“It from Bits”: Managing Massive Data as a Critical Challenge For Biomedical R&D and Healthcare Delivery

Dr. George Poste
Chief Scientist, Complex Adaptive Systems Initiative and Del E. Webb Chair in Health Innovation
Arizona State University  
george.poste@asu.edu  
www.casi.asu.edu

Keynote Presentation:  
The Burrill Personalized Medicine Meeting  
Burlingame, CA • Oct. 3-4, 2011
Slides available @ http://casi.asu.edu/
understanding how encoded genome information creates complex multiscale biological systems ("It")

and

defining health and disease in terms of patterns of information flow in biological information networks ("bits")
Medical Progress:
From Superstitions to Symptoms to Signatures
Mapping The Molecular Signatures of Disease: The Intellectual Foundation of Rational Diagnosis and Treatment Selection

Genomics

Proteomics

Molecular Pathways and Networks

Network Regulatory Mechanisms

ID of Causal Relationships Between Network Perturbations and Disease

Patient-Specific Signatures of Disease or Predisposition to Disease
The Intellectual Horizons and Aspirations of Modern Biomedicine

Defining The Molecular Taxonomy of Diseases as Altered Patterns of Dysregulated Biological Networks

- molecular mechanisms of disease
- molecular diversity of disease arising in same organ/cell – disease subtyping
- germ line genome/epigenome – pharmacogenetic risk(s)
  - disease predisposition risk(s)

Improved Disease Diagnosis and Rational Therapy

- precision diagnostics based on alterations in molecular targets/pathways/networks
- rational (targeted) Rx selection based on underlying molecular pathology
- risk mitigation

Systems Biology

Multiscale Biology

Personalized Medicine
Challenges in Making Personalized Medicine a Reality: Key Themes

- Silos
- Standards
- Scale
- Systems
- Structures
The Evolution of Biomedical R&D and Clinical Medicine as Data-and Computation-Intensive Disciplines

Advanced Computing and Knowledge Management (ACKM): Building Critical Competencies to Sustain Progress in Biomedical R&D and Healthcare Delivery
Profiling Platforms for Mapping Molecular Networks: The Accelerating ‘Data Deluge’

Low Cost Exome- and/or Whole Genome Sequencing

Transcriptomics

miRNAs

Proteomics

Protein Interaction Networks (PIN)
daunting complexity of defining signals and signatures across massive combinatorial space
- 230 different cell types + body fluids
- pre-and post-translational gene regulation
- SNPs, copy number variants, mutations, rearrangements
- at least 200 PTMs and multiple PTMs in same biological pathway
- protein expression, abundance and interactomes
- localization, trafficking, turnover
- dynamic range (from attomole to millimole)
- physiological homeostasis
- dysregulation and disease pathogenesis
Rapid Growth of Human Genome Sequencing Data

- Evaluation of all combinations of two SNPs for 1 million SNPs represents nearly 500 billion possibilities.
- dbSNP contains over $20 \times 10^6$ validated SNPs.
- Human Gene Mutation Database contains over 76,000 mutations from 2,900 genes.
- COSMIC (The Catalogue of Somatic Mutations in Cancer) contains over 25,000 unique mutations.
- PharmGKB database lists over 40 pharmacogenes and over 3,400 annotated drug-response variants.
The Human Mitochondrial Transcriptome

From: T. R. Mercer et al. (2011) CELL 146, 645
The Epigenome

Modulation of Gene Expression/Regulation by Environmental Factors/Xenobiotics/Rx (The Exposome)

Effect of Maternal Diet/Stress on Germ Line Genome (+ trans three-generational?)

