Retrieving Information from the Book of Humanity: the Personalized Medicine Data Tsunami crashes on the beach of Jeopardy

Daniel R. Masys, M.D.
Affiliate Professor
Biomedical and Health Informatics
University of Washington, Seattle
Topics

- Biology as literature: an information retrieval challenge
- The personal genome data tsunami forecast
- Incorporation of personal molecular variation into Electronic Health Records: Promise and Peril, Joy and Jeopardy
Science of Genetics

Watson & Crick, 1953

A Structure for Deoxyribose Nucleic Acid

We wish to suggest a structure for the salt of deoxyribose nucleic acid (DNA). This structure has model features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey. They kindly made their model available to us in advance of publication. Their model consists of three internested chains, with the phosphates near the fiber axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons. (1) We believe that the material which gives the X-ray pattern is not the basic acid. Without the bases, hydrogen bonds are not close enough to form the complex; especially on the sugar-phosphate fragment near the neck, where the bases are absent. In addition, the outer Watson-Crick distances appear to be too small.

(2) Another important feature is that the orientation of the bases is too regular, held together by hydrogen bonds.

We wish to put forward a completely different structure for the salt of deoxyribose nucleic acid. This structure, with the bases held planar, is not consistent with the X-ray data. We have made the usual model complexes, namely, the complex results, the complex of DNA, and the complex of DNA with RNA. We have found that the bases are held in position by hydrogen bonds instead of the usual phosphodiester bonds. In this model, the bases are arranged in a plane, and the phosphates are arranged in a helical manner. The model is consistent with the X-ray data, and can be built up from the experimental results.

The structure we have suggested is a new one for DNA, and is not consistent with the model proposed by Pauling and Corey.
The Central Dogma of Molecular Biology

- **Replication**: DNA duplicates
- **Transcription**: RNA synthesis
- **Translation**: Protein synthesis
Partial list of protein translation ‘rules’:

<table>
<thead>
<tr>
<th>Alphabet</th>
<th>mRNA Nitrogen Bases (Codons)</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>GCA, GCC, GCG, GCU</td>
<td>ALANINE</td>
</tr>
<tr>
<td>B</td>
<td>AGA, AGG, CGA, CGC, CGG, CGU</td>
<td>ARGININE</td>
</tr>
<tr>
<td>C</td>
<td>AAC, AAU</td>
<td>ASPARAGINE</td>
</tr>
<tr>
<td>D</td>
<td>GAC, GAU</td>
<td>ASPARTIC ACID</td>
</tr>
<tr>
<td>E</td>
<td>UGC, UGU</td>
<td>CYSTEINE</td>
</tr>
<tr>
<td>F</td>
<td>GAA, GAG</td>
<td>GLUTAMIC ACID</td>
</tr>
<tr>
<td>I</td>
<td>CAC, CAU</td>
<td>HISTIDINE</td>
</tr>
<tr>
<td>Z</td>
<td>UCU</td>
<td>SERINE</td>
</tr>
<tr>
<td>END WORD</td>
<td>UAA, UAG, UGA</td>
<td>STOP</td>
</tr>
</tbody>
</table>
An approximate simile for the problems of information retrieval from nucleotide sequence
Zyxwuuuuuuuuuuuaxxmsdddsdas callmeishma elseoyearsagonevermindhowaaaaaaaaaaxxlong preciselyacfsdgdsdxhavingleornomoneyinmy purseandnothingparticulartoalksdhhgxinterestmeonshoreacfsdgsdxxithoughtiwouldsail aboutalittleandseethewaterypartoftheworld aaxxmsdddsdas

Introns: noncoding DNA

Flanking sequences: noncoding DNA
Call me Ishmael. Some years ago--never mind how long precisely--having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world.

Herman Melville

Moby Dick
Mutations

The Cat in the Hat

point mutation

The Rat in the Hat

Red Blood Cells

Sickle cell

Normal red blood cell

Hand with deformed fingers
Insertion Mutations

The quick brown fox jumped over the fence to greet the lazy dog

Single letter insertion causes Frame Shift mutation

Thb equic kbrow nfo xjumpe dove rth efenc et ogree tth elaz ydog
## Language Assembly Heirarchies

<table>
<thead>
<tr>
<th>Human Language</th>
<th>Genetic Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>Triplet codons</td>
</tr>
<tr>
<td>Multiword terms</td>
<td>Oligonucleotides, oligopeptides</td>
</tr>
<tr>
<td>Phrases</td>
<td>Individual genes</td>
</tr>
<tr>
<td>Sentences</td>
<td>Pathways of coordinated genes</td>
</tr>
<tr>
<td>Paragraphs</td>
<td>Tissues</td>
</tr>
<tr>
<td>Chapters</td>
<td>Organ Systems</td>
</tr>
<tr>
<td>Book</td>
<td>Intact organisms</td>
</tr>
</tbody>
</table>
The Human Genome Project
1986- >2005
Harbinger of the Data Tsunami
The Genome Sequence is at hand...so?

