Genotypes and Phenotypes (dbGaP)
  (**Mindy Ross, UC San Diego**)
- Identification of Novel Biomarkers for Early Detection of Ovarian Cancer
  (**Linda Szabo, Stanford**)

***Translational and Bioinformatics**:*

- Ovarian Cancer and the Hematopoietic Household
  (**Morgan Taylor, Rice/MD Anderson Cancer Center)**
- Novel Amino Acid Counting and Machine Learning Product Ion Identification Algorithms Improve Mass Spectrometry-Based Protein Identification
  (**Catherine Vincent, U of Wisconsin-Madison**)
- Understanding Interferon-Stimulated Genes Dynamic by Mathematical Modeling of Liver Cell Transcriptional Response to Interferon Alpha and Beta Stimuli
  (**Jialiang Wu, Yale**)
- DNA Methylation Modifications and Role in Fetal Growth Restriction
  (**Denise McGinnis, Harvard**)
- Commonality of Frequently Mutated Functional Domains Across Cancers
  (**Huy Vuong, Vanderbilt)**
- RPPAth: Integrate, Visualize, and Model Time-Series Data in Cancer Cell Lines
  (**Ted Laderas, OHSU**)

4:15 – 5:15 PM **Parallel Focus Sessions B, four 10 minute talks plus discussion**

HSEB 1700

Focus Session B1; Moderator: **George Demiris, U Washington**

- Chronic Noncancer Pain in Iraq/Afghanistan Veterans
  (**Lawrence Lyon, Veterans Admin**)
- Deriving a Modified Charlson Comorbidity Index from Clinical Narratives Using Natural Language Processing
  - **Hojjat Salmasian, Columbia**)
- Decision Support System to Improve Positive Predictive Value in Mammography
  (**Francisco Gimenez, Stanford**)
- Decision Support Impact on Physician Adherence to Appropriate Use of ED Imaging
  (**Anurag Gupta, Harvard**)

HSEB 1730

Focus Session B2; Moderator: **Larry Hunter, U Colorado**

- A New Method to Store, Classify and Retrieve Electrophysiological Signals
  (**Antonio Garcia-Hernandez, U of Utah**)
- Modeling a Decision Rule to Guide CT Usage in Pediatric Blunt Abdominal Trauma
  (**Edna Shenvi, UC San Diego**)
- Random Forest Classification of Acute Coronary Syndromes
  (**Jacob VanHouten, Vanderbilt**)
- Vital Sign Quality Assessment Using Ordinal Regression of Time Series Data
  (**Risa Myers, Rice University**)

5:15 PM Faculty and students walk to housing

6:15 PM *Shuttle Bus Service to Utah Museum of Natural History
Rally at Heritage Center (Officer Circle) for buses

6:30 – 10:00 PM Dinner/Reception; live Jazz ensemble
Utah Museum of Natural History

*Shuttle Bus Service will begin at 6:15 pm and will operate continuously until 11:00 pm

WEDNESDAY, JUNE 19, 2013

7:00 AM – 8:15 AM Breakfast
Students - Heritage Center
Faculty - Guest House
NLM – Marriott

8:30 AM – 9:30 AM **Plenary Session #3, four 15 minute talks; Moderator: George Hripcsak, Columbia**

- Data Accuracy in Journal Abstracts for Use in Informing Clinical Decisions (**Raymond Sarmiento, NLM**)
- Characterizing MS Disability with Accelerometer-Based Motion Capture in the Field (**Matthew Engelhard, U of Virginia**)
- Grid and Public Health: Using Grid Service to Share Resources **Kailah Davis, U of Utah**)
- Early Detection of Statin Adherence: Surveillance for a Quality Measure (**Andrew Zimolzak, Harvard**)

9:30 AM – 10:30 AM **Poster Break -- Day 2 Group, 14 posters**

***Health Care and Public Health*:**

- The Walkability Index: Modeling Local Weather as a Predictor of Exercise Behavior (**Daniel Schulman, Veterans Admin**)
- Building a Preference Repository to Facilitate Sharing Decision Making (**Robert Dunlea, U of Utah)**
- Creating a Usability Testing Ontology for Biomedical Information Retrieval Tools (**Jing Zhang, UC San Diego**)
- Inferring Functional Human Language Pathways (**Meagan Whaley, Rice University**)
- Topic Models for Mortality Modeling in Intensive Care Units (**Tristan Naumann, Harvard**)

***Clinical/Translational*:**

- Crowdsourcing Ontologies (**Jonathan Mortensen, Stanford**)
- User-Centered Design, Implementation and Evaluation of a Clinician-focused Pharmacogenomics Information Resource **(Katrina Romagnoli, U of Pittsburgh**)
- Characterization of the Biomedical Query Mediation Process (**Gregory Hruby, Columbia**)

***Translational and Bioinformatics**:*

- Evaluation of de Novo Transcriptome Assemblies from RNA-Seq Data
  (**Nathanael Fillmore, U of Wisconsin Madison)**
- Identifying Clonally Related Sequences in Immunoglobin Repertoires
  (**Namita Gupta, Yale**)
- Discovering Functional Modules Using Spectral Clustering and the Gene Ontology
  (**Henry Ogoe, U of Pittsburgh**)
- Gibbon Chromosomal Breakpoints Display Distinctive Epigenetic States
  (**Nathan Lazar, OHSU**)
- Optimizing the Orbitals Score Function for Protein-Ligand Interface Design in Rosetta
  (**Morgan Harrell, Vanderbilt**)
- Evaluating Docking Scoring Methods Using the 2012 CSAR Benchmark
  (**Sam Grinter, U of Missouri-Columbia**)

10:30 – 11:45 AM **University of Utah Showcase**: "***Visualizing the Future of Biomedicine"***
Skaggs Hall **Dr. Chris Johnson**, Distinguished Professor and Director of the Scientific Computing and Imaging Institute, University of Utah

11:45 – 1:00 PM Introduction to Career Development Awards and New Investigator R01 Grants, box lunch. ***Webinar for Trainees In or Nearing Their Final Year**
HSEB 3515C ***Webinar starts at Noon

Heritage Center Lunch for rest of students, faculty, and NLM staff

11:45 – 2:00 PM NLM – 2013 Biomedical Informatics Training Program Overview,
HSEB 4100C Lunch and Networking. ****Webinar for Training Program Administrators**
****Webinar starts at 1:00 pm

1:00 – 2:00 PM **Parallel Focus Sessions C, three 10 minute talks plus discussion**
HSEB 1700 ***Focus Session C1; Moderator:* Peter Embi, Ohio State University**

- Patient, Caregiver and Provider Perceptions of a Colon Cancer Personal Health Record
  (**Thomas Carr, Veterans Admin**)
- Looking for Paradigm Shifts in the Biomedical Literature (
  Bastien Rance, NLM)
- Defining Completeness of Electronic Health Records for Secondary Use: How Big is Your Database?
  (**Nicole Weiskopf, Columbia**)

HSEB 1730 ***Focus Session C2; Moderator: Karen Eilbeck, U of Utah***

- Data Driven Optimization of Gene Expression in Gynecological Cancer
  (**Kevin McDade, U of Pittsburgh**)
- Arpeggio: Separation of Technical and Biological Variability in ChIP-Seq for Classification and Peak Detection
  (**Kelly Stanton, Yale**)
- Predicting Aspects of Protein Structure with Deep Networks

(**Jesse Eickholt, U of Missouri-Columbia**)

2:00 – 3:00 PM **Poster Break -- Day 2 Group, 14 posters**
HSEB Atrium

***Health Care and Public Health*:**

- The Walkability Index: Modeling Local Weather as a Predictor of Exercise Behavior
  (**Daniel Schulman, Veterans Admin**)
- Building a Preference Repository to Facilitate Sharing Decision Making
  (**Robert Dunlea, U of Utah)**
- Creating a Usability Testing Ontology for Biomedical Information Retrieval Tools
  (**Jing Zhang, UC San Diego**)
- Inferring Functional Human Language Pathways
  (**Meagan Whaley, Rice University**
- Topic Models for Mortality Modeling in Intensive Care Units
  (**Tristan Naumann, Harvard**)

***Clinical/Translational*:**

- Crowdsourcing Ontologies
  (**Jonathan Mortensen, Stanford**)
- User-Centered Design, Implementation and Evaluation of a Clinician-focused Pharmacogenomics Information Resource
  (**Katrina Romagnoli, U of Pittsburgh**)
- Characterization of the Biomedical Query Mediation Process
  (**Gregory Hruby, Columbia**)

***Translational and Bioinformatics**:*

- Evaluation of de Novo Transcriptome Assemblies from RNA-Seq Data
  (**Nathanael Fillmore, U of Wisconsin Madison**)
- Identifying Clonally Related Sequences in Immunoglobin Repertoires
  (**Namita Gupta, Yale**)
- Discovering Functional Modules Using Spectral Clustering and the Gene Ontology
  (**Henry Ogoe, U of Pittsburgh**)
- Gibbon Chromosomal Breakpoints Display Distinctive Epigenetic States
  (**Nathan Lazar, OHSU**)
- Optimizing the Orbitals Score Function for Protein-Ligand Interface Design in Rosetta
  (**Morgan Harrell, Vanderbilt**)
- Evaluating Docking Scoring Methods Using the 2012 CSAR Benchmark
  (**Sam Grinter, U of Missouri-Columbia**)

3:00 – 3:30 PM Closing Session and Awards, John Hurdle and Valerie Florance
Skaggs Hall

PODIUM PRESENTATIONS (Listed Alphabetically By Author)			
Presenter	**Institution**	**Title**	**Page**
Allen, Mary	University of Colorado-Boulder	The p53 Transcriptome	42
Carr, Thomas	Department of Veterans Affairs, VA Medical Center	Patient, Caregiver, and Provider Perceptions of a Colon Cancer Personal Health Record	76
Chase, Dian	Oregon Health and Science University	Technology and Communications Between Providers – A Frontline Report	40
Culbertson, Adam W.	Regenstrief Institute & Indiana University	SemRep to Extract Meaning from FDA Black Box Warnings: a Feasibility Study	22
Davis, Kailah T.	University of Utah	Grid and Public Health: Using Grid Service to Share Resources	60
Eickholt, Jesse L.	University of Missouri	Predicting Aspects of Protein Structure with Deep Networks	81
Engelhard, Matthew	University of Virginia	Characterizing MS Disability with Accelerometer-Based Motion Capture in the Field	59
García-Hernández, Antonio	University of Utah	A New Method to Store, Classify, and Retrieve Electrophysiological Signals	50
Gibbs, David L.	Oregon Health and Science University	Multi-Omic Co-Expression Network Signatures of Disease	43
Gimenez, Francisco	Stanford University	Decision Support System to Improve Positive Predictive Value in Mammography	48
Gupta, Anurag	Harvard Medical School	Decision Support Impact on Physician Adherence to Appropriate Use of ED Imaging	49
Hebbring, Scott J.	University of Wisconsin-Madison	PheWAS – A Novel Method that Combines Genetics and Medical Informatics	38
Hebert, Courtney L.	Ohio State University	Administrative vs. Problem List Data for Comorbidity Detection: Re-Use Implications	21
Hillenmeyer, Sara	Stanford University	Identification of Gene-Level Associations in Exome Data: Application to Warfarin	41
Kendall, Logan	University of Washington	Physician Handoffs and Cross-Cover Chart Biopsy: A Descriptive Approach	39
Le, Thai	University of Washington	Stakeholders' Evaluation of Visualization Approaches of Wellness	18
Levy, Eric	University of California, San Diego	Using Distributed Data to Enhance Predictive Models	45

Liu, Kevin J.	Rice University	Pleiotropy and Polygenes in Adaptive Hybridized Gene Clusters in Mice	44
Lyon, Lawrence	Department of Veterans Affairs, Puget Sound Healthcare System	Chronic Noncancer Pain in Iraq/Afghanistan Veterans	46
McDade, Kevin K.	University of Pittsburgh	Data Driven Optimization of Gene Expression in Gynecological Cancer	79
Myers, Risa B.	Rice University	Vital Sign Quality Assessment Using Ordinal Regression of Time Series Data	53
Pei, Baikang	Yale University	Identification of Transcribed Pseudogenes in Human Genome	37
Rance, Bastien	National Library of Medicine	Looking for Paradigm Shifts in the Biomedical Literature	77
Salmasian, Hojjat	Columbia University	Deriving a Modified Charlson Comorbidity Index from Clinical Narratives Using Natural Language Processing	47
Sarmiento, Raymond Francis	National Library of Medicine	Data Accuracy in Journal Abstracts for Use in Informing Clinical Decisions	58
Shenvi, Edna C.	University of California, San Diego	Modeling a Decision Rule to Guide CT Usage in Pediatric Blunt Abdominal Trauma	51
Smith, Joshua C.	Vanderbilt University	Adverse Drug Effect Discovery from Data Mining of Clinical Notes	20
Stanton, Kelly	Yale University	Arpeggio: Separation of Technical and Biological Variability in ChIP-Seq for Classification and Peak Detection	80
Stokes, Matthew E.	University of Pittsburgh	The Application of Network Label Propagation for Biomarker Ranking in Genome-Wide Data	36
VanHouten, Jacob P.	Vanderbilt University	Random Forest Classification of Acute Coronary Syndromes	52
Weiskopf, Nicole G.	Columbia University	Defining Completeness of Electronic Health Records for Secondary Use: How Big is Your Database?	78
Weiss, Jeremy C.	University of Wisconsin-Madison	Multiplicative-Forest Continuous-Time Disease Prediction from EHRs	19
Zimolzak, Andrew J.	Harvard Medical School	Early Detection of Statin Adherence: Surveillance for a Quality Measure	61