A. Transcriptionally active chromatin

- DNA
- Transcription
- HATs
- DNMTs
- MBPs
- HDACs

B. Transcriptionally inactive chromatin

- DNA
- Transcription
We Are Not Alone: Variation in the Human Microbiome as a Potential Factor in Health and Disease
Data-Intensive Imaging and High Content Analysis of Cellular Architecture and Dynamics

Chromatin Loop Domains

Modeling of Nucleosome Folding

Digital Pathology

Nature (2011) 470, 292
Science (2009) 326, 289
Computational Chemistry and Molecular Modeling
Data-Intensive Imaging Technologies
# Data-Intensive Biomedical R&D and ‘The Data Deluge’

<table>
<thead>
<tr>
<th>Patient Stratification For Clinical Trials</th>
<th>Pharmacogenomics</th>
<th>m.Health</th>
<th>Remote Health Monitoring</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Patient Stratification" /></td>
<td><img src="image2.png" alt="Pharmacogenomics" /></td>
<td><img src="image3.png" alt="m.Health" /></td>
<td><img src="image4.png" alt="Remote Health Monitoring" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Microbial Diagnostics</th>
<th>Biosurveillance and Public Health</th>
<th>High Performance Computing</th>
<th>Health IT and EMRs</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5.png" alt="Microbial Diagnostics" /></td>
<td><img src="image6.png" alt="Biosurveillance and Public Health" /></td>
<td><img src="image7.png" alt="High Performance Computing" /></td>
<td><img src="image8.png" alt="Health IT and EMRs" /></td>
</tr>
</tbody>
</table>
From Linneas to Life Codes: Mapping Biological Diversity and a New Digital Taxonomy

GenBank release 185.0
14 August 2011

- 131 gigabases of data from 142 million entries of non-whole genome sequencing
- 208 gigabases from 65 million WGS
- additional 970,764 records updated
Data: The Fastest Growing Resource on Earth
Exabyte World

- Projected 1800 exabytes of new global data in 2011 (10x more than 2006)
- Routine multi-petabyte data sets emerging in national security and big science
- Large Hadron Collider estimated 15 petabytes/year
- Smart electricity grid: 100 million customers ≡ 50 petabytes/year before compression
- Walmart: 1 MM transactions/hr = 2.5TB
- Boeing Dreamliner jet engines produce 10TB operational status/30mins
- Twitter ecosystem: 8TB data/day vs NYSE 1TB/day
- Meta-data (information about information)
  - Growing as fast as data in big data environments
Managing “Mega-Data” in Biomedicine

- Volume
- Computational scale
- Global networks

Bench to bedside: multiscale heterogeneity
Integration
Building Large Scale Biomedical Collaborative Research Networks and Commitment to Standardization of Experimental Platforms, Models and Data Reporting

Website: IPSOGEN (http://www.ipsogen.com/corporate/)
The Proliferation of Poorly Standardized, Non-Reproducible and Statistically Flawed Research Data


Reliability of ‘new drug target’ claims called into question

Bayer halts nearly two-thirds of its target-validation projects because in-house experimental findings fail to match up with published literature claims, finds a first-of-a-kind analysis on data irreproducibility.

JAMA (2011) 305, 2200

Comparison of Effect Sizes Associated With Biomarkers Reported in Highly Cited Individual Articles and in Subsequent Meta-analyses

John P. A. Ioannidis, MD, DSc
Orestis A. Panagiotou, MD

Many new biomarkers are continuously proposed as potential determinants of disease risk, prognosis, or response to treatment. The plethora of statistically significant associations increases expectations for improvements in risk appraisal. However, many markers get evaluated only in 1 or a few stud-

Context Many biomarkers are proposed in highly cited studies as determinants of disease risk, prognosis, or response to treatment, but few eventually transform clinical practice.

Objective To examine whether the magnitude of the effect sizes of biomarkers proposed in highly cited studies is accurate or overestimated.

Data Sources We searched ISI Web of Science and MEDLINE until December 2010.