“The good news is that we have the human genome. The bad news is it’s just a parts list”
The Promise (joy)

- Molecular and clinical biomarkers for health conditions individuals either have or are susceptible to
  - Includes traditional healthcare history, physical findings, diagnostic imaging, standard clinical laboratories
  - Increasingly: large volumes of molecular data
    - Structural genomics: DNA in residence (~22,000 genes)
    - Functional genomics: genes switched on (1-2% active)
    - Proteomics (400,000 proteins from 22,000 genes)
The Promise (joy), cont’d

- Precision Health Care
- Pharmacogenomics
  - “The right dose of the right drug for the right patient at the right time”
  - Drug development:
    - Avoid drugs likely to cause side effects
    - Re-investigate “back-burner” drugs
    - Develop entirely new drugs targeting fundamental disease processes

“Here’s my sequence…”

New Yorker, 2000
Tsunami Forecast: Big Data Ahead in Healthcare

NEW YORK – A biotechnology company has announced it can decode a person's DNA in a day for $100. The resulting person's genome useful for medical care.
Tsunami problem #1: getting the data into Electronic Health Records in optimally usable format(s)

Technical desiderata for the integration of genomic data into Electronic Health Records


Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, WA 98195-7240, United States
Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195-7220, United States
Department of Genome Sciences, University of Washington, Seattle, WA 98195-7220, United States
Division of Prevention and Population Sciences, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, United States
Advanced Technologies and Surgery Branch, Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, United States
Cerner Corporation, Kansas City, MO, United States
Harvard-MIT Division of Health Sciences and Technology, Biomedical Informatics, Cambridge, MA, United States
Division of General Internal Medicine and McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, United States

Abstract

Output of workshop on “Integration of Genetic Test Results into Electronic Medical Records” convened by the National Heart Lung and Blood Institute, Bethesda, MD August 2-3, 2011
Most common current method for delivery of DNA analysis into clinical operations.

The KRAS gene (12p12) is a member of the Ras family of proto-oncogenes, and encodes a protein containing guanosine nucleotide triphosphate hydrolysis activity (known more commonly as a GTPase). These proteins are active when bound to guanosine triphosphate (GTP) and inactive when bound to guanosine diphosphate (GDP). KRAS is membrane-bound, is activated by growth factor receptors, and through BRAF, stimulates the MAPK/ERK pathway resulting in transcription and cell proliferation. KRAS mutations are observed in colon cancer (40-50%), lung cancer (20-30%) and pancreatic cancers (90%). Conserved missense mutations in codons 12 and 13 result in prolonged binding of GTP and constitutive activation of RAS proteins, thereby leading to uncontrolled cell proliferation.

Progressive and/or metastatic non-small cell lung adenocarcinomas are often treated with inhibitors of the EGFR receptor as a second line therapy. However, it has been shown that tumors, which harbor mutations in codons 12 and 13 of KRAS...
Issues with treating genomic analysis in same fashion as other professionally interpreted clinical data

- Lossy compression: many DNA features observed, only a few clinically relevant reported, remainder discarded
- Interpretation inextricably bound together with primary observations in a document format
- Document reporting format not easily amenable to parsing for automated machine interpretation, decision support, and finely granular information retrieval
- Much more unknown than known about genomic effects, and science changing rapidly
7 desiderata for molecular variation data in EHRs

1. Lossless data compression from (high volume) primary observations to clinically relevant subsets.
2. Since methods will change, molecular lab results carry observation methods with them.
4. Simultaneously support for information retrieval of human-viewable formats (with links to interpretation) and formats interpretable by decision support rules.
5. Separate primary sequence data (remains true if accurate) from clinical interpretations of them (will change with rapidly changing science).
6. Anticipate the boundless creativity of Nature: multiple somatic genomes, multiple germline genomes for each individual over their lifetime.
7. Support both individual care and discovery science.
Step 1: A principled approach to reducing the data tsunami to manageably-sized components