POSTER PRESENTATIONS (Listed Alphabetically By Author)			
Presenter	**Institution**	**Title**	**Page**
Dunlea, Robert	University of Utah	Building a Preference Repository to Facilitate Shared Decision Making	63
Fillmore, Christopher L.	University of Utah	Systematic Review of CDS with Potential for Inpatient Cost Reduction	26
Fillmore, Nathanael	University of Wisconsin-Madison	Evaluation of de Novo Transcriptome Assemblies from RNA-Seq Data	70
Grinter, Sam Z.	University of Missouri-Columbia	Evaluating Docking Scoring Methods Using the 2012 CSAR Benchmark	75
Gupta, Namita	Yale University	Identifying Clonally Related Sequences in Immunoglobulin Repertoires	71
Harrell, Morgan	Vanderbilt University	Optimizing the Orbitals Score Function for Protein-Ligand Interface Design in Rosetta	74
Hruby, Gregory W.	Columbia University	Characterization of the Biomedical Query Mediation Process	69
Laderas, Ted	Oregon Health & Science University	RPPAth: Integrate, Visualize, and Model Time-series Data in Cancer Cell Lines	35
Lazar, Nathan H.	Oregon Health & Science University	Gibbon Chromosomal Breakpoints Display Distinctive Epigenetic States	73
McCarthy, Ted	University of Washington	A Mobile App for Hypothesis Testing Symptom Triggers in Tourette Syndrome	24
McGinnis, Denise	Harvard Medical School	DNA Methylation Modifications and Role in Fetal Growth Restriction	33
Mortensen, Jonathan M.	Stanford University	Crowdsourcing Ontologies	67
Naumann, Tristan	Massachusetts Institute of Technology	Topic Models for Mortality Modeling in Intensive Care Units	66
Ogoe, Henry A.	University of Pittsburgh	Discovering Functional Modules Using Spectral Clustering and the Gene Ontology	72
Romagnoli, Katrina M.	University of Pittsburgh	User-Centered Design, Implementation, and Evaluation of a Clinician-focused Pharmacogenomics Information Resource	68
Ross, Mindy K.	University of California, San Diego	UMLS Based Ontology of Phenotypes in the Database of Genotypes and Phenotypes (dbGaP)	28
Schulman, Daniel	Department of Veterans Affairs, VA Boston Healthcare System	The Walkability Index: Modeling Local Weather as a Predictor of Exercise Behavior	62

Szabo, Linda	Stanford University	Identification of Novel Biomarkers for Early Detection of Ovarian Cancer	29
Taylor, Morgan	University of Texas, MD Anderson Cancer Center	Ovarian Cancer and the Hematopoietic Household	30
Vincent, Catherine E.	University of Wisconsin-Madison	Novel Amino Acid Counting and Machine Learning Product Ion Identification Algorithms Improve Mass Spectrometry-Based Protein	31
Vuong, Huy	Vanderbilt University	Commonality of Frequently Mutated Functional Domains Across Cancers	34
Whaley, Meagan	Rice University	Inferring Functional Human Language Pathways	65
Woollen, Janet	Columbia University	Engaging Hospitalized Patients with an Inpatient Personal Health Record	23
Wu, Jialiang	Yale University	Understanding Interferon-Stimulated Genes Dynamic by Mathematical Modeling of Liver Cell Transcriptional Response to Interferon Alpha and Beta Stimuli	32
Yim, Wen-wai	University of Washington	Real-Time Identification of Pneumonia Using Clinical Data	27
Zhang, Jing	University of California, San Diego	Creating a Usability Testing Ontology for Biomedical Information Retrieval Tools	64
Zirkle, Maryan	Department of Veterans Affairs, Portland VA Medical Center	Developing Manually Annotated Corpus of VA EMR Notes for Symptoms and Treatments of PTSD	25

Open Mic Session Presenters

Health Care & Public Health Informatics		
Presenter	**Institution**	**Title**
Bergman, Margo	University of Washington	Emotions in Medical Decision Making
Cole-Lewis, Heather	Columbia University	Participatory Knowledge Acquisition for Tailored Diabetes Patient Decision Support
Doberne, Julie	Oregon Health & Science University	Systematic Review of Usability Evaluation in EHRs
Huh, Jina	University of Washington	Social Media for Health
Kudesia, Valmeek	Harvard Medical School	Real-world Application of Natural Language Processing for Surveillance of Healthcare
Romagnoli, Katrina M.	University of Pittsburgh	Clinician-focused Pharmacogenomics Information Resource: An Initial Prototype
Walsh, Colin	Columbia University	Towards an Ontology for Hospital Readmissions

Translational Bioinformatics and Clinical Research Informatics		
Presenter	**Institution**	**Title**
Behkami, Nima	Harvard Medical School	Predicting Adverse Events and Patient Non-Compliance for Smarter Behavioral Health Screening
Fan-Minogue, Hua	Stanford University	Trial-Trak: Development of a User-Oriented Smart System for Identification of Eligible Clinical Trials
Planey, Katie	Stanford University	Meta-unsupervised Molecular Subtyping of Breast Cancer Patients from Multiple Clinical Trials
Putnam, Daniel	Vanderbilt University	GPU Accelerated Protein Structure Prediction using Small Angle X-ray Scattering Experimental Restraints
Zaneveld, Jacques	Rice University	LRP6 May Cause a Novel Human Developmental Disease
Zhao, Xueheng	Stanford University	Computational Approach to Identify Motifs of Major Histocompatibility Complex (MHC) Class I Binding Peptides

Stakeholders' Evaluation of Visualization Approaches of Wellness

Authors:
Thai Le, Blaine Reeder, Daisy Yoo, Rafae Aziz, Hilaire Thompson, George Demiris, University of Washington

Abstract:
There exists a broad range of literature examining information visualization, defined as the use of computer-supported, interactive visual representations of data to amplify cognition. Appropriately applied to health information systems, these visualizations support consumers for both health assessment and shared decision-making. Though design guidelines and cognitive theories on information visualization exist, they are often understudied for older adults who face challenges associated with the aging process. We describe the design of prototype visualizations representing the complex construct of wellness (consisting of several underlying sub-constructs such as physical, psychological, spiritual, and social dimensions of well-being) and their evaluation through multiple focus groups with older adults (N=31) and health care providers (N=10). We selected these two groups as primary stakeholders for the presentation of wellness data collected from a longitudinal pilot study of a community based health monitoring system for older adults.

We performed a qualitative descriptive analysis to identify themes associated with the visualization process. We report on similarities and differences across both stakeholder groups related to perceived value, process, and effectiveness of the various visualization approaches. In addition, we provide a set of recommendations for the design of information visualizations of comprehensive health related parameters targeted towards older adults.

Multiplicative-Forest Continuous-Time Disease Prediction from EHRs

Authors:
Jeremy C Weiss, University of Wisconsin-Madison, Sriraam Natarajan, Wake Forest University
David Page, University of Wisconsin-Madison

Abstract:
Accurate prediction of future onset of disease from Electronic Health Records (EHRs) has important clinical and economic implications. In this domain the arrival of data comes at semi-irregular intervals and makes the prediction task challenging. Continuous-time Bayesian networks effectively model such processes but are limited by the number of conditional intensity matrices, which grows exponentially in the number of parents per variable. We develop a partition-based representation using regression trees and forests whose parameter spaces grow linearly in the number of node splits. Using a multiplicative assumption we show how to update the forest likelihood in closed form, producing efficient model updates. Our results show multiplicative forests can be learned from few temporal trajectories with large gains in performance and scalability particularly when risk factors are known to be independent, or multiplicative in nature. We apply the multiplicative-forest idea to EHR data and show that it improves our ability to predict the future onset of disease.

Adverse Drug Effect Discovery from Data Mining of Clinical Notes

Authors:
Joshua C Smith, Randolph A Miller, Vanderbilt University

Abstract:
Only postmarketing surveillance can detect some types of ADEs (adverse drug effects, e.g., effects appearing after years of exposure). Pharmacovigilance projects, such as the FDA Sentinel Initiative, and institutionally-based research, such as that of Friedman and colleagues at Columbia and Altman, et al., at Stanford, demonstrate the utility of using electronically stored health information to enhance drug safety. We studied the feasibility of using data mining techniques on EMR-based comprehensive History and Physical Exam (H&P) notes to detect novel ADEs. A previous Vanderbilt study validated accuracy of mentions of past medications in H&P notes. The Vanderbilt Synthetic Derivative, a de-identified version of Vanderbilt's EMR, provided a larger corpus of notes than used in many previous studies. We used NLP to extract current medications and clinical findings (including diseases) from 360,000 H&P notes. We identified ~35,000 statistically "interesting" drug-finding associations from among ~600 unique drugs and ~2000 unique findings in the notes. Additionally, we constructed a knowledge base using UMLS and FDA prescription drug labels to classify those correlated drug-finding pairs as *known ADEs* (drug causes finding, n≈10,500), *known indications* (drug treats/prevents finding, n≈3500), or an *unknown relationship* possibly indicating a new ADE (n≈20,000). Our preliminary results illustrate both the problems and potential of using data mining of EMR notes for ADE discovery.

Administrative vs. Problem List Data for Comorbidity Detection: Re-Use Implications

Author:
Courtney L Hebert, Chaitanya Shivade, Philip R O Payne, Peter J Embi, The Ohio State University

Abstract:
To develop a prospective risk-assessment tool for 30-day readmission, we required data on comorbidities that could be analyzed in real-time from the electronic health record (EHR). We hypothesized that comorbidities documented by clinicians in the “problem list” area of the EHR would be sufficiently complete and accurate to use for risk assessment when compared to administrative data, which are not available in real-time.

Administrative and problem list data were collected from patients admitted with congestive heart failure between 10/15/11 and 1/30/13. Administrative and problem list comorbidity data from the index hospitalization were compared using Kappa statistics and generally showed good to excellent reproducibility, although varied considerably by individual comorbidity. Further analysis illustrated availability and accuracy changes in these data when physicians began to use the problem list for billing purposes. Finally, we compared administrative and problem list data to a gold standard of manual chart review on a subset of the charts. Specific findings for each comorbidity will be discussed.

This study illustrates the benefits as well as limitations of using problem list data for real-time comorbidity detection. This research adds to a limited literature and informs a framework for the leveraging of different types of electronic data for re-use.

SemRep to Extract Meaning from FDA Black Box Warnings: a Feasibility Study

Authors:
Adam W Culbertson, Marcelo Fiszman, Thomas Rindflesch, Regenstrief/Indiana University

Abstract:
Introduction: As many as 770,000 patients yearly experience adverse drug reactions resulting in injury or death. Package insert labels (in particular black box warnings) provide valuable drug information, such as administration, adverse reactions, and population at risk. However, their free-text format presents a challenge to automated access. We evaluated the effectiveness of SemRep, a semantic processor, to extract useful information from black box warning labels.

Methods: We randomly selected 2,000 drug labels from the DailyMed website, 900 of which had black box warnings. Rules were developed and applied to the output of SemRep (semantic relations) to extract *drug reactions*, *conditions* (diseases) at risk, and *populations* at risk. 100 labels were manually annotated to create a gold standard. SemRep output was compared to the gold standard, and precision, recall, and F-Score were calculated.

Results: Precision, recall, and F-score were, 94%, 52%, 0.67 for *drug reactions*; 80%, 53%, and 0.64 for *conditions*; and 95%, 44%, 0.61 for *population.* Overall performance was 90% precision, 51% recall, and 0.65 F-Score.

Discussion: SemRep was effective in extracting information from black box warnings. The low recall was mostly due to anaphora. With moderate enhancement SemRep could be used to create a structured resource of drug information.

Engaging Hospitalized Patients with an Inpatient Personal Health Record

Authors:
Janet Woollen, Alexander Sackeim, Jenny Prey, David Vawdrey, Columbia University

Abstract:
Engaged patients have better health outcomes; however, patient engagement is often poorly addressed in the hospital. An inpatient personal health record (PHR) may improve patient engagement. The objectives of this research were (1) to explore hospitalized patients' information needs and (2) to assess such patients' usage and perceptions of usefulness of an inpatient PHR for enhancing engagement.