Study Selection We included biomarker studies that had a relative risk presented in their abstract. Eligible articles were those that had received more than 400 citations in the ISI Web of Science and that had been published in any of 24 highly cited biomedical journals. We also searched MEDLINE for subsequent meta-analyses on the same associations (same biomarker and same outcome).
poor access to rigorously annotated biospecimens from stringently phenotyped sources

insufficient control of pre-analytical parameters and variable analytical standards

idiosyncratic individual investigator methods

‘small N’ studies lacking statistical power

chaotic data reporting formats and poor dbase interoperability

pressure to publish and poor compliance with funding agency/journal policies on data sharing

failure to work to industry standards
Mapping the Human Variome: Defining the Molecular Taxonomy of Individuality and Correlations With Phenotypic Traits and Disease Processes
When Will Partial- and Whole Genome Sequencing Become ‘Just Another Laboratory Value’ in Patient Care?
The Expansion of Human Genome Sequencing Projects
What is A Complete and Accurate Analysis of Genome Architecture and Regulation?
The Scale and Complexity of Human Genome Sequencing Data

- **Accuracy and Comprehensiveness**
  - need for consensus metrics for these parameters
  - population-based studies
    - pooled samples with low depth coverage (<10x)
  - personal genomes
    - greater accuracy and confidence for base calling for clinical diagnostics and care decisions
    - regulatory oversight of QA/QC and analytics algorithms
  - current technologies
    - 30-40x coverage to ID 92-95% of both alleles
    - 50-100x coverage to ID 99.9% sequence and rare variants
    - final sequence with only 1 error/10^6 bases will still contain 6000 errors
What Is A Healthy Genome?

Early Members of the 3 Gigabyte WGS Club

Consumer Genomics: Hype or Personal Freedom?

Analysis of Probabilistic Risk(s)
Interpretation of the Functional Effects of Variants

- prediction of deleterious missense SNPs
  - significant fraction in non-coding regulatory regions
- new methods needed to evaluate functional effects of synonymous and intronic SNPs, insertions, deletions and CNVs
- ‘phasing’
  - ID on which of the two chromosomes a variant is located
- emerging evidence of widespread RNA and DNA sequence differences (RDD) in human transcriptome
  - RNA sequences do not match DNA
  - some RNA variants at RDD sites translated into proteins
  - unknown role in biology and/or disease
The Perils of Assembly and Annotation: Total Base Pair Discrepancies in Published WGS of Han Chinese (YH) and Yoruban (Y) Individuals Versus Reference Genome

- C. Alkan et al. (2011) Nature Methods 8, 61
- de novo assemblies were 16.2% shorter than reference genome
- 420.2 megabases of common repeats and 99.1% duplicated sequences missing
- over 2,337 coding exons completely missing
- 136,613 bp in Y genome had high sequence identity to EpsteinBarr virus
- Y sequence generated from cell line vs YH from blood DNA
The Pervasive Problem of Poor Standardization and Annotation in Large Scale ‘Global Profiling’ of Genomes and Gene Expression

“The expertise and motivation to sequence genomes to a high quality are disappearing….
…if genome manuscripts’ can now be published without accounting for the 20% that is missing
…what incentive remains to spend the additional effort and cost to sequence genomes well?”

“The balance between quantity and quality of genomes (must be) reestablished.”

C. Alkan et al. (2011) Nature Methods 8, 61

“The excitement of having more and more tools always brings us back to the very important question of having to validate or replicate. I worry that that’s getting lost as everyone gets so excited about the next really cool tool.”

Dr. Stephen Chanock
Chief, Translational Genomics, NCI
Genome Technology April 2011 p. 31
The Cost of Sequencing
Versus
The Cost of Computational Analysis and Storage

- the $1000 genome,
  - the $? analysis and interpretation cost
  - the $? storage, retrieval and security costs
- turn around time (TAT) and analysis for clinical value cost
- regulatory and reimbursement policies
The Data Storage Challenge:
The Price of Sequencing is Falling Faster Than Computer Storage Costs and Availability

- data ‘triage’: store only data deemed relevant and/or with differences to reference set
  - risk of bias/ignorance about value of discarded data elements
- data compression and ‘loss of precision’
  - different compression methods depending on desired end use/reuse needs
- unmapped reads cannot be compressed using current alignment frameworks
  - 10-40% of reads remain unmapped to traditional reference genomes
  - 60-70% for short RNA sequencing reads
- many samples may not be reacquirable/renewable
  - cancer
“Clinical Grade” Genome Sequencing: Ready for Prime Time?