Layered classes of EHR-relevant data

- Primary Observations. If accurate, keep forever
- Consensus full personal germline and somatic sequence(s) and metadata: a few gigabytes each
- 8x -30x nextgen reads: hundreds of gigabytes/terabyte
- Personal molecular differences represented in EHR as computed offset from a Clinical Standard Reference Genome (CSRG) =~1% of genome/proteome. A few megabytes.
- Diagnostic Interpretations (PDF reports). A few kilobytes each.
- Structured keywords for clinical decision support (e.g., C*2*2CLM fires decision rule for CYP2C19*2 homozygotes at time of clopidigrel prescribing). A few tens of bytes each.
- Interpretive codes
Step 2: Retrieving data from EHRs for advancing correlation science: Genomes and Phenomes
The eMERGE Network
electronic Medical Records & Genomics
A consortium of biorepositories linked to electronic medical records data for conducting genomic studies

URL: www.gwas.org
General elements of a clinical phenotype description derived from EHR data

- Diagnostic and procedure codes (ICD9, CPT)
- Laboratory values
- Medications
- NLP: Natural Language Processing of clinical documents (H&P, Operative notes, Discharge Summaries, Progress Notes).
General algorithm for determining EHR-derived phenotype

Definite Cases (algorithm-defined)  Possible Cases (require manual review)  Excluded (algorithm-defined)  Controls (algorithm-defined)

- Iteratively refine case definition through partial manual review until case definition yields PPV \( \geq 95\% \)
- For small case sizes (~100), hand curate cases but use automated case definitions for others
- For samples with inadequate counts of “Definite Cases”, manually review possible cases to determine true positives
- For controls, exclude all potentially overlapping syndromes and possible matches, iteratively refine such that NPV \( \geq 98\% \)
Example of EHR phenotype selection logic
(Rheumatoid Arthritis)

ICD 9 codes (any of the below)
- 714  Rheumatoid arthritis and other inflammatory polyarthropathies
- 714.0  Rheumatoid arthritis
- 714.1  Felty’s syndrome
- 714.2  Other rheumatoid arthritis with visceral or systemic involvement

AND
Medications (any of the below)
- methotrexate [MTX][amethopterin] sulfasalazine [azulfidine]; Minocycline [minocin][solodyn]; hydroxychloroquine [Plaquinil]; adalimumab [Humira]; etanercept [Enbrel] infliximab [Remicade]; Gold [myochrysine]; azathioprine [Imuran]; rituximab [Rituxan] [MabThera]; anakinra [Kineret]; abatacept [Orencia]; leflunomide [Arava]

AND
Keywords (any of the below, asserted rather than negated)
- rheumatoid [rheum] [reumatoid] arthritis [arthritides] [arthriris] [arthristis] [arthritus] [arthrtis] [artritis]
RA case definition – 2: exclusions

AND NOT

ICD 9 codes (any of the below)

- 714.30 Polyarticular juvenile rheumatoid arthritis, chronic or unspecified
- 714.31 Polyarticular juvenile rheumatoid arthritis, acute
- 714.32 Pauciarticular juvenile rheumatoid arthritis
- 714.33 Monoarticular juvenile rheumatoid arthritis
- 695.4 Lupus erythematosus
- 710.0 Systemic lupus erythematosus
- 373.34 Discoid lupus erythematosus of eyelid
- 710.2 Sjogren's disease
- 710.3 Dermatomyositis
- 710.4 Polymyositis
- 555 Regional enteritis
- 555.0 Regional enteritis of small intestine
- 555.1 Regional enteritis of large intestine
- 555.2 Regional enteritis of small/large intestine
- 555.9 Regional enteritis of unspecified site
- 564.1 Irritable Bowel Syndrome
- 135 Sarcoidosis
- 696 Psoriasis and similar disorders
- 696.0 Psoriatic arthropathy
- 696.1 Other psoriasis and similar disorders excluding psoriatic arthropathy
- 696.8 Other psoriasis and similar disorders
- 099.3 Reiter's disease
- 716.8 Arthropathy, unspecified
- 274.0 Gouty arthropathy
- 358.0 myasthenia gravis
- 358.00 myasthenia gravis without acute exacerbation
- 358.01 myasthenia gravis with acute exacerbation
- 775.2 neonatal myasthenia gravis
- 719.3 Palindromic rheumatism
- 719.30 Palindromic rheumatism, site unspecified
- 719.31 Palindromic rheumatism involving shoulder region
- 719.32 Palindromic rheumatism involving upper arm
- 719.33 Palindromic rheumatism involving forearm
- 719.34 Palindromic rheumatism involving hand
- 719.35 Palindromic rheumatism involving pelvic region and thigh
- 719.36 Palindromic rheumatism involving lower leg
- 719.37 Palindromic rheumatism involving ankle and foot
- 719.38 Palindromic rheumatism involving other specified sites
- 719.39 Palindromic rheumatism involving multiple sites
- 720 Ankylosing spondylitis and other inflammatory spondylopathies
- 720.0 Ankylosing spondylitis
- 720.8 Other inflammatory spondylopathies
- 720.81 Inflammatory spondylopathies in diseases classified elsewhere
- 720.89 Other inflammatory spondylopathies
- 720.9 Unspecified inflammatory spondylopathy
- 721.2 Thoracic spondylitis without myelopathy
- 721.3 Lumbosacral spondylitis without myelopathy
- 729.0 Rheumatism, unspecified and fibrositis
- 340 Multiple sclerosis
- 341.9 Demyelinating disease of the central nervous system unspecified
- 323.9 transverse myelitis
- 710.1 Systemic sclerosis
- 245.2 Hashimoto's thyroiditis
- 242.0 Toxic diffuse goiter
- 443.0 Raynaud's syndrome