We observed twenty post-cardiothoracic surgery patients using tablet computers with access to an inpatient PHR. Semi-structured interviews were conducted and audio-recorded by the research team and transcribed, coded, and analyzed for emerging themes. The survey data were analyzed using descriptive statistics.

Fourteen out of the 20 patients used the inpatient PHR and completed the interview. All patients who used the PHR responded favorably to having access to their clinical data. Engagement, empowerment, and information needs emerged as major themes. Patients reported a desire to view diagnoses, labs, radiology reports, progress notes, and procedure notes, during and after hospital stay. Patients found medication information and links to educational materials especially helpful.

An inpatient PHR can be useful for increasing patient engagement and identifying and addressing patients' information needs. We plan to follow up with a larger study in the future.

A Mobile App for Hypothesis Testing Symptom Triggers in Tourette Syndrome

Authors:
Ted McCarthy, Anne M Turner, Sean Munson, University of Washington

Abstract:
Tourette Syndrome (TS) is marked by high symptom (tic) variability, and can lead to immense daily distraction, social stigmatization, and even pain and injury. While some contextual factors affecting tics are known, there likely exist a number of unconfirmed factors that affect symptoms[1]. Mobile applications have been developed to help patients monitor symptoms for other neurological disorders, including epilepsy[2]. We are beginning research and development of a mobile application to allow patients to monitor tics and factors that may affect them. The proposed application will allow patients to systematically test suspected symptom triggers and provide a visual, interactive graph of tic severity and triggers to track trends over time. Once this framework has been created and verified for TS, the software can be modified for use in tracking symptoms and exacerbating factors in other chronic disorders.

[1] C. A. Conelea and D. W. Woods, "The influence of contextual factors on tic expression in Tourette's syndrome: a review," *J Psychosom Res*, vol. 65, no. 5, pp. 487–496, Nov. 2008.
[2] S. Le, P. O. Shafer, E. Bartfeld, and R. S. Fisher, "An online diary for tracking epilepsy," *Epilepsy & Behavior*, vol. 22, no. 4, pp. 705–709, Dec. 2011.

Developing Manually Annotated Corpus of VA EMR Notes for Symptoms and Treatments of PTSD

Authors:
Maryan Zirkle[1,2], Bryan T Gamble[2], Dezon Finch[3],
1.Department of Veterans Affairs, Portland VA Medical Center, Portland, OR, 2.Oregon Health & Science University, 3.Department of Veterans Affairs James A. Haley Veterans' Hospital, Tampa, FL

Abstract:
Mental health notes contain implicit and explicit concepts difficult to extract from the narrative note. This is an ongoing challenge for use with advanced NLP (Natural Language Processing) tools. This project will construct a corpus of clinical text manually annotated for symptoms and treatments documented in the Veterans Health Administration (VHA) for Post Traumatic Stress Disorder (PTSD). We describe and discuss the annotation technique with regard to process and content. This includes defining a proper annotation schema, training annotators, and determining the level of information to be annotated. Clinical documents from the VHA EMR for patients known to have a diagnosis of PTSD are used. Protege Knowtator was used with a specialized schema to capture the documentation of free text that describes the concepts. We report annotator agreement on varying levels and are creating a reference standard for future tasks in NLP, machine learning, and text mining. Approximately 600 clinical notes have been annotated thus far to yield more than 1000 terms and /or phrases describing symptoms and treatments of PTSD.

Systematic Review of CDS with Potential for Inpatient Cost Reduction

Authors:
Christopher L Fillmore, Bruce E Bray, Kensaku Kawamoto, University of Utah

Abstract:
The purpose of this study was to systematically review trials of clinical decision support (CDS) interventions with the potential to reduce inpatient costs, so as to identify promising interventions for more widespread implementation and to inform future research in this area. MEDLINE was searched up to September 2012, and relevant studies were identified using titles and abstracts. Full text articles were reviewed to make a final determination on inclusion. Relevant characteristics of the studies were extracted and summarized.

Following a screening of 6,978 articles, 60 manuscripts were included. The majority of manuscripts were published during or after 2007. 63.3% of studies were pre-post comparisons, and 13.3% were randomized controlled trials. 56.7% of the studies were focused on pharmacotherapy. 71.7% of the studies resulted in statistically and clinically significant improvements in an explicit financial measure or a proxy financial measure. Only 15% of the studies directly measured the financial impact of an intervention, whereas the financial impact was inferred in the remainder of studies. Data on cost effectiveness was available for only one study.

Given these results, it is apparent that further research is needed on the cost impact and cost effectiveness of CDS in the inpatient setting.

Real-Time Identification of Pneumonia Using Clinical Data

Authors:
Wen-wai Yim, Cosmin A Bejan, Lucy Vanderwende, Fei Xia, Heather L Evans, Mark M Wurfel, Meliha Yetisgen-Yildiz, University of Washington

Abstract:
Pneumonia is one of the most common severe infections in critically-ill patients where early intervention is crucial. Even small treatment delays can lead to higher mortality, longer-term mechanical ventilation, and excessive hospital costs. Comprehensive electronic medical records (EMRs), which capture patient vital information in both structured and free-text format, provide unique opportunities to leverage natural language processing (NLP) for pneumonia surveillance—a task that is resource-intensive and requires real-time assessment.
Our lab has developed several clinical NLP tools to process medical records and to apply assertion analysis for states of interest [1,2]. We use these tools for pneumonia detection on a dataset of 5313 ICU physician notes (including admit, discharge, and daily progress notes) generated for 426 patients at the University of Washington. Our previous work has shown successful pneumonia identification using clinical notes [3]. Here, we extend previous work by incorporating new data describing the clinical state of each patient (e.g. white blood cell count, temperature) in addition to the corresponding free-text data. Features are ranked employing statistical significance testing and the most relevant ones are used in a machine learning framework based on support vector machine. The information is then projected and analyzed according to different time-spans [4].

References
[1] Tepper M, Capurro D, Xia F, Vanderwende L, Yetisgen-Yildiz M. Statistical Section Segmentation in Free-Text Clinical Records. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), 2012.
[2] Bejan CA, Vanderwende L, Xia F, Yetisgen-Yildiz M. Assertion modeling and its role in clinical phenotype identification. J Biomed Inform. 2013 Feb;46(1):68-74.
[3] Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Pneumonia identification using statistical feature selection. J Am Med Inform Assoc. 2012;19(5):817-23.
[4] Bejan CA, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Assessing pneumonia identification from time-ordered narrative reports. AMIA Annu Symp Proc. 2012;2012:1119-28.

UMLS Based Ontology of Phenotypes in the Database of Genotypes and Phenotypes (dbGaP)

Authors:
Mindy K Ross, Neda Alipanah, and Hyeoneui Kim, University of California, San Diego

Abstract:
DbGaP is a repository for genome-wide associated studies with abundant phenotypic data. However, its utility for study retrieval is limited because phenotype variables are not standardized.
To standardize the phenotype variables, we first programmatically reduced the 130,000+ phenotype variables in dbGaP to 65,191 unique meaningful variables. We then ran these variables through the standardization pipeline, which first mapped them to UMLS concepts using MetaMap then selected Concept Unique Identifiers (CUIs) representing key concepts of the variables based on predetermined rules. This process resulted in 5,096 unique CUIs.
We are now building an ontology to organize the CUIs into a formal semantic representation to support phenotype search through hierarchical expansions. We filtered 50 million UMLS records to 5,815 subclass relations and utilized the 5,096 CUIs in UMLS.MRREL table triple formats: source concept [A], target concept [B], and their relation [PARENT]. Filtering triples was a challenge because of the multiple, conflicting, and redundant hierarchies between semantic types,[2] which required a domain expert to refine the ontology structure. The ontology is organized by using the UMLS semantic types as top-level categories and organ systems/specific diseases as the sub-classes.
We expect this ontology will improve accuracy and efficiency in phenotype-based dbGaP study retrieval.

[1]Aronson A. [2012]. The UMLS Metathesaurus, MetaMap Portal [Online], Available: http://metamap.nlm.nih.gov/.
[2]Gu, et al. Questionable Relationship Triples in the UMLS. Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2012).

The authors were funded in part by NIH grants UH2HL108785 and T15LM011271 (MKR)

Identification of Novel Biomarkers for Early Detection of Ovarian Cancer

Authors:
Linda Szabo, Purvesh Khatri, Xiaodan Liu, Bruce Ling, Atul J Butte, Stanford University

Abstract:
Ovarian cancer is the leading cause of gynecologic cancer death and the fifth leading cause of cancer death in North American women, primarily due to lack of early symptoms or effective screening. Several early detection markers have been proposed, but to date all have failed validation.

First, in order to discover genes differentially expressed across ovarian cancer studies regardless of stage, we performed a meta-analysis of 7 microarray data sets consisting of 396 samples in the NCBI Gene Expression Omnibus (GEO) and validated results in 2 independent cohorts. We then used clinical and genomic data from The Cancer Genome Atlas to identify the subset of these genes that are differentially expressed in early stage ovarian cancer. Pilot ELISA tests for these early stage candidates whose protein products have been previously identified in international serum proteome projects confirmed that these proteins are differentially detectable in the blood of ovarian cancer patients compared to healthy controls. We are currently conducting a clinical trial at Stanford to validate these results in a larger cohort and identify the panel of proteins and scoring model with the best predictive value.

Ovarian Cancer and the Hematopoietic Household

Authors:
Morgan Taylor[1], Behrouz Zand[1], Wah Chiu[2], Anil K Sood[1], [1] University of Texas MD Anderson Cancer Center, [2]Baylor College of Medicine

Abstract:
One-third of women with epithelial ovarian cancer have paraneoplastic thrombocytosis at initial diagnosis, and affected women are significantly more likely to have advanced disease and shortened survival. The mechanisms by which platelets promote tumor progression have not been completely elucidated. Cancer cells are known to be present in platelet aggregates in the bloodstream, and growing data suggest an important role for platelets in tumor metastasis. However, the role of platelets in rescuing cancer cells circulating in the blood or ascites from anoikis is not known. We added platelets to a diverse panel of ovarian cancer cell lines and observed aggregate formation followed by decreases in anoikis rates of up to 42%. Inhibition of aggregate formation with heparin attenuated platelet-mediated effects.

These data support the hypothesis that platelets promote metastasis by forming aggregates around cancer cells that promote their anchorage-independent survival. The underlying mechanisms will be investigated under the following aims: 1) to examine the cell-platelet interface with either electron cryomicroscopy or conventional electron microscopy; 2) to quantify the ultrastructural differences between normal and platelet-exposed cells; and 3) to set up a database to archive the large 3-D platelet-cell datasets for later basic and clinical research purposes.

Novel Amino Acid Counting and Machine Learning Product Ion Identification Algorithms Improve Mass Spectrometry-Based Protein Identification

Authors:
Catherine E Vincent, Chris M Rose, Alicia L Richards, Derek J Bailey, Michael S Westphall, Joshua J Coon, University of Wisconsin-Madison

Abstract:
Mass spectrometry has become a powerful tool for the identification and relative quantification of proteins within biological samples. Proteomics studies typically employ a “bottom-up” strategy for protein identification (i.e. whole proteins are enzymatically digested into peptides prior to analysis). Two components are crucial for peptide sequence determination using this approach: 1. intact peptide mass (MS^1 scan) and 2. the masses of characteristic peptide fragments generated upon collision with inert gas molecules (tandem MS scan). Automated database searching compares this spectral data against a list of peptide candidates (generated based on prior genetic annotation of the species of interest); the strongest matches result in peptide identifications.

Unfortunately, fast and accurate database searching is often hindered by noisy, complex spectra and/or large lists of candidate peptides. We introduce a strategy which utilizes stable isotope labeling by amino acids in cell culture (SILAC) and two novel post-acquisition filtering algorithms to increase the speed and accuracy of peptide identification. The first algorithm employs an amino acid counting approach to enable peptide identification from intact peptide mass alone. When MS^1 identification isn’t feasible, however, our second algorithm facilitates peptide classification by rapid annotation of fragment ions from tandem MS scans using a machine learning approach.

Understanding Interferon-Stimulated Genes Dynamic by Mathematical Modeling of Liver Cell Transcriptional Response to Interferon Alpha and Beta Stimuli

Authors:
Jialiang Wu, Christopher R Bolen, Steven H Kleinstein, Yale University

Abstract:
Interferons (IFNs) are an integral part of the mammalian innate immune system, and their varied anti-viral and anti-bacterial effector functions are controlled through the activation of several hundred interferon-simulated genes (ISGs). The study of the ISGs and their protein products is important for elucidating antiviral and antitumor mechanisms, and as a general model for cell signaling and communication. Interferon stimulated response element (ISRE), a common DNA motif found in the promoters of ISGs, are responsible for regulating the expression and the amplifying effects of ISGs. In order to understand the mechanics involved in the signaling of IFNs, we use microarray-based gene expression profiling of hepatic Huh7 cells to demonstrate that IFN-α and β create a stable hierarchy of ISG expression, where IFN-β shows significantly higher activity after 8 hours of stimulation. To explain these patterns, we build mathematical models to estimate the production and degradation rates of the mRNAs for the individual ISGs. We use these identified parameters to quantify the characteristics of ISGs, and to provide a numerical measure on how the ISRE location and distance to the ISG promoter affect the ISGs expression.