Disparate Views on Timing

“…pie in the sky ideas about what the clinician could do if they had at their disposal a genome and a set of analytical tools.”

….we don’t know how to do any of this.”

….it will take years, if not a decade or so, to implement widely and effectively.”

Dr. Les Biesecker
Chief, National Human Genome Research Institute, Genetic Disease Branch
IOM Symposium 2011

“Next generation sequencing is truly changing the way we’re treating patients.”

Francois Ferre
CEO, Althea Dx
Genome Web 22 March 2011
23 June 2011 workshop

- accuracy, depth of coverage, validation set, impact of pre-analytic/analytic variables
- CLIA/CAP facilities
- sequencers as Class III devices?
- RUO materials
- source computer code(s) for analytical algorithms
- performance thresholds and QA/QC requirements for error detection (instrumentation + analytics)
Integration of Genome Sequencing, Gene Expression and Transcriptomics Data With The Dynamics of Biological Pathways and Networks
Individual Variation, Genomic Complexity and the Challenge of Genotype-Phenotype Prediction

Junk No More!
- pervasive transcription
- SNPs, CNVs
- indels, SVs
- ncRNAs
- phasing
- epistasis
- imprinting

Molecular Interaction Networks

recognition of increasing organizational and regulatory complexity

Disease Perturbations
“An Algebra for Theoretical Genetics”  
Ph.D. Thesis, MIT (1939)

“A Mathematical Theory of Communication”  
Bell System Technical Journal (1945)

“It is obvious from the analysis of these (bacterial genetic regulatory) mechanisms that their known elements could be connected into a wide variety of ‘circuits’ endowed with any desired degree of stability.”

François Jacob and Jacques Monod (1962)  
● the current ‘black box’ of major knowledge voids that create poor predictability in accurate ID and selection of biomarkers/Rx targets with resulting high failure rates in clinical trials

● has the gap between basic science and realizable therapeutic applications widened?

● how can systems complexity be deconvoluted to identify tractable approaches for new molecular diagnostics, targeted therapy out and disease risk predisposition assessment?
Mapping Modules, Pathways and Subnetworks in Biological Systems: The TCGA Glioblastoma Multiform Dataset and Pathway Analysis

From: C. J. Vaske et al. (2011)
Bioinformatics 26, i237

From: J. H. Morris et al. (2010)
Molec. Cell. Proteomics 9, 1703
Molecular Interaction and Signaling Pathway Resources

- Biocarta
- KEGG
- Reactome
- BioGRID
- MetaCyc
- PANTHER
- PathScan® ELISA
- Pfam
- PATIKA
- IMID
- PID
- ChiBe
- Human Protein Reference Database
- Pathway Interaction Database
- PathCase
- WikiPathways
- NetPath
- CST
- VisANT
Mapping Pharmacological “Interaction Space” in Biological Pathways and Subnetworks

DrugBank

Therapeutic Targets Database

PUBLISHERS OF WOMBAT

PDT [Potential Drug Target Database]

CTSA Pharmaceutical Assets Portal

chem2bio2rdf
Reducing The Failure Rate of Investigational Drugs in Clinical Trials

- targeted therapies, YES!

- improved success requires targeting network modules, pathways and subnetworks **not single targets**

- complexity of linked and overlapping modules and pathway “cross-talk”
  - long range pleiotropic effects
  - weak indirect effects
Mapping of Protein Interaction Network in Alzheimer’s Disease (AD)

From: M. Soler-Lopez et al. (2011) Genome Res. 21, 364

- 200 high confidence 2P interactions
  - 8 confirmed AD – related genes
  - 66 additional candidates
  - 31 in chromosome regions containing putative susceptibility loci
  - 17 dysregulated in AD

Place Your Bets!
Network Pharmacology

- same drug: interaction with multiple targets
- same target: interaction with multiple drugs
- mapping structural chemotypes to pathways and subnetworks for targeted (poly)pharmacology

From: M. J. Keiser et al. (2011) Nature 462, 180
Network Based Perturbations in Disease: Implications for Biomarker Discovery and Validation

- ‘publish and vanish’: poor productivity due to failure to evaluate network effects
- if disease involves multi-loci perturbations in modules/subnetworks then multiple parameters will need to be measured
  - different multiplex biomarkers in different disease subtypes
  - changes in biomarker profile with disease progression/Rx response/Rx resistance

G. Poste (2011) Nature 469, 156

Bring on the biomarkers

The dismal patchwork of fragmented research on disease-associated biomarkers should be replaced by a coordinated ‘big science’ approach, argues George Poste.
"It takes a network to stop a network."