AND NOT Keywords (any of the below, asserted)

juvenile [juv] rheumatoid [rheum] [reumatoid] [rhumatoid] arthritis [arthritides] [arthrhis] [arthristis] [arthritis] [arthritis] [arthritis]
juvenile [juv] arthritis arthritis [arthritides] [arthritis] [arthristis] [arthritis] [arthritis] [arthritis]
juvenile chronic arthritis [arthritides] [arthritis] [arthristis] [arthritis] [arthritis]
juvenile [juv] RA; JRA
Inflammatory [inflamatory] [inflam] osteoarthritis osteoarthrosis [OA]
Reactive [psoriatic] arthritis [arthropathy] [arthritides] [arthritis] [arthristis] [arthritis] [arthritis] [arthritis]
Information Retrieval Challenges for Phenotype extraction from EHRs

- No ‘all purpose’ phenotype extraction algorithm; must be tuned by experts for eccentricities of care, coding and documentation

- Some tasks are require more than NLP:
  - Two records disagree
  - Evolution of diagnoses over time
  - Phenotypes with definitions other than just “yes”/“no” (e.g., smoking status)
  - Severity of disease
Step 3: Retrieving and Using the Data for Healthcare Decisions

Jeopardy: the Perils Begin
Peril #1: 
Systems Design Issues in Healthcare

- Current practice depends upon the clinical decision making capacity and reliability of autonomous individual practitioners, for classes of problems that routinely exceed the bounds of unaided human cognition.

The molecular tsunami crashes on the beach of human cognitive capacity for decision making...

- Decisions by clinical phenotype
  - i.e., traditional health care

- Human Cognitive Capacity

- Proteomics and other effector molecules
- Functional Genetics:
  - Gene expression profiles
- Structural Genetics:
  - e.g. SNPs, haplotypes

Structural Genetics:
- e.g. SNPs, haplotypes

Functional Genetics:
- Gene expression profiles
General observations about clinical genomics

- Genomic data is the current poster child for complexity in healthcare
- No practitioner can absorb and remember more than a tiny fraction of the knowledge base of human variation
- Therefore, computerized clinical decision support is the only effective way to insert genomic variation-based guidance into clinical care (but less than 10% of hospitals have this capability)
The face of personalized medicine
Vanderbilt PREDICT: Pharmacogenomic Resource for Enhanced Decisions in Care and Therapy
(Go-live date: Sept 22, 2010)

- Use data mining methods in Electronic Medical Record (EMR) to identify individuals at increased likelihood of a future prescription of a drug for which pharmacogenetics has relevance
- Prospectively acquire 200 marker DNA panel and put selected subset of data in electronic medical record
- At moment of prescribing, use decision support rules to guide drug selection and correct dosing
Decision support alert for personalized molecular guidance at the teachable moment of drug prescribing

Clopidogrel Poor Metabolizer Rules

Genetic testing has been performed and indicates this patient is at risk for inadequate anti-platelet response to clopidogrel (Plavix) therapy.