DNA Methylation Modifications and Role in Fetal Growth Restriction

Author:
Denise McGinnis, Harvard Medical School

Abstract:
Fetal development epigenetic markers are a growing area of research that aims to identify associations between the intrauterine environment and disease (1). Fetal growth restriction (FGR) has been associated with the development of adverse health outcomes such as cardiovascular disease in adulthood (2) with growing evidence that environmental exposures drive these changes (3). This nested study analyzes the DNA methylation patterns in the genomic DNA of four target tissues (umbilical cord blood, placentas, umbilical arteries, and umbilical vein) from 150 newborns currently enrolled in ELEMENT, The Early Life Environment in Mexico and NeuroToxicology Study. Sequencing will be performed using the Illumina 450K human methylation chip to determine the highest ranking CpG sites that predict baseline and environmentally influenced fetal growth. Environmental exposure of lead, air pollution, social stress, and environmental tobacco smoke are collected by the ELEMENT study at developmental time points starting in the 1st trimester of pregnancy. These exposures are known risk factors for reduced fetal growth (4-6)and offer the potential of testing whether environmental exposures on fetal growth are mediated by DNA methylation modifications. This study provides a unique opportunity to analyze the methylomics of fetal tissue offering insight into gene-environment based mechanisms of adult disease.

Commonality of Frequently Mutated Functional Domains Across Cancers

Authors:
Huy Vuong, Peilin Jia, Zhongming Zhao, Vanderbilt University

Abstract:
Given the large number of cancer mutations detected by next generation sequencing studies, one central aim in cancer research is to identify recurrent driver mutations that are critical to cancer phenotypes. Current approach is primarily gene-based, which defines driver mutations as both occurring in frequently mutated genes and being involved in cancer-associated pathways. Recent studies have reported several limitations of this gene-based approach and proposed an alternative domain-based approach by considering the position of mutations within conserved domains[3].

By applying this domain-based approach to the Catalogue of Somatic Mutations In Cancer (COSMIC) database[4], we mapped over 630,000 SNVs and indels to ~12,000 domains in the Pfam[5] database and identified significantly mutated genes and domains across 42 tumor types with false discovery rate below 0.1. We showed the number of shared domains between two tumor types varied significantly due to the inherently complex heterogeneity between cancer phenotypes. We also performed hierarchical clustering using the Jaccard similarity coefficient to identify closely related tumors and GO term enrichment analyses. Preliminary results indicated the top shared domains included trans-membrane, DNA binding and kinase domains. Thus, our analysis demonstrates the domain-based approach can provide testable hypotheses for driver mutations within significantly mutated domains in cancer.

[3] Nehrt N, Peterson T, Park D, Kann M: **Domain landscapes of somatic mutations in cancer.** *BMC genomics* 2011, **13 Suppl 4**:S9.

[4] Forbes S, Bindal N, Bamford S, Cole C, Kok C, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague J, Campbell P, Stratton M, Futreal P: **COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.** *Nucleic acids research*2010, **39**:D945–50.

[5] Finn R, Tate J, Mistry J, Coggill P, Sammut S, Hotz H-R, Ceric G, Forslund K, Eddy S, Sonnhammer EL, Bateman A: **The Pfam protein families database.** *Nucleic acids research* 2007, **36**:D281–8.

RPPAth: Integrate, Visualize, and Model Time-series Data in Cancer Cell Lines

Authors:
Ted Laderas, Laura Heiser, Joe Gray and Kemal Sonmez, Oregon Health & Science University

Abstract:
Reverse Phase Protein Arrays (RPPA) are a technique that allows biologists to examine the timecourse response of signaling proteins and their phosphorylated states in response to a stimulus. However, this set of proteins spans a wide range of signaling pathways and determining which pathways are activated is difficult. We present a framework called RPPAth that 1) highlights statistically significant pathway activity, 2) visualizes the timecourse response onto pathway and network structures, and 3) integrates known genomic alterations through protein-protein interactions. This framework acts as a bridge between the RPPA data and constructing dynamic ODE-based models. We demonstrate RPPAth's utility on a breast cancer cell line experiment with the drug Lapatinib applied as a stimulus. RPPAth allows for between-cell line comparison, highlighting pathways that are significantly activated in Lapatinib response. Visualizing RPPA data on known pathways shows concordance and discordance between pathways and the data. Finally, integrating known genomic alterations highlights possible targets for combination drug therapies. We show the utility of RPPAth by building a dynamic ODE model of the high sensitivity cell line response. Although the framework is applied towards RPPA data, RPPAth can also be used to visualize pathway/network response of other timecourse data such as microarray data.

The Application of Network Label Propagation for Biomarker Ranking in Genome-Wide Data

Authors:
Matthew E Stokes and Shyam Visweswaran, University of Pittsburgh

Abstract:
Typically, univariate ranking methods (such as chi square) are used to rank biomarkers that are associated with a phenotype or disease in high-dimensional genome-wide single nucleotide polymorphism (SNP) data. While such methods are computationally efficient they suffer from several drawbacks including poor reproducibility of top-ranked biomarkers across datasets and inability to account for interactions among genetic variants.

To address these challenges we applied a computationally efficient semi-supervised label propagation (LP) algorithm to rank SNPs in genome-wide association (GWA) datasets. LP represents SNPs and samples in GWA data as nodes in a bipartite graph and propagates information about nodes to neighboring nodes through the graph edges. We applied LP to over 100,000 SNPs in each of two Alzheimer's disease GWA datasets, and evaluated the top-ranked SNPs for predictive performance and reproducibility between datasets.

Compared to chi square (a univariate ranking method) and ReliefF (a multivariate ranking method), the top-ranked SNPs obtained from LP had significantly better predictive performance when used in a k-nearest neighbor classifier. Moreover, with LP significantly more SNPs were found to be common in the top-ranked SNPs between the two datasets. Our results suggest that application of LP to GWA data has strong potential to improve the effectiveness of ranking of SNPs.

Identification of Transcribed Pseudogenes in Human Genome

Authors:
Baikang Pei[1], Cristina Sisu[1], Alexandre Reymond[2], Jennifer Harrow[3], Mark Gerstein[1]
[1]Yale University, [2]University of Lausanne, [3]Wellcome Trust Sanger Institute

Abstract:
Pseudogenes are genomic loci with high sequence similarity to functional genes but lacking coding potential due to the presence of disruptive mutations. They have long been considered as nonfunctional sequences. However, recent evidence suggests some pseudogenes can perform regulatory roles through their RNA products. One example is the pseudogene of PTEN, a crucial tumor suppressor, resurrects as lncRNA and regulates the expression of its parent gene.

To examine the pervasiveness of pseudogene activity, we identified pseudogene transcription on a genome-wide scale by using GENCODE annotation of human pseudogenes and RNA-Seq data from BodyMap tissues and ENCODE cell lines. We showed about 10% of human pseudogenes are potentially transcribed, and interestingly, a subset of these show discordant expression patterns compared to their parents. Pseudogene transcription exhibits tissue specificity, and very few pseudogenes are broadly transcribed in multiple tissues and cell lines. Some transcribed pseudogenes predicted with high-throughput data were further evaluated by experiments specifically targeted to those loci. The transcription information of each pseudogene, together with other functional genomics data, are stored in an resource file, psiDR, which will be useful for the initial identification of potentially functional pseudogenes.

PheWAS – A Novel Method that Combines Genetics and Medical Informatics

Authors:
Scott J Hebbring, Steven J Schrodi, Zhiyi Zhou, Zhan Ye, Murray H Brilliant, David C Page, University of Wisconsin-Madison and Marshfield Clinic

Abstract:
A priority in research is to understand the genetics and etiology of common diseases so that genetics may be used in "personalized medicine." As such, significant resources have been dedicated to the study of complex diseases through genome-wide association studies (GWAS). However, GWAS has provided few medically actionable results. To address this challenge, we applied a novel method that combines genetics with the power of medical informatics. This method is called the Phenome-Wide Association Study (PheWAS). In our PheWAS, we genotyped 120 SNPs, selected for biological function and/or known disease involvement, in 4,235 Marshfield Clinic patients. Each SNP was tested for association with 4,841 clinically significant phenotypes defined by ICD9 coding from patient electronic medical records (EMRs). Numerous significant findings were observed. For every SNP that had pre-described disease associations, that SNP was significantly associated ($p<0.05$) with the ICD9 code(s) that defined the pre-described disease – demonstrating proof-of-principle that PheWAS can be used to rediscover gene-disease associations. In addition to these associations, many novel clinically significant phenotypes were also associated and replicated. This study not only builds on the feasibility/utility of the powerful PheWAS approach, but demonstrates that PheWAS may help expand our understanding of the genetic etiology of common diseases.

Physician Handoffs and Cross-Cover Chart Biopsy: A Descriptive Approach

Authors:
Logan Kendall, Katherine Blondon, Justin Iwasaki, Jennifer Best, Andrew White, University of Washington, University of Washington Medical Center

Abstract:
Background. Clinical handoffs involve a rapid transfer of information from one provider or team to another, through processes that may introduce communication errors and affect care delivery. The process of cross-covering in particular requires quickly getting up to speed on an unfamiliar set of patients with management plans established by another medical team.

Aim. In this study, we describe the information seeking approaches employed by physicians from internal and family medicine within an electronic medical record system at a major academic medical center.

Method. We conducted simulated handoff sessions and interviews with 21 physicians across a set of standardized patient cases. We collected screen capture data, mouse movements, and audio recordings of their use of an electronic medical record to perform rapid chart biopsies.

Results. Our findings describe physicians' information seeking pathways and strategies during chart review and their responses to interruptions created during the simulation, stratified by physician experience.

Implications. Understanding how physicians seek out and assimilate data about patients can inform the design of handoff tools and suggest strategies for explicitly supporting chart biopsies within the EMR.

Technology and Communications Between Providers – A Frontline Report

Authors:
Dian Chase, Jennifer Hall, Deborah Cohen, Joan Ash and David Dorr, Oregon Health and Science University

Abstract:
Poor communication and collaboration between providers is a root cause for the majority of sentinel (adverse) events in the U.S. Ideally, health information technology (HIT) can support inter-provider communication and improve patient outcomes. This qualitative study (part of a larger study on EHR safety) examined communication processes at five sites with mature EHR implementations. These sites were broadly representative of U.S. healthcare as they differed in size, region, ownership, and composition. Data were collected using semi-structured interviews and field observation. Findings include problems and successes using HIT to support inter-provider communication, the variety of available communication channels, use of technology to support referrals, medication reconciliation, and discharge, and experiences with Patient Centered Medical Homes. Selecting and using the right communication channel is important to effective patient care. Developers will need to further evolve HIT related systems to support newer models of collaboration.

Identification of Gene-Level Associations in Exome Data: Application to Warfarin

Authors:
Sara Hillenmeyer, Stanford University, Roxana Daneshjou, Stanford University, Luke Miratrix, Harvard University, Konrad Karczewski, Stanford University, Russ Altman, Stanford University

Abstract:
The anticoagulant warfarin has a narrow therapeutic window and severe thrombotic and hemorrhagic consequences for under- and over-dosing. Inter-patient dose varies more than 10 fold, and patients require weeks of monitoring and dose adjustment to achieve stable anticoagulation. Predicting stable dose using clinical and genetic covariates reduces the incidence of adverse events and shortens the time to stable anticoagulation. Although genetic algorithms explain up to 55% of warfarin dose variation in European and Asian patients, these algorithms explain less than 45% of dose variation in African Americans. We sequenced the exomes of African Americans with both high and low stable warfarin doses. Using these exome data, I have developed a method that aggregates rare and common variants from a gene into a single gene score. Including the scores from the top five genes in the dosing algorithm improves dosing prediction in African Americans. When tested in a held out dataset, this dose prediction algorithm has average squared error that is half of the previous best model.

The p53 Transcriptome

Authors:
Allen, Mary A[1], Freeman, JA[2], Mellert, H[2], Espinosa, JM[2], Dowell, RD[3]
1. Computational Bioscience Program at the University of Colorado Anschutz Medical Campus, 2. Molecular, Cellular and Developmental Biology at the University of Colorado-Boulder and HHMI, 3. Molecular, Cellular and Developmental Biology and BioFrontiers at the University of Colorado-Boulder

Abstract:
The guardian of the genome is a transcription factor p53 that can activate transcription of target genes in response to cellular stress. p53 plays a central role in the suppression of cancer. Yet, despite the vast amount of research on p53, the precise transcriptional response to activating p53 is not understood. I have used Global Nuclear Run On-sequencing (GRO-seq) to directly measure changes in transcription genome-wide at early time points after p53 activation. My data contains several exciting results. Surprisingly, only one hour after p53 activation, transcription of several well-known target genes is up-regulated, something novel and undetectable by traditional methods. In addition, many previously unannotated transcripts are regulated by p53, including a intriguing percentage of potential p53 binding sites that are transcribed. These data demonstrate the potential of this approach to yield new insights into p53 and cancer regulation via p53. And open the field to new questions about the effect of transcription over a p53-binding site.