Lt. Gen. S. McChrystal
(on combating the Al-Qaeda and Taliban in Afghanistan)
Understanding the Internal Circuit Diagrams of Cells and Identification of the Disruption(s) Caused by Disease

Disease Profiling to Identify Subtypes (+ or - Rx Target)

ID Molecular Targets for Rx Action and Blockade of Compensatory “By pass” Pathways
Initial Response (A/B) of BRAF-V600 Positive Metastatic Miliary Melanoma After 15 Weeks Therapy with Vemurafenib (Zelboraf® - Roche) Followed by Rapid Recurrence of Rx-Resistant Lesions with MEKI C1215 Mutant Allele After 23 Weeks Therapy

From: N. Wagle et al. (2011)
J. Clin. Oncol. 29, 3085
Network Pharmacology

- Elucidation of definitive ‘chokepoints’ as optimum targets
  - Subvert adaptive cellular options to use alternate compensatory pathways

- The design challenge for multi-target polypharmacology
  - Multi-agent therapy (patient tolerance?)
  - Controlled multi-target promiscuity in a single moiety

- Does chronic progression in complex, multigenic diseases amplify pathway/subnetwork dysregulation?
  - Greater complexity of multi-target Rx for homeostatic ‘reset’
  - Role of Rx in driving selection of variants with Rx-resistance (‘escape’) pathways (e.g. oncology)
Improving the Productivity and Proficiency of Biomedical R&D and Clinical Medicine: An Unavoidable (But Essential) Transition to Data-and Computation-Intensive Methods
An Unavoidable Transition to Data-and Computation-Intensive Methods

Current Era

- a high opinion, low robust information content world
- “silos” of research/clinical activities and slow adoption of “systems-based” cross-disciplinary integration
- proliferation of poorly standardized and fragmented data, semantic anarchy and incompatible databases
- poor predictability of the behavior of complex biological networks and accompanying ‘rude shocks’
  - clinical trial failures (biomarkers, Rx)
  - inaccurate diagnosis and flawed clinical decisions
  - highly variable treatment selection and uncertain clinical outcomes
  - extravagant waste and risk to patients
An Unavoidable Transition to Data-and Computation-Intensive Methods

Current Era

The Imperative to Move from Descriptive Phenomenology to Mechanism-Based Knowledge of the Behavior of Complex Biological Networks to Achieve Precision Diagnosis, Rational Targeted Rx Design and Improved Clinical Outcomes
An Unavoidable Transition to Data-and Computation-Intensive Methods

- “massive data”: automated, high throughput, massively parallel profiling tools to study body function
- technology acceleration and convergence (life sciences, engineering, telecommunications, computing)
- adoption of stringent analytical QC/QA and consistent ontologies for data reporting
- new annotation tools for ‘big data’ and interoperable database designs for facile cross-disciplinary/cross-sector integration
- new computational architectures and services for mega-data and machine-based mining and analysis
The Imperative for Integrated Inter-Operable ACKM Capabilities Across the Full Continuum from Discovery to Patient Care

- Discovery
- Translation
- Clinical trials
- Clinical validation
- Optimized healthcare delivery
- EHRs
- Consumers/patients
- mHealth
- Regulatory
- Reimbursement
- Outcomes analysis
- CER
- Remote health status monitoring
IT-enabled Ecosystems for Discovery and Translational Research

Adapted from: K. Buetow, NCI 2011
http://sagecongress.org/Presentations/Buetow.pdf
Data Integration and Exchange Standards