This patient has been tested for CYP2C19 variants, and the presence of the "2/2" genotype has identified this patient as a poor metabolizer of clopidogrel. Poor metabolizers treated with clopidogrel at normal doses exhibit higher rates of stent thrombosis/other cardiovascular events.

Treatment modification is recommended:
- Prescribe prasugrel (EFFIENT) 10mg daily and stop clopidogrel (Plavix) startdate, 10AM

Due to increased risk of bleeding, prasugrel should not be given to patients:
- Those with a history of stroke or transient ischemic attack
- Those greater than 75 years of age
- Those whose body weight is less than 66 kg

Click here for more information

If prasugrel (EFFIENT) not selected, please choose desired action:
- Increase maintenance dose of clopidogrel (Plavix) 150 mg daily, startdate, 10AM
- Maintain requested daily dose of clopidogrel (Plavix) 75 mg daily, startdate, 10AM

- Contraindicated
- Expected effects (e.g., nuisance bleeding)
- Patient preference
- Other

Click here for more information

NOTE: The Vanderbilt P&T Committee has recommended that prasugrel (if not contraindicated) should replace clopidogrel for poor metabolizers; if this is not possible consider doubling the standard dose of clopidogrel (or, use standard dose clopidogrel). However, there is not a national consensus on drug/dose guidance in this population.
Genomic data as viewed by patients
This test examines your gene known as CYP2C19 (sounds like "sip-2-C-19"). CYP2C19 can affect your response to a drug called clopidogrel (sounds like "klöh-PID-oh-grel"). Clopidogrel has the brand name Plavix. Clopidogrel is used to help prevent harmful blood clots from developing, such as for people who have had a recent heart attack, or a stroke.

CYP2C19 Results: Your result is *1/*2. This means you may not respond as well to clopidogrel.

Many factors, including this test, help your doctor decide if taking clopidogrel is right for you.
Operational Implementation of Prospective Genotyping for Personalized Medicine: The Design of the Vanderbilt PREDICT Project

JM Pulley¹, JF Peterson²,³, JF Peterson²,³, GR Bernard¹,⁴, CL Vencak-Jones⁵,⁶, AH Ramirez³, JT Delaney³, E Bowton⁴, K Brothers⁵, K Johnson²,⁵, DC Crawford⁷,⁸, J Schildcrout⁹, DR Mays⁵,³, HH Dilks⁵, RA Wilke³, EW Clayton⁵,¹⁰, E Shultz²,³, M Laposata³,⁶, J McPherson³, JN Jirjis²,³ and DM Roden³,¹¹

The promise of "personalized medicine" guided by an understanding of each individual's genome has been fostered by increasingly powerful and economical methods to acquire clinically relevant information. We describe the operational implementation of prospective genotyping linked to an advanced clinical decision-support system to guide...
Peril #2

- Our ability to acquire person-specific DNA data far exceeds our understanding of its meaning.
- Genetic data conclusively explains the basis for only a tiny set of the 8000+ diseases of humans and responses to therapy.
- As a result DNA data acquired now will likely need to be accessed and re-interpreted many times over in the future as DNA science unfolds, accommodating the health literacy of the reader.
Peril #3

- DNA is in some cases a (probabilistic) ‘future diary’ of events yet to occur.
- It may be used to predict future health risks that affect:
  - Employability
  - Insurability
  - Social standing
- Low science literacy and relatively high paranoia related to genetics among lay public.
So, in Information Retrieval for Personalized Healthcare...

We is confronted by an insurmountable tidal wave of opportunity

Pogo
**SIG-IR: Your Job**

- Create better approaches to retrieval and linkage of specific molecular patterns with a large, rapidly evolving corpus of interpretive knowledge.

- Retrieve information in a way that is:
  - Context sensitive for intended use
  - Adaptive to variable health literacy of users
  - Culturally aware
The lesson for molecular information retrieval from Jeopardy!

- Hint: Not Watson
The lesson for molecular information retrieval from Jeopardy!

- Question to Alex Trebek: Is Jeopardy about rewarding smart people with money?
- Answer: Jeopardy is about fostering a lifelong love of learning
SIG-IR: Your Job

- Create better approaches to retrieval and linkage of specific molecular patterns with a large, rapidly evolving corpus of interpretive knowledge that are:
  - Context sensitive for intended use
  - Adaptive to variable health literacy of users
  - Culturally aware

...and fosters a lifelong love of learning