Multi-Omic Co-Expression Network Signatures of Disease

Authors:
David L Gibbs, Shannon K McWeeney, Oregon Health and Science University

Abstract:
Motivation: Integrative multi-omic methods are needed to analyze high dimensional data sets to gain insight into the complex and dynamic relationship among diverse cellular components both for mechanistic studies, as well as to develop integrated biomarkers for disease processes.
Results: Using publically available data measuring the temporal host response to SARS-CoV infection in mice, we analyzed the relationship between the transcriptome and proteome in terms of co-expression structure. While there is significant overlap at the sub-network (or module) level across data types, there is little concordance with regard to the rankings of individual nodes (based on key network measures such as centrality and connectivity), making prioritization across individual analyses difficult. We proposed a novel strategy for joining independent transcript and peptide level co-expression modules that results in an integrated omic signature of disease progression, allowing downstream prioritization across data types. Integrated modules, in addition to showing related functional enrichments, suggest novel pathways activated in response to infection. These disease and platform independent methods are a way forward to realize the full potential of multi-omic network signatures of disease.

Pleiotropy and Polygenes in Adaptive Hybridized Gene Clusters in Mice

Authors:
Kevin J Liu, Ying Song, Michael H Kohn, Luay K Nakhleh, Rice University

Abstract:
In 2011, Song *et al*. (Current Biology 21(15):1296-1301) first reported natural hybridization of the VKORC1 gene from *Mus spretus* into *Mus musculus*, conferring rodenticide resistance. The canonical rodenticide of choice is warfarin - the most widely used anticoagulant in clinical practice - and human orthologs to VKORC1 have been successfully targeted to personalize warfarin therapy.

Major questions still remain. Are adaptive hybridized genes other than VKORC1 present in the mouse genome? If so, what do they do and what else can they tell us about human health?

We created a new computational pipeline to investigate the first question. Between one-tenth and one-fifth of each mouse autosome displayed patterns of hybridization, much of it adaptive. Our genomic catalog can inform the principled construction of new laboratory strains, similar to the Collaborative Cross.

Regarding the second question, we detected adaptive co-hybridization of large, densely connected interacting gene clusters with translationally promising functional roles. We also examined two mysteries concerning pleiotropic and polygenic effects of adaptive hybridization: how can mutations conferring anticoagulant resistance persist despite deleterious side effects, and why was VKORC1 hybridized but not genes in related pathways? Our answer relied on a fine-scale comparative association study using natural variational data.

Using Distributed Data to Enhance Predictive Models

Authors:
Eric Levy, Xiaoqian Jiang, Jihoon Kim, Lucila Ohno-Machado, University of California, San Diego

Abstract:
It is important to share both clinical and genomic data in order to obtain sample sizes that are large enough to classify complex or rare patterns. However, there are many institutional and legal barriers that make it difficult or impossible to create pooled datasets, which could facilitate construction of more robust models. For these reasons, we would like to create models that can work with distributed data, and only operate on summary statistics.

We have already shown that a distributed binary logistic regression model can produce the same results as a centralized one. Support vector machines with linear kernels can also be distributed. We will develop distributed non-linear SVMs. In addition, we would like to ensure our approach scales and preserves privacy with high-dimensional data.

We will demonstrate the use of these models for prediction of pre-microRNA sequences from short read sequencing data.

Chronic Noncancer Pain in Iraq/Afghanistan Veterans

Authors:
Lawrence Lyon, Kenric Hammond, Department of Veterans Affairs, Puget Sound Healthcare System

Abstract:
Objective: Describe and characterize treatment of Iraq/Afghanistan Veterans (IAV) with chronic noncancer pain (CNCP) at one VA facility.

Methods: Fiscal years 2003-2012 primary care patient data were extracted, synthesized and analyzed from a data warehouse using SQL queries. Four markers of CNCP were developed: a chronically painful diagnosis (CPDx), chronic opioid therapy (COT), chronic elevation of pain score (CPS), and documented pain agreement (DPA).

Results: Of 4599 patients in our cohort, 48% had a CPDx; 32% of patients with CPDx were treated with COT; 54% of those lacked a DPA. Half of CPDx patients had CNCP when classified by CPS. Painful diagnosis distribution was similar in all groups. 7% of COT patients' doses exceeded 120 mg/day morphine equivalent. Hydrocodone/Acetaminophen (HC-A) was the most commonly prescribed chronic opioid.

Discussion: CPDx and COT are the most useful parameters characterizing this patient population. CPS is of limited use in establishing suitability of COT and monitoring outcome.

Conclusion: Chronic noncancer pain characterizes almost half of these IAV patients. A chronically painful diagnosis is the most inclusive parameter, whereas one based on COT is most relevant to safety. Patient function, a key factor for monitoring pain therapy outcome, cannot be determined from available EHR data.

Deriving a Modified Charlson Comorbidity Index from Clinical Narratives Using Natural Language Processing

Authors:
Hojjat Salmasian, Daniel Freedberg, Carol Friedman, Columbia University

Abstract:
The Charlson index, a vital tool for evaluating patient comorbidities, is a well-validated composite measure formulated by assigning weights from 1 to 6 for 19 chronic medical conditions and is usually computed via manual chart review. We used a natural language processing (NLP) system, MedLEE, to structure information in inpatient notes and developed queries to identify the concepts representing each Charlson condition from the structured data. We computed a modified Charlson index and tested it in a study concerning the association between proton pump inhibitors (PPIs) and recurrence of *Clostridium difficile*–associated diarrhea (CDAD). Using admission notes for 894 inpatients, we successfully extracted the data corresponding to 17 of 19 (89%) items in the index. The mean index for inpatients was 3.0 (standard deviation = 3.1). As anticipated, a higher Charlson index was significantly associated with a higher probability of recurrent CDAD after adjusting for PPIs and other factors (Hazard Ratio 1.09, 95%CI 1.04-1.14). After adjusting for the NLP-derived Charlson index, the study showed no association between PPIs and recurrent CDAD (HR 0.83, 95% CI 0.58-1.17). We conclude that comorbidity scores derived using NLP can be utilized to automatically adjust the outcome of clinical studies for the effect of comorbidities.

Decision Support System to Improve Positive Predictive Value in Mammography

Authors:
Francisco Gimenez, Daniel Rubin, Stanford University

Abstract:
Breast cancer is the most prevalent cancer amongst women and their second leading cause of cancer deaths. As a result, the American Cancer Society recommends that all women older than 40 receive yearly mammograms to screen for breast cancer. Despite this recommendation and a reduction in breast cancer mortality, there is controversy over the efficacy of mammography screening. One well studied factor contributing to suboptimal care is the variability of radiologists in screening mammography. Such suboptimal care can lead to (1) false positive cancer detections, which lead to unnecessary and invasive testing and (2) false negative cancer detections, which lead to missed cancers during their most treatable stages. Furthermore, these errors are inversely proportional, where improving one involves aggravating the other. We propose using a probabilistic decision support system to improve the positive predictive value (PPV) of radiologists without significantly impacting false negative detections. This system achieved a 34% increase in PPV of radiologists with no significant effect on the false negative rate.

Decision Support Impact on Physician Adherence to Appropriate Use of ED Imaging

Authors:
Anurag Gupta, Ali S Raja, Ivan K Ip, Angela M Mills, Ramin Khorasani, Harvard Medical School

Abstract:
Background: While clinical decision support (CDS) has been shown to decrease the overall use of CT pulmonary angiography (CTPA) in ED patients with suspected pulmonary embolism (PE), its effect on physician adherence to evidence-based guidelines (EBG) is yet unknown.

Methods: This prospective study was performed in our ED, located in a 777-bed, academic Level-1 trauma center. We evaluated the 12-month periods prior and subsequent to the quarter during which a CDS, based on Well's criteria, was implemented. 200 random records were reviewed (100 pre and post) based on a sample size calculation to detect a 20% effect size with power of 0.8 (alpha = 0.05) and an estimated baseline proportion of 70%. We used chi-square tests with proportional analyses to assess pre- and post-intervention differences.

Results: 1,155 patients with suspected PE were evaluated by CTPA during the 12 month period prior to implementation of CDS (9/1/09-8/31/10), and 1,292 patients in the 12 months following (12/1/10-11/30/11). Adherence to EBG increased from 55% to 73% (p=0.008).

Conclusion: Implementation of CDS significantly increases adherence to EBG for the use of CT in ED patients with suspected PE. Even with CDS, nearly one quarter of CTPA studies performed deviated from EBG.

A New Method to Store, Classify, and Retrieve Electrophysiological Signals

Authors:
Antonio García-Hernández[1], Luis V Colom[2], Julio C Facelli[1], [1]University of Utah; [2]UT-Brownsville

Abstract:
Electrophysiological signals (ES) displays activity that allows physicians and researchers to differentiate between physiological and pathological conditions. ES are non-stationary and fast transient; hard to characterize and describe, making it difficult to compare to each other inside a database. ES data require high sampling rates to be digitally recorded and stored; therefore, the potential benefits for re-use and sharing these data cannot be exploited by most physicians and researchers. To overcome these limitations a signal could be decomposed over basis functions (wavelets) that are well concentrated in time and frequency. Wavelet theory has had a growing impact on signal processing theory, but has not been applied to this problem. We propose to use wavelet transforms to convent analog neuro-electrophysiological signals into a discrete representation. These discrete representations will be used to demonstrate how to build a neuro-electrophysiological database that takes digitized ES (raw data) from the user and returns a rank list of similar signals already existing in the data base. This new method will allow biomedical researchers to group patients with similar electrophysiological pathological recordings (*i.e.* cardiac and neurological patients), which can be used for different biomedical research, which includes but not limited to genetic, epidemiological, and/or pharmaceutical studies.

Modeling a Decision Rule to Guide CT Usage in Pediatric Blunt Abdominal Trauma

Authors:
Edna C Shenvi, Adela Grando, Mary Hilfiker, Hyeon-Eui Kim, University of California, San Diego

Abstract:
Diagnostic computed tomography (CT) exposes patients to large radiation doses, and a decision rule[i] has been developed based on clinical history and exam findings to identify children with blunt abdominal trauma in whom CT may not be necessary. We modeled the components of this decision rule as an ontology in Web Ontology Language (OWL) format using Protégé[ii] with three output recommendations (CT recommended, consider CT, and CT not necessary). We populated it with eleven scenarios provided by a pediatric trauma surgeon and tested the decision rule using an OWL-DL (description logic) reasoner called FaCT++. The output yielded ten of eleven in perfect agreement with clinical expert opinion, demonstrating the feasibility of incorporating this rule in a clinical decision support (CDS) system. We intend to further refine this rule to capture subtle differences in severity of trauma, test it with actual patient cases, and implement a prospective CDS study at a pediatric hospital.

Random Forest Classification of Acute Coronary Syndromes

Authors:
Jacob P VanHouten, John M Starmer, Thomas A Lasko, Vanderbilt University

Abstract:
Acute Coronary Syndrome (ACS) is a collection of diseases related to acute myocardial ischemia, and includes unstable angina and myocardial infarction. Missing a diagnosis of ACS can lead to potentially life-threatening health consequences; as a result, many patients who present to the emergency department with a chief complaint of chest pain are subjected to an expensive battery of tests. Up to 85% of these presenting patients do not receive a final diagnosis of ACS, but despite the extensive testing, 2 to 8% of patients with myocardial infarction are included among them and mistakenly sent home. Better methods of risk estimation are necessary for improved diagnostic accuracy and more efficient use of scarce medical resources. Using a dataset of de-identified patient records, we explored the relationship between electronically available clinical observations and an ultimate diagnosis of ACS. Using a random forest model, we achieved an AUC of 0.851 under cross validation for predicting the presence of ACS. Further work is needed before this algorithm will be clinically useful.

Vital Sign Quality Assessment Using Ordinal Regression of Time Series Data

Authors:
Risa B Myers[1], John C Frenzel[2], Christopher M Jermaine[1], [1]Rice University, [2]University of Texas MD Anderson Cancer Center

Abstract:
With the increased use of EHRs, it is important to develop new approaches to abstracting patient data to provide clinical decision support. In particular, we look at modeling vital signs. Such data are typically represented as a time series. Our goal is to label these time series with one or more integer labels that are equivalent to an expert-supplied evaluation of the quality, volatility, or level of alarm that should be associated with the time series.

Unfortunately, we do not expect classical shape- or pattern-based time-series classification methodologies to perform well, as, in practice, human experts assessing vital sign data seem to be evaluating the statistical properties of the series, rather than searching for specific patterns. We propose a novel, Bayesian statistical model, the AR-OR model, for labeling time series data and compare our approach to traditional time series classification methods.