- Chaotic state of discovery stage semantics, standards
- Limited research dbase inter-operability with industry/regulatory standards for clinical trials
- Leveraging existing HL7 standards for clinical trials
  - Draft Standards for Trial Use (DSTU)
  - CDISC, ICH
- Digital Imaging and Communications in Medicine (DICOM)
- Seamless federation with healthcare system and reimbursement databases
  - CPT, ICD (USA)
- Certification of compliance with proposed HITECH EHR Standards (HIMSS, AHIMA)
“In one 24 hour period, 465 children were admitted to the hospital with fever. Their fever-like symptoms were recorded in the EMR in 278 ways.”

Dr. C.B. Forrest, Professor of Pediatrics, Children’s Hospital of Philadelphia cited in: Registries-RemedyMD (2011)
A Learning Healthcare System

Proliferation of Clinical Computational Systems

Clinical Decision Support Systems:
State of the Art

HITECH Mandates
Incentives
EHR and Smart Cards
Informed Consumers/Patients

Overview:
Federal Health IT Strategic Plan
2011-2015

Go Paperless and Get Paid
Register NOW for CMS Electronic Health Record Incentives

Informed Patients
Just Ahead
The Only Valuable Data is Validated, Actionable Data
Mining EHRs to Identify Disease Correlations with Molecular Profiling Datasets and Improved Clinical Stratification (Phenotyping) of Patient Cohorts

- 18.688 million medical members
- 13.953 million dental members
- 10.410 million pharmacy members
- 966,000 healthcare professionals
- 543,000 primary care doctor specialists
- 5,200 hospitals
- 71 billion health records
- 75 TB storage (50% occupied)

From: Health Data Sept. 2011
Mapping of 26 Clusters in a Phenotype Networks in 1497 Danish Psychiatric Patients By Combing Structured (ICD Codes) and NLP Processing of Medical Records

What Is?
The Evolution of Computation Capabilities for Natural Language Q&A in Large Datasets

- IBM’s Watson
  - 2880 CPUs
  - natural language questions

- Prototype for intelligent systems for biomedicine beyond keyword IR searches

Jeopardy 16 February 2011
The Status Quo is Neither Desirable Nor Sustainable

• continued undisciplined, free form tagging and idiosyncratic researcher/clinician-centric approaches
  - fragmented data silos and poor database interoperability/integration

• adoption of highly engineered ontologies and formal logic
  - reference structures, vocabularies and semantics
  - automated machine-based algorithmic or rules-based scoring
Informatics and High Performance Computing

- modeling and simulation of biological networks of escalating complexity

- development of new mathematical, statistical and computing tools for analysis and modeling of non-linear phenomena in complex networks

- application of advanced machine learning tools, avatars, robots and automated data production suites for customized data to promote optimum decision-actions
New Visualization Tools, Interactive Interfaces and Rapid Customization Formats
Technology Acceleration and Convergence: The Escalating Challenge for Professional Competency, Decision-Support and Future Education Curricula
The Imminent Collapse of the Genome Informatics Ecosystem?

- **Moore’s Law**
  - # transistors/circuit board doubles c.18 months
- **Kryder’s Law**
  - hard disk capacity doubles c.12 months
- **Butter’s Law**
  - cost of sending a bit of information over optical network halves every 9 months
- **NGS**
  - sequence data doubles every 6 months (other ’omics will follow)
Managing Massive Data: Scale, Storage and Cost

- the ‘lottabyte’ problem
- widespread strategic failure to assess scale and cost requirements for high performance computing (HPC) needs
- intelligent total-cost-of-ownership analysis
  - cost per gigabyte and user allocated charges
  - platform diversity (supercomputers to smartphones)
  - physical versus virtual storage
  - impact of new technologies (solid-state drives, deduplication)
  - data retention policies
  - cloud services versus internal operations
Managing Big Data in Biomedicine:
Learning Precedents from Other Research Domains and Corporate Capabilities
To The Cloud:

GSA

TechAmerica Foundation
Biocomputing in the Cloud

- shared ‘virtual commons’ for elastic (burst), scalable scientific computing
  - CaaS, SaaS, IaaS
- internet not designed for transfer of large scale datasets (terabyte +)
  - typical research lab. connectivity 1-10 gigabit/sec (0.25 to 1.25 gigabytes/sec) requires 1 week/1 day to transfer 100 gigabyte NGS file
  - more cost-effective to FedEx hard drive to CC provider
  - requirement to constantly move same datasets (and refresh them) and mirror them in multiple local storage systems
BGI Cloud on the Horizon

- “Amazon is slow”
  Evan Xiang, BGI Shenzhen
  Bio-IT World August 2011 p.8

- launch of new platforms
  - Hecate: de novo assembly
  - Gaea: SOAP, BWA, Samtools, Dindel, reals-FS algorithms

- November 2011 launch of new journal with BioMed Central
  - ‘big data’ studies
  - host citable public datasets on BGI cloud
  - each with permanent digital object identifiers
Personal Privacy In An Era of Pervasive Computing and Multi-Parameter Profiling
Homomorphic Encryption: Creating a Secure Cloud?

- Commercial objective to build secure virtual vault for medical data
- Mathematical proof of feasibility by IBM in 2009 (Craig Gentry)
- Cloud-based tools for statistical analysis of encrypted data without decryption
- Results remain fully encrypted for interpretation only by data owner’s key(s)
- Any escape/leak/hack would only be in encrypted form
- Next phase to increase speed and use in search and indexing functions
Development of Vanguard Capabilities in ACKM: A Fundamental Requirement for Sustained Competitiveness

REPORT TO THE PRESIDENT
AND CONGRESS

DESIGNING A DIGITAL FUTURE:
FEDERALLY FUNDED RESEARCH
AND DEVELOPMENT IN
NETWORKING AND INFORMATION
TECHNOLOGY

Executive Office of the President
President’s Council of Advisors on
Science and Technology

DECEMBER 2010
“If the scientific community can justify billions of dollars, 100MW of power and thousands of staff in order to fire tiny particles that most people have never heard of around a big ring of magnets for a fairly narrow science purpose that most people will never understand…..

…..then how come we can’t make the case for facilities needing half the resources that can do wonders for a whole range of science problems and industrial applications?”

Andrew Jones
Vice-President, Numerical Algorithms Group
HPC Wire 29 August 2011
Major New Initiatives in Supercomputing for Analysis of Research and Clinical Molecular Profiling Data

Chan Soon-Shiong
Institute for Advanced Health (CSS Institute)
and
National Coalition for Health Integration
The Principal Forces Shaping Progress in Biomedical R&D and Clinical Practice in An Era of Data-Intensive Methods

- new complexities
- new competencies
- new computation-intensive requirements
- new consumer and communication services
- new competitors
- new cross-sector alliances and business models
- new cultures and organizational structures
The Sociology of Integration of Computational Science as a Core Component of Biomedical R&D

- bridging three cultures
  - biomedical specialities, software engineering and scientific computing
- new ‘hybrid’ competencies/specialities
- building sufficient expertise (individuals/communities)
  - training, funding, incentives, rewards
- designing workflows and interfaces for e.science and federated virtual research environments
- increasing dependence/contribution on open-source datasets
- mapping data provenance in large multi-source datasets
- life in ‘the perpetual beta’
Technological Innovation in Different Domains: A Study in Cultural, Methodological and Organizational Contrasts

Physics

Chemistry

Engineering

Computing

Telecommunications and Digital Media
**Technological Innovation in Different Domains:**
*A Study in Cultural, Methodological and Organizational Contrasts*

<table>
<thead>
<tr>
<th>Physics</th>
<th>Chemistry</th>
<th>Engineering</th>
<th>Computing</th>
<th>Telecommunications and Digital Media</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Physics Image" /></td>
<td><img src="image2" alt="Chemistry Image" /></td>
<td><img src="image3" alt="Engineering Image" /></td>
<td><img src="image4" alt="Computing Image" /></td>
<td><img src="image5" alt="Telecommunications Image" /></td>
</tr>
</tbody>
</table>