This project is supported in part by a training fellowship from the Keck Center NLM Training Program in Biomedical Informatics of the Gulf Coast Consortia (NLM Grant No. T15LM007093).

Data Accuracy in Journal Abstracts for Use in Informing Clinical Decisions

Authors:
Raymond Francis Sarmiento, Alex Gavino, Paul Fontelo, National Library of Medicine, NIH

Abstract:
The abstract is the most frequently read section of a research article. Using consensus abstracts for informing clinical decisions was recently proposed; however, inaccuracies between abstracts and corresponding full-text articles have been previously observed. Through the years, efforts have been made to improve quality. We compared data between 60 abstracts and full-text articles from six highly read medical journals. Data inaccuracies were identified then classified as either clinically significant or not significant. Discrepancies were observed in 53.33% of articles ranging from 3.33% to 45.00% based on the IMRAD structure. The Results section showed the highest discrepancies (45.00%) although these were deemed to be mostly clinically not significant except in one. The two most common discrepancies were mismatched numbers or percentages (11.67%), and numerical data or calculations in abstracts but not mentioned in the full-text (35.00%). There was no significant relationship between journals and discrepancies (Fisher's exact p-value = 0.3405). Although we still found a high percentage of inaccuracies between abstracts and full-text articles, these were not significant clinically. The inaccuracies do not seem to affect the conclusion and overall interpretation. We found structured abstracts to be informative and may be useful to practitioners as another resource for guiding clinical decisions.

Characterizing MS Disability with Accelerometer-Based Motion Capture in the Field

Authors:
Matthew Engelhard, Stephen D Patek, John Lach, Myla D Goldman – University of Virginia
Kristina Sheridan – MITRE Corporation

Abstract:
Gait impairment is common in multiple sclerosis (MS). It reduces mobility and limits independence, leading to decreased quality of life. Accelerometer data gathered during clinical walking tests has been shown to provide meaningful information about gait impairment in MS, distinguishing between subjects at different levels of disability. Because it requires only a single pocket-sized device that may be used in an unsupervised fashion, collection of accelerometer data is promising as a mobile health assessment tool contributing to a better understanding of the factors that relate to disease progression.

To investigate the utility of long-term field collection of accelerometer data, we have designed a pilot study that will evaluate two different consumer-grade wearable accelerometer systems in a broad MS cohort over a six-month period. We will attempt to determine subject disability level using acceleration data alone, without an accompanying activity record. Patient-reported outcomes will be collected through an online interface, accessible to both patients and their physicians.

Post-study surveys will assess the value of the resulting data to patients and physicians. We believe that by supplementing patient reports with objective measures of community mobility, we may help patients better understand their disease process and assist physicians in making informed clinical decisions.

Grid and Public Health: Using Grid Service to Share Resources

Authors:
Kailah T Davis, Julio Facelli, Department of Biomedical Informatics, University of Utah

Abstract:
Real-time public health surveillance systems are crucial for the timely detection and response to public health threats. Historically, public health departments are organized vertically, and system interoperability is limited, leading to numerous "siloed" surveillance systems. This lack of integration creates many inefficiencies in the areas of analysis and communication of information, which become most apparent during public health threats and emergencies. This centralized model has proven to be difficult to maintain and enhance, therefore, here we present the rationale of creating a federated model by leveraging grid technology concepts and tools for the sharing and epidemiological analyses of public health data. As a case study for this approach, we used tools which were developed by the caGrid project team to create a secured virtual organization where users are able to access a grid data service, death certificates from the Utah Department of Health, and two analytical grid services, MetaMap and R, to create a public health surveillance infrastructure that uses data and services under the control of different administrative domains. This research fills an important need of developing prototypes that provides insight on how a public health infrastructure could be developed as a dynamically evolving ecosystem of grid enabled applications.

Early Detection of Statin Adherence: Surveillance for a Quality Measure

Authors:
Andrew J Zimolzak, Kenneth D Mandl, Harvard Medical School

Abstract:
Medication nonadherence costs $300 billion annually in the US. The Centers for Medicare and Medicaid Services (CMS) awards bonus payments to Medicare plans based in part on population adherence to medications. We sought to build an individualized surveillance model that detects early which beneficiaries will fall below the CMS adherence threshold for statins.

We retrospectively studied 210,000 beneficiaries in a private insurance claims database. Regression models were constructed to predict statin nonadherence at one year, using adherence from days 1-90, as well as 15 additional characteristics. In a sensitivity analysis, we varied the number of days of adherence data used for prediction.

Lower adherence in days 1-90 was the strongest associate of one-year nonadherence (odds ratio 25.0, 95% confidence interval 23.6-26.4). Models were highly predictive of one-year nonadherence, with areas under the receiver operating characteristic curve of 0.84 to 0.80, and positive predictive value 87.7%. The sensitivity analysis showed that individualized surveillance improves prediction as early as 31 days after statin initiation.

To preserve their CMS quality ratings, plan managers can use individualized surveillance to identify members who would benefit from adherence improvement programs, recognizing the tradeoff between improvement of model performance over time and the advantage of earlier detection.

The Walkability Index: Modeling Local Weather as a Predictor of Exercise Behavior

Authors:
Daniel Schulman[1] , Timothy Bickmore[2], Michael Paasche-Orlow[3], Katherine Waite[3],
1. Department of Veterans Affairs, VA Boston Healthcare System, 2. Northeastern University,
3. Boston University School of Medicine

Abstract:
The promotion of walking is a common behavioral target as it is effective across a wide range of ability, and objective measures can be easily obtained with pedometers. For example, over 60 ongoing trials registered on ClinicalTrials.gov assess walking by pedometer. Poor weather is among the most commonly reported barriers to daily walking. In this project we will develop a *walkability index* - representing the effects of the quality of local weather conditions on the likelihood and quantity of walking performed.

In a set of recent exercise promotion interventions, a manual walkability index was recorded each day based on weather parameters including temperature, humidity, wind, and precipitation, and taking into account the immediate local weather, recent past weather, and local climactic norms. We will attempt to validate the index and evaluate its predictive power compared to a simple seasonal variable. We will then attempt to develop an automated index with predictive power matching or exceeding the manual index, using publicly available weather data as predictors and applying statistical and machine learning techniques.

Potential applications of the walkability index include an improved ability to analyze results of exercise interventions, and the ability to dynamically tailor intervention content based on forecasted weather.

Building a Preference Repository to Facilitate Shared Decision Making

Authors:
Robert Dunlea, Leslie Lenert, Department of Biomedical informatics, University of Utah

Abstract:
Background:
Shared Decision Making (SDM) is a widely held ideal for clinical practice; however, SDM is difficult to integrate into practice because of the time constraints present in a typical clinical encounter. We propose create a technical infrastructure to integrate patients' preferences and values in EHRs and to drive SDM Clinical Decision Support (CDS) technologies using a standards based approach. The first stage of the research focuses on recording patient preferences and identifying preference subgroups for use with rule based systems and for collaborative decision making. The initial context of our work is choice of a specialist for referral.
Methods: Adaptive conjoint analysis is being used to measure the preferences of 500 patients in the US for eight factors related to the referral process.
Results: The aggregate data indicates that insurance coverage and specialist to primary care physician communication are key factors considered when selecting a specialist. We identified several* common preference pattern subgroups.
Conclusions: The preference patterns identified will serve as a "preference knowledge repository" to drive a SDM CDS system. By identifying a patient's preference type, we hypothesize that we can use that type to support SDM using standard CDS tools such as rules, reminders, and alerts.

Creating a Usability Testing Ontology for Biomedical Information Retrieval Tools

Authors:
Jing Zhang, Rebecca Walker, Hyeon-Eui Kim, University of California, San Diego

Abstract:
With an increasing numbers of biomedical information retrieval tools being developed, usability testing has to become accessible to ensure user satisfaction. *Phenotype Discover* was developed at UCSD as a web-based search tool that allows users to search content in the database of Genotypes and Phenotypes (**dbGaP**). Advanced features were added to support effective information retrieval, based upon user requirement analysis. In order to evaluate the *Phenotype Discover'*s user interface and to address the general challenge of providing a fast, simple, and cost-effective way to perform usability testing, we proposed developing a one-stop-shop tool for system evaluators with the end goal of generating a customized questionnaire-based usability test according to specific goals and constraints.

The proposed research composes of 3 stages:
Stage 1. Constructing the usability testing knowledge base
Stage 2. Building the usability testing tool
Stage 3. Evaluating the usability testing tool

We have currently completed Stage 1 where an ontology was constructed based on key concepts identified in Human Computer Interaction textbooks, usability studies literature, as well as industry best practices. 200 testing questions from usability questionnaires have been included in the ontology. They were annotated and categorized using testing attributes and domains.

Inferring Functional Human Language Pathways

Authors:
Meagan Whaley[1], Steven Cox[1], Nitin Tandon[2], Chris Conner[2]
[1]Rice University, Houston, Texas, [2]University of Texas Health Science Center at Houston

Abstract:
This ongoing research is providing insight about fundamental networks involved in human language processes by applying a statistical method to time series data recorded from intracranial electrodes implanted in human subjects. Language processes involve networks of neurons sending and receiving information through convoluted systems of fiber pathways distributed throughout the cortex, and due to the unavailability of temporally and spatially precise clinical data, understanding these functional pathways has proven to be a fundamental problem in medical and biological research.

This work applied a well-developed, adaptable statistical tool, Granger Causality, to hundreds of time series recordings taken directly from the cortex (electrocorticography or ECoG) of human subjects while they participated in a verb completion task. Granger Causality was used to analyze the precise ECoG data by delivering intelligible results illustrating the direction and relative magnitude of interactions that occurred between time series over periods of time and at specific frequency values. These results are interpreted as indicating how and when the brain regions located underneath the electrodes interact during precise stages of language processes, and thus far, they have been consistent with modern language theory.

This project is supported in part by the NLM Training Program in Biomedical Informatics T15LM007093.

Topic Models for Mortality Modeling in Intensive Care Units

Authors:
Tristan Naumann, Marzyeh Ghassemi, Rohit Joshi, Anna Rumshisky, Peter Szolovits,
Massachusetts Institute of Technology

Abstract:
Mortality prediction is an important problem in the intensive care unit (ICU) because it is helpful for understanding patients' evolving severity, quality of care, and comparing treatments. Most ICU mortality models primarily consider structured data and physiological waveforms (Le Gall et al., 1993). An important limitation of structured data approaches is that they omit much vital information captured in providers' free text notes and reports. We propose an approach to mortality prediction that incorporates the information from free text notes using topic modeling.

Topic models and their variants are becoming popular in inferring the latent structure behind a collection of documents (Blei et al., 2003; Arnold, 2010). The goal of this work is to identify common topics from the unlabeled free-text observations in patients' care notes and to determine whether the distribution of topic membership for each patient can be used as a predictive feature for mortality in-hospital, 30 days post discharge, and 6 months post-discharge when combined with the expect mortality for each topic. Our preliminary results show that SVM classifiers trained on derived topics perform better than the baseline clinical features typically used for patient severity prediction in the ICU.

Crowdsourcing Ontologies

Authors:
Jonathan M Mortensen, Paul R Alexander, Mark A Musen, and Natalya F Noy, Stanford University

Abstract:
Biomedical ontologies are becoming increasingly large and complex. A single user cannot easily develop or maintain them. Researchers have developed various automated techniques to assist with ontology development and engineering at scale. However, these solutions are not always complete. Microtask crowdsourcing, wherein workers are paid small amounts to complete simple, short tasks, may be one technique to alleviate some of the development difficulties. Previously, we developed a method to verify an ontology hierarchy using microtask crowdsourcing. In this work, we investigated the finer details of the design and configuration of a hierarchy- verification task. For example, when we provided definitions and required qualifications, workers performed with 82% accuracy on the hierarchy verification task, compared to 50% without. We showed that to achieve reasonable performance on such a task, workers require context via definitions, tasks require qualifications that select a worker with proper domain knowledge, and a question must be phrased with the least cognitive load (i.e., in the simplest way).

User-Centered Design, Implementation, and Evaluation of a Clinician-focused Pharmacogenomics Information Resource

Authors:
Katrina M Romagnoli, Richard Boyce, Philip Empey, Harry Hochheiser, University of Pittsburgh

Abstract:
Personalized medicine is only possible with comprehensive, relevant, and actionable information about an individual's genetic makeup and medications. Clinicians need clinically relevant information to help them integrate pharmacogenomics, and genetic tests into their practice. Information about pharmacogenomics is dispersed among multiple resources not designed for clinical decision support. Resources integrating information from these diverse sources and presenting it appropriately have the potential to support informed clinical pharmacogenomics decision-making.

As the first step towards developing pharmacogenomics information tools, we developed a model of pharmacogenomics-related information, and then used that model to write a problem scenario illustrating the current workflows of physicians, clinical pharmacists, and nurses during pharmacogenomics-related medication prescription. This model was used by pharmacists to annotate structured product labels of medications about pharmacogenomics-related information.
Five pharmacists annotated 213 pharmacogenomics-related statements in 29 sections of structured product labels. Their annotations demonstrated the possibility of making unstructured, relevant but inaccessible pharmacogenomics information actionable, and verified the model we developed.

We will use the model and problem scenario to inform the design of highly usable information tools to provide clinically relevant pharmacogenomics information to clinicians, as a form of clinical decision support regarding medications that have pharmacogenomics implications.

Characterization of the Biomedical Query Mediation Process

Authors:
Gregory W Hruby, Mary Regina Boland, James J Cimino, Junfeng Gao, Adam B Wilcox, Julia Hirschberg, Chunhua Weng, Columbia University Medical Center

Abstract:
Expanding life science data access for research uses is critical for achieving learning health systems. However, little is known about the query mediation process (QMP).
We transcribed query mediation dialogues between a query expert (QE) and medical researchers (MRs) and used them to develop a classification schema (Problem Statement, Clinical Process, EHR Data Location, Study Design, Research Workflow, Review Data Extraction, IRB and Policy Review, and Confirm Results) for dialogue acts, each being one exchange of speech.
Three raters (G Hruby, M Boland, and J Gao) independently annotated all the 3160 dialogue acts. The kappa inter-rater agreement was 0.61. The minimum, median, and maximum numbers of dialogue acts in a project were 27, 134, 323, respectively. We identified temporal trends in topic flow in the QMP. The following figure shows the topic flows during the QMP. The movement between an abstract description of data needs to a practical data extraction plan is an iterative exchange occurring between these two classes (Study Design and Research Workflow) of dialogue acts.
This study contributes early knowledge of the QMP and confirms that query formulation is not a straightforward translation of a researcher's needs, but rather an iterative process for information needs elaboration.

Evaluation of de Novo Transcriptome Assemblies from RNA-Seq Data

Authors:
Nathanael Fillmore*, Bo Li*, Colin Dewey, University of Wisconsin-Madison
*co-first authors

Abstract:
High-throughput RNA sequencing (RNA-Seq) is a powerful tool for studying a cell's transcriptome, i.e., the collection of RNA transcripts present in the cell. RNA-Seq is especially useful in the de novo setting, when no reference sequences are available beforehand; in this setting, the short RNA-seq reads can be assembled from scratch into putative full-length transcripts by a computer program. However, current computational approaches to de novo RNA-Seq assembly rely on heuristics and ad-hoc parameter settings; different heuristics or parameter settings may result in substantially different assemblies, and it is difficult to judge which assembly is “best”, since in the de novo setting the true sequences are not known.
In order to overcome this problem, we have developed a transcriptome assembly scoring function, which can be used to choose the best assembly from a collection of candidate de novo assemblies, even when no ground-truth reference is available. The score is based on a probability model of the process of RNA-Seq read generation and of ideal transcriptome assembly. We have also developed several simple reference-based scores, and we have used these to carry out a large-scale meta-evaluation of our de novo scoring function on real and simulated data.

Identifying Clonally Related Sequences in Immunoglobulin Repertoires

Authors:
Namita Gupta, Jason A Vander Heiden, Steven H Kleinstein, Yale University

Abstract:
B cells use their Immunoglobulin (Ig) receptors to neutralize foreign pathogens. Characterization of an individual's repertoire of Ig receptors has been shown to have clinical implications, such as response to infection or efficacy of a vaccine. The advent of high-throughput sequencing technologies allows for comprehensive analyses of Ig repertoires, but methods are needed to explore the relationships between large numbers of these sequences as well as further impacts on human health. A key step in such analyses is identifying clonally related sequences, or Ig receptors that share a common ancestor. The difficulty with clonal grouping methods lies in finding biologically accurate criteria to be used in assigning clones. We are developing methods to address these issues and correctly identify B cell clones within an Ig repertoire.

Our methods play a central role in ongoing collaborations with experimentalists and clinicians. One project seeks to determine whether auto-antibodies found in the brains of multiple sclerosis patients are the result of clonal expansion inside or outside of the brain. Another consists of searching a repertoire for antibody sequences found to bind West Nile Virus in single-cell experiments. Using our computational framework we are uncovering new biological information present in human Ig repertoires.

Discovering Functional Modules Using Spectral Clustering and the Gene Ontology

Authors:
Henry A Ogoe, Vanathi Gopalakrishnan, University of Pittsburgh

Abstract:
Recent studies have shown that there exist a many-to-many relationship between certain diseases and causative genes, which are functionally related. These functionally related genes, which we refer to as functional modules (FMs), work in tandem to drive a biochemical pathway or processes. Several studies have linked the dysfunctional of FMs to some diseases. Revealing FMs within a list of genes in an automated fashion could help researchers gain insight into the etiology of diseases.

We conducted an exploratory study on a method to identify FMs using spectral clustering in combination with the Gene Ontology. By using genes extracted from some metabolic pathways of *saccharomyces cerevisiae*, we identify FMs that were associated with stress response, cell aging, maintenance of stability, and host of others. In addition, our method revealed pleiotropic genes such as *ZWF1* and a co-occurrence of several families of genes in some FMs. These results support claims by other studies that, for most complex diseases caused by genes, FMs exist.
Results from our study shows that the spectral clustering algorithm in combination with the Gene Ontology can identify FMs. Subject to further refinement, our method could develop into a framework to identify FMs from a list of arbitrary genes.

Gibbon Chromosomal Breakpoints Display Distinctive Epigenetic States

Authors:
Nathan H Lazar, Larry Wilhelm, Elizabeth Terhune, Thomas J Meyer, Lucia Carbone, Oregon Health & Science University

Abstract:
Chromosomal rearrangements (CRs) are gross mutations with dramatic consequences. They may create fusion genes or alter the expression of intact genes (as is common in cancer) or generate reproductive barriers and promote speciation. Despite their biologically relevant role, mechanisms of CRs are still mostly unknown. Although the relationship between chromosomal breaks and genomic features has been investigated, links with epigenetic factors have not yet been explored. Epigenetic modifications are inherited signals not reflected in the DNA sequence that have been associated with genome instability.

We used the gibbon genome as a model to address these questions. Gibbons are arboreal apes found in Southeast Asia which exhibit an unusually high rate of CRs. Next-generation sequencing techniques examining DNA methylation (the addition of methyl groups to cytosine bases) and histone modifications (DNA organizing agents) revealed differences in regions surrounding CRs. Transposable elements, in particular subfamilies of *Alu* active during the time of the radiation of gibbons, showed significant differences in CR regions indicating historically looser DNA packaging and possibly higher activity for these 'jumping genes'. This finding is significant, given the abundance of these retrotransposons in the human genome (10.7%) and the fact that cancer genomes often display disrupted epigenetic states.

Optimizing the Orbitals Score Function for Protein-Ligand Interface Design in Rosetta

Authors:
Morgan Harrell, Steven Combs, Jens Meiler, Vanderbilt University

Abstract:
Proteins that bind small molecules can sequester ligands, stimulate/extinguish signaling pathways, and transport molecules. Control over these events through protein-ligand interface design could yield new *in vivo* diagnostics. Rosetta is a widely-used software suite for modeling macromolecular structures. To model ligand binding, Rosetta calculates short-range interactions with partial covalent character (e.g. hydrogen bonds, salt bridges, and cation-π-interactions). These interactions are determined by properties of the orbitals attached to the interacting functional groups. We developed a method to computationally place orbitals on atoms for the purpose of identifying partial covalent interactions. We aim to show that optimizing the new partial covalent interaction score function will improve docking and design for the protein-ligand interface. We compared the new orbitals score function with three established score functions by designing the protein-ligand interface of 43 structures with each score function. Then, we evaluated design results through a series of metrics: sequence recovery, packing statistics, root mean square distance to native ligand, and the number of unsatisfied hydrogen bonds in each structure.

Evaluating Docking Scoring Methods Using the 2012 CSAR Benchmark

Authors:
Sam Z Grinter, Chengfei Yan, Sheng-You Huang, Lin Jiang, and Xiaoqin Zou, University of Missouri–Columbia

Abstract:
Computational methods of reliably predicting protein-ligand interactions would have a great impact on drug design. Publicly-accessible databases that pair accurate protein-ligand structures with their associated binding affinities are invaluable tools for assessing the docking methods used to predict such interactions. In this work, we use the 2012 Community Structure-Activity Resource (CSAR) to evaluate ITScore and STScore, two knowledge-based scoring functions developed in our laboratory. We also evaluate a simple force-field-based potential as a reference. The full CSAR Dataset is used to evaluate these scoring functions on binding affinity prediction and active/inactive compound discrimination. The CSAR subset that includes crystal structures is also used to evaluate the scoring functions' binding mode predictions. We investigate the importance of accurate ligand and protein conformational sampling and find that the binding affinity predictions are less sensitive to the bound ligand and protein conformations than the binding mode predictions. We also show the full CSAR dataset to be more challenging in making binding mode predictions than the structure subset. We offer scripts that prepare the CSAR dataset for large-scale docking studies to the academic community. We also provide perspectives for future work.

Patient, Caregiver, and Provider Perceptions of a Colon Cancer Personal Health Record

Authors:
Thomas A Carr, Michael Weiner, David A Haggstrom, Department of Veterans Affairs, Richard L Roudebush VA Medical Center

Abstract:
Personal Health Records (PHRs) are useful for communication, information exchange, and disease self-management. With the appropriate access architecture, a PHR could be used by patients, caregivers, and healthcare providers. As each group's role in care is different, their perception of the functions and usefulness of a PHR might also vary. In this qualitative study, we explored each group's perception of a colorectal cancer (CRC) PHR prototype. Participants completed scenario-based testing across eight use cases, and semi-structured follow-up interviews after each scenario. Video/audio tapes were collected in the setting of a Human-Computer Interaction laboratory. Veteran cancer patients (n=6), their caregivers (n=6), and VHA providers (n=7) were enrolled. Discrete observations were transcribed and underwent grounded theory affinity analysis to identify themes. All groups agreed with the added value of tethering the PHR to an electronic health record; usefulness of tracking treatment and follow-up testing; and the most appropriate uses of secure messaging. Patients and caregivers valued the journal as a memory aid, tool for reflection, and vehicle for non-verbal communication. Healthcare providers were concerned about accuracy of patient-entered data and time burden for both the journal and secure messages. While perceptions differed by role, all groups found significant value in the PHR.

Looking for Paradigm Shifts in the Biomedical Literature

Authors:
Bastien Rance, Michael Cairelli, Olivier Bodenreider, National Library of Medicine, NIH

Abstract:
Contradictive statements are not uncommon in the biomedical literature. For example, “exercise” *cause* “fall” in the 90s and *prevent* them in the 2010s. Some of these changes may simply reflect different experimental conditions, while others indicate paradigm shifts.

Paradigm shifts are significant change in trend between positive and negative statements over time. We derive such statements from predications extracted from the biomedical literature by SemRep, e.g., interferon *inhibits* TNFRSF10B (positive) vs. interferon *stimulates* TNFRSF10B (negative). We selected 8 pairs of contradictive predicates (e.g., *inhibits* / *stimulates*) and studied the evolution of positive and negative statements between 1950 and 2012. We used linear regression to model trends. Sharp slopes, ascending or descending, suggest paradigm shift candidates. The p-value assesses the fitness of the model.

Of the 65M predicates extracted from 20M of MEDLINE titles and abstracts, 3.3M contradictive statements were identified, among which we found 185 paradigm shift candidates. For example, arginine *stimulates* nitric oxide (vs. *does not stimulates*) gradually switched from a majority of negative statements (65% in 1990) to a majority of positive statements (80% in 2012), with a p-value of 1.8x10e-9. The other candidates are under review.

Defining Completeness of Electronic Health Records for Secondary Use: How Big is Your Database?

Authors:
Nicole G Weiskopf, George Hripcsak, Sushmita Swaminathan, Chunhua Weng, Columbia University

Abstract:
To demonstrate the importance of explicit definitions of electronic health record (EHR) data completeness and how different conceptualizations of completeness may impact findings from EHR-derived datasets.
We derived four prototypical definitions of EHR data completeness from the literature on EHR data quality: documentation, breadth, density, and predictive completeness. Each definition dictated a different method of assessing completeness. These methods were computationally applied to representative data including visit information, diagnoses, narrative notes, laboratory results, medication orders, and basic demographic information from New York - Presbyterian Hospital's clinical data warehouse.
This study may have important repercussions for researchers and clinicians engaged in the secondary use of EHR data. The maximum percentage of complete records ranged from 18.5% to 55.4%, depending upon the definition of completeness used. According to any definition, no more than approximately half of records could be considered complete. The proportion that met criteria for completeness was heavily dependent on the definition of completeness used, which is determined by the research task at hand. Lastly, different subsets of records were selected for different goals.
We urge data consumers to be explicit in how they define a complete record and transparent about the limitations of their data.