- Increasingly mature knowledge of underlying design laws and principles
- Mechanism-based, standardized methods and highly predictable outcomes
- Problem-oriented and focus on robust solutions
- Organizational/institutional structures address systems scale and complexity
Biomedical R&D and Clinical Medicine Are Methodological and Cultural Outliers in the Science and Technology Universe

Performance Parameters to Archive High Levels of Predictable Product Performance

- mature knowledge of underlying design laws and principles
- mechanism-based, standardized methods and predictable outcomes
- problem-oriented and focus on solutions using integrated cross-disciplinary teams
- organizational/institutional frameworks address systems scale and complexity

Current Predominant Culture/Methods in Biomedicine

- largely phenomenological experiments/clinical interventions (‘black box’)
- largely descriptive data, poor standardization, replication and low predictive power
- hypothesis-based, organizational silos and inefficient transfer processes
- ‘cottage industry’ era and slow adoption of systems-based approaches and requisite scale
An Unavoidable (But Essential) Transition to Data-and Computation-Intensive Methods

**Strategic Needs**

- systems-based approaches to define physiology and pathology in terms of molecular information networks
  - multiscale and hierarchical (cells to person)
  - spatiotemporal breadth (psec to lifespans)
- comprehensive knowledge of the topologies, dynamics and (dys)regulation of molecular networks to increase the predictability and productivity of all aspects of biomedicine
  - reduced failure rate of Dx/Rx assets in clinical trials
  - improved clinical decisions and health outcomes
  - risk mitigation and sustainable health
Changing the Sociology and Organization of Biomedical R&D

- ‘flying blind’: current ‘voids’ in understanding biological network dynamics and disease-associated perturbations
  - limited prediction of system behavior
  - unacceptable high failure rates in clinical trials

- pre-competitive 3P consortia to define ‘rules’ for biological network behavior
  - ‘connectome dynamics’
  - ID optimum loci for Rx
  - multiplex biomarkers for Dx
  - patient stratification
    - enrichment/ adaptive trials
    - rational Rx selection and improved outcomes
    - new analytical services and data aggregators

- ‘voids’ in understanding biological network dynamics and disease-associated perturbations
  - limited prediction of system behavior
  - unacceptable high failure rates in clinical trials
IT Future of Medicine

Barcelona Supercomputing Center
European Sequencing and Genotyping Infrastructure

ESGI

Infrastructure for Systems Biology – Europe

ISBE

EATRIS

PASCAL²

ELIXIR

European Life Sciences Infrastructure for Biological Information

INSTRUCT

Integrated Structural Biology Infrastructure

BBMRI

Biobanking and Biomolecular Resources Research Infrastructure

PRACE

Partnership for Advanced Computing in Europe

IBM

Microsoft

SICEENS

INSTITUTE

Jüllich Supercomputing Centre

xerox

amazon

ORACLE

IBM

Jülich Forschungszentrum online

ITFOM

European Advanced Translational Research Infrastructure for Medicine
The Rise of Data-Intensive Biomedicine: Disruptive Change and New Value Drivers for Improved R&D Productivity and Healthcare Quality

- biomedical R&D and healthcare delivery are under siege
  - conceptually limited, operationally inefficient, wasteful and economically unsustainable
- radical (disruptive) changes are required across the entire continuum from discovery to product development to patient care
- mapping disease-induced perturbations in biological systems and networks as a unifying theme for improved R&D strategies, rational clinical decisions and improved health outcomes
- realization of this objective will require new technical capabilities and organizational models to assemble and analyze biomedical data on an unprecedented scale
new capabilities for engaging escalating complexity
- new public funding priorities for large scale, team-based research
- new education and training curricula
- new 3P consortia to solve current major knowledge ‘voids’ in network dynamics that result in high product failure rates in clinical development, poor clinical decisions and less than optimal outcomes

new opportunities to exploit technology convergence and data-centric platforms to build new business models and alliance networks
Slides available @ http://casi.asu.edu/