Data Driven Optimization of Gene Expression in Gynecological Cancer

Authors:
Kevin K McDade, Uma Chandran, Roger Day, University of Pittsburgh

Abstract:
One of the most central assumptions in the clinical relevance of cancer gene expression studies is that low expression of tumor suppressor mRNA or high expression of oncogene mRNA is indicative of disease stage/diagnosis. This assumption overlooks the high complexity in post-transcriptional regulation since protein expression, not mRNA expression, is more often the molecular etiology of disease. A more appropriate question may be: which mRNA expression data points are concordant with the functional protein expression? We utilize two gynecological data sets; 1) 98 endometrial samples with LC-MS/MS proteomic data and Affymetrix U1332Plus expression data, 2) 538 ovarian serous cystadenocarcinoma samples with reverse phase protein assay and Affymetrix U133A mRNA TCGA data. Correlations, standard deviation and bias are determined through a 200 replicate bootstrapping procedure on 887 mRNA-protein pairs using the IdMappingAnalysis R package. The Expectation-Maximization algorithm determines the posterior probability of concordance between mRNA-protein pairs based on bootstrap values.

In order to determine the optimal set of pairs that represents concordance, 9 expression bioinformatics resources such as ENCODE, Geneannot, and Jetset are utilized in the model selection. A greedy search through all intersections and unions of the 9 resources maximizes the estimated proportion of concordance from .303 to .503 after 6 decision tree levels, giving a 20% increase in molecular concordance for the endometrial dataset.

Arpeggio: Separation of Technical and Biological Variability in ChIP-Seq for Classification and Peak Detection

Authors:
Kelly Stanton (1), Fabio Parisi (1), Francesco Strino (1), Neta Rabin (1), Patrik Asp (2), and Yuval Kluger (1), (1) Yale University, (2) Albert Einstein College of Medicine

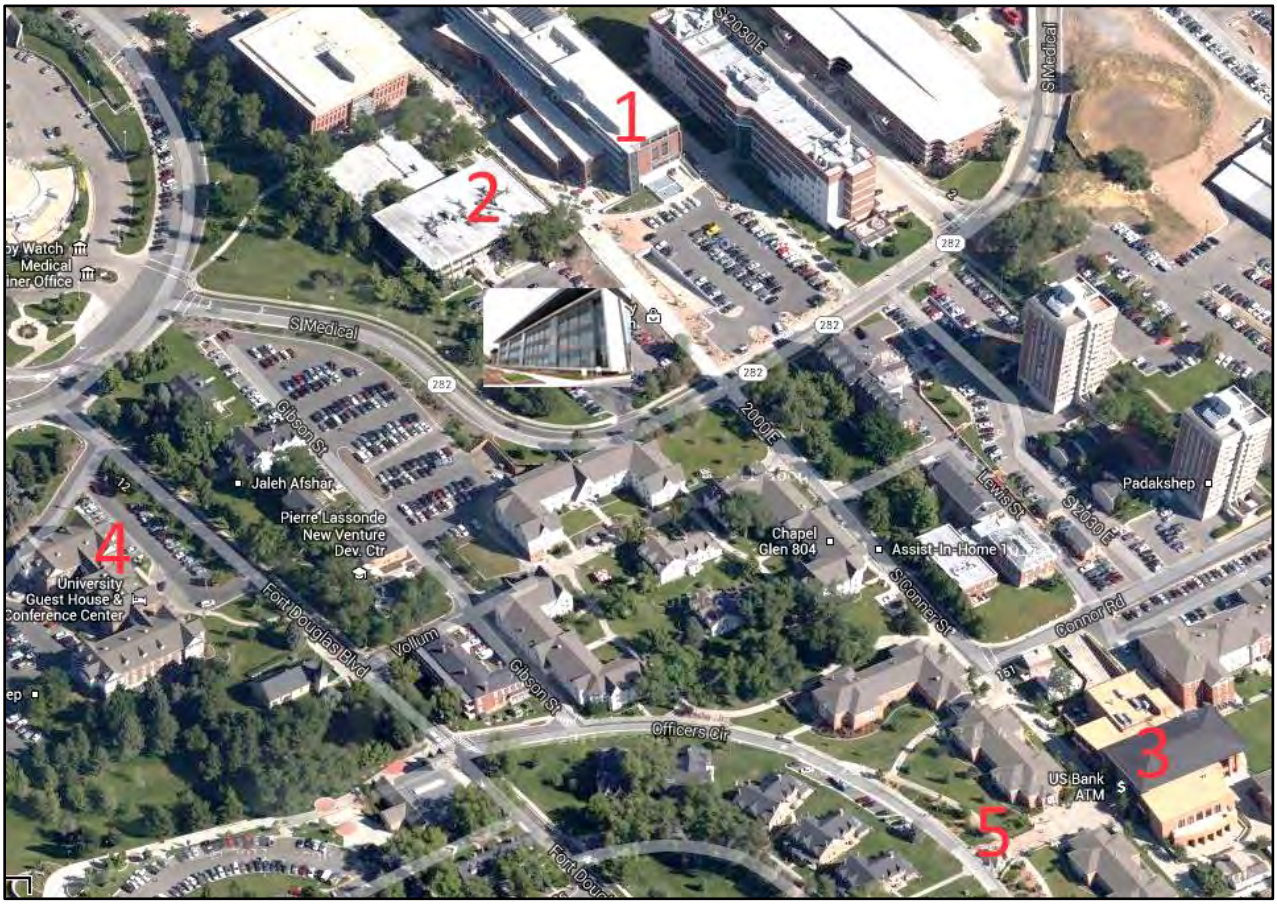
Abstract:
A very large number of proteins such as histones, polymerases, and transcription factors interact with chromatin which together with epigenetic modifications give rise to highly complex regulatory programs that ultimately determine the biological phenotype. It is only recently that we have begun to probe this immense network in an attempt to untangle the mechanisms and understand functional outcomes. Chromatin immuno-precipitation has emerged as the method of choice to map these complex interactions on a genome-wide scale but as with most high throughput sequencing approaches, results can be obscured by noise, fragment sampling biases and chromatin accessibility. To separate technical noise from biological variability, we have developed Arpeggio which utilizes Fourier transformation to deconvolve the local distribution of true ChIP signals from systematic noise and bias due to variations in chromatin accessibility that are also present in total DNA-input and IgG samples, commonly used controls for ChIP-Seq experiments. We show that Arpeggio local IP signals can be used for accurate classification and characterization of ChIP-Seq binding profiles for a dataset of 806 experiments and also demonstrate its use for peak detection.

Predicting Aspects of Protein Structure with Deep Networks

Authors:
Jesse L Eickholt, Matt C Spencer, Taeho Jo, Jianlin Cheng, University of Missouri, Columbia

Abstract:
Deep networks and deep learning allow complex models to be trained on large datasets in a reasonable amount of time. Due to a semi-supervised learning process, deep networks first attempt to learn to model the structure of the data itself and then map labels onto a transformed representation of the data. In a sense, this process allows the deep network to learn the features to be used for classification. The training process also lends itself to network architectures which are very deep and very wide and this enhances the modeling power of the network.
Here, we apply and analyze the use of deep networks to a number of facets of protein structure prediction. In particular, we use deep networks to predict secondary structure, solvent accessibility and a protein's fold from sequence. This type of predicted information is often quite useful in modeling a protein's tertiary structure and in assessing the quality of generated models.

NLM Informatics Training Conference 2013
University Buildings



1 **School of Medicine Health Sciences Education Building (HSEB)**

2 **College of Pharmacy Skaggs Hall**

3 **Heritage Commons**

4 **University Guest House**

5 **Tuesday night shuttle Pick-up/Drop-off for trainees**

Transportation and Area Information

University of Utah NLM Trainees will be your guides on Tuesday and Wednesday. They will all be wearing red polo shirts with the UtahBMI logo. Look for them, as they are available to walk with you to the various locations and answer any questions.

All sessions will be held in the College of Pharmacy's Skaggs Hall and the Health Sciences Education Building (HSEB), with the exception of the Tuesday night dinner at the Utah Museum of Natural History.

Trainees:
The Residence Dorms are located directly to the east of Heritage Commons. Travel distance from Heritage Commons to Skaggs Hall and the Health Sciences Education Building HSEB) is 0.2 miles and takes approximately 5 minutes to walk. From the front of Heritage Commons immediately walk north on South Connor Street. Connor Street changes names to South 2000 East. Continue walking north, cross South Medical Drive, the College of Pharmacy Skaggs Hall is on the left, HSEB is on the right.

Attendees at the Guest House:
Travel distance from the University Guest House to Skaggs Hall and the HSEB is 0.3 miles and takes approximately 6 minutes to walk. From the front of the Guest House cross Fort Douglas Blvd and walk south to Vollum. Follow Vollum until it dead ends into South 2000 East and turn left. Continue walking on South 2000 East, cross South Medical Drive, the College of Pharmacy Skaggs Hall is on the left.

NLM Staff at the University Park Marriott:
On Tuesday and Wednesday mornings, shuttles operated by the hotel will bring you down to Skaggs Hall. Request shuttle service the night before (this is a very short trip; ask to be delivered to "HSEB" or "The College of Pharmacy." After the close of Tuesday's last Focus Session, Utah staff will be available to drive you back to the hotel. Speak with Linda Galbreath or John Hurdle to indicate when you would like to leave. The staff will also drive you to the Museum from the hotel, where the festivities begin at 6:30, and back again when you wish to return to the hotel (we will be "on call"). If you need to contact us for any reason, please call John Hurdle at 801-884-9071 or Linda Galbreath at 801-369-2210.

All Attendees: Tuesday, June 18, 2013 – Utah Museum of Natural History
Shuttle Buses will pick-up participants from the top of Officer's Circle (a minute walk west from the front of Heritage Commons—see the map) then drive to the Guest House to pick-up participants and transport everyone to the Utah Museum of Natural History. The buses will begin at 6:15 pm and will make continuous loops with the last bus leaving the Natural History Museum at 11:00 pm. NLM staff see the paragraph above for transport details.

2013 NLM INFORMATICS TRAINING CONFERENCE

EVALUATION SURVEY

http://goo.gl/I4JhQ

Evaluation Survey will be active immediately following the conference on June 20, 2013

Agenda at a Glance

Tuesday, June 18, 2013

7:00 – 8:00	Breakfast Students – **Heritage Center** Faculty – **Guest House** NLM - **Marriott**
	Poster Setup – Day 1 Group **Health Sciences Education Bldg. Atrium** (HSEB)
8:00 – 8:45	Welcome **College of Pharmacy, Skaggs Hall**
8:45 – 10:00	Plenary Session #1 **Skaggs Hall**
10:00 – 11:00	Poster Break Day 1 Group ***Health Care and Public Health*** ***Clinical/Translational*** ***Translational and Bioinformatics*** **HSEB Atrium**
11:00 – 11:45	Parallel Focus Sessions A *Focus Session A1* **HSEB 1700** *Focus Session A2* **HSEB 1730**
11:45 – 1:00	LUNCH Executive Session of Training Directors **HSEB 2100**
	Birds of a Feather for students, remaining faculty and NLM staff **Heritage Center**
1:00 – 2:00	Open Mic Session **Skaggs Hall**
2:00 – 3:15	Plenary Session #2 **Skaggs Hall**
3:15 – 4:15	Poster Break (as above) **HSEB Atrium**
4:15 – 5:15	Parallel Focus Sessions B *Focus Session B1* **HSEB 1700** *Focus Session B2* ***HSEB 1730***
6:30 – 10:00	Dinner/Reception **Natural History Museum of Utah**

Wednesday, June 19, 2013

7:00 – 8:15	Breakfast Students – **Heritage Center** Faculty – **Guest House** NLM - **Marriott**
	Poster Setup – Day 2 Group **Health Sciences Education Bldg. Atrium** (HSEB)
8:30 – 9:30	Plenary Session #3 **Skaggs Hall**
9:30 – 10:30	Poster Break Day 2 Group ***Health Care and Public Health*** ***Clinical/Translational*** ***Translational ad Bioinformatics*** **HSEB Atrium**
10:30 – 11:45	University of Utah Showcase: "Visualizing the Future of Biomedicine" Dr. Chris Johnson **Skaggs Hall**
11:45 – 1:00	LUNCH Intro to Career Development Awards And New Investigator R02 Grants **HSEB 3515C**
	Birds of a Feather for students, remaining faculty and NLM staff **Heritage Center**
12:45 – 2:00	LUNCH - Administrators NLM 2013 Biomedical Informatics Training Program Overview **HSEB 4100C**
1:00 – 2:00	Parallel Focus Sessions C *Focus Session C1* **HSEB 1700** *Focus Session C2* **HSEB 1730**
2:00 – 3:00	Poster Break (as above) **HSEB Atrium**
3:00 – 3:30	Closing Session and Awards **Skaggs Hall**