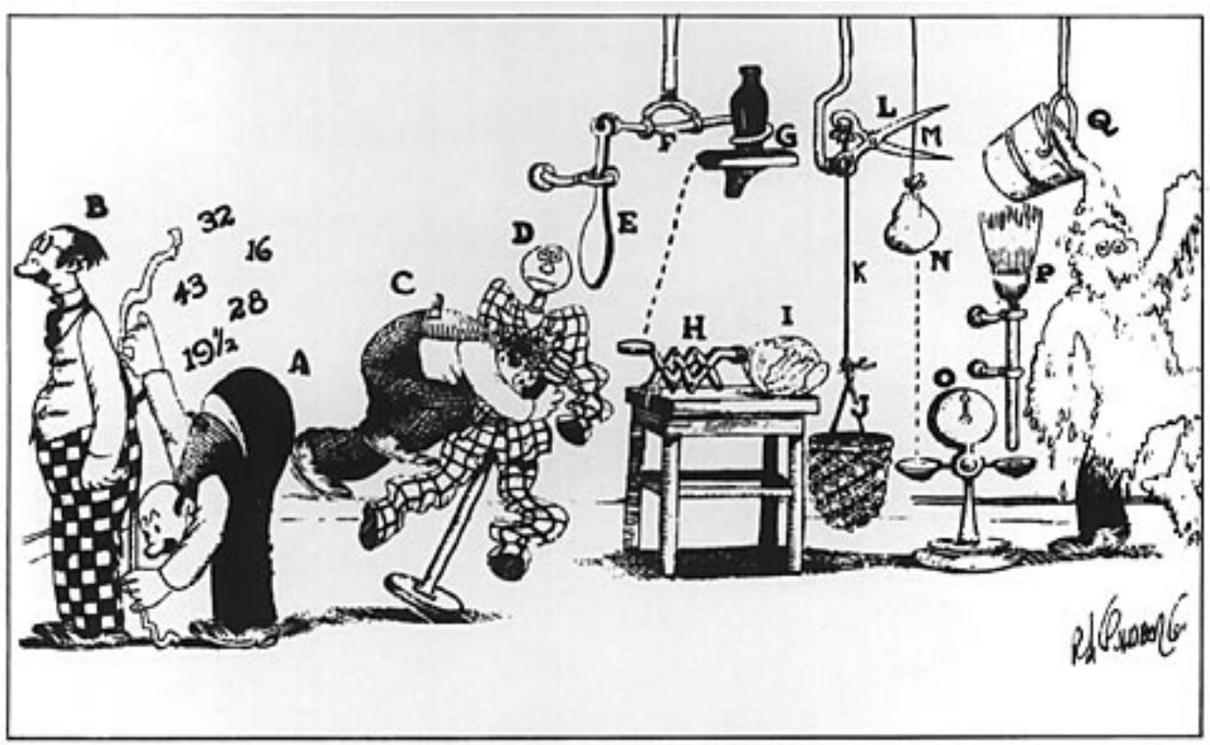
ThoughtWorks®

Lisp & Cancer



Idea For Dodging Bill Collectors RUBE GOLDBERG (tm) RGI 046

Ola Bini

computational metalinguist

ola.bini@gmail.com

http://olabini.com/blog

698E 2885 C1DE 74E3 2CD5 03AD 295C 7469 84AF 7F0C

The problem

Genomics in one slide

The human genome: nuclear DNA and mitochondrial DNA

Nuclear DNA: 22 chromosomes * 2 + (XX || XY)

DNA is a helix spiral, each side is complementary to the other side. (ACGT, A complements T, C complements G)

DNA gets *transcribed* into mRNA

Actually, it transcribes into precursor RNA, then splicing happens

mRNA gets *translated* into proteins (polypeptides)

Proteins do stuff (including transcription and translation)

1 codon = 1 amino acid

1 codon = 3 bases of DNA, which means a 6bit byte code machine

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC } Phe UUA, UUG } Leu	UCU, UCC, UCA, UCG } Ser	UAU, UAC } Tyr UAA Stop UAG Stop	UGU, UGC } Cys UGA Stop UGG Trp	U C A G
C	CUU, CUC, CUA, CUG } Leu	CCU, CCC, CCA, CCG } Pro	CAU, CAC } His CAA, CAG } Gln	CGU, CGC, CGA, CGG } Arg	U C A G
A	AUU, AUC, AUA } Ile AUG Met	ACU, ACC, ACA, ACG } Thr	AAU, AAC } Asn AAA, AAG } Lys	AGU, AGC } Ser AGA, AGG } Arg	U C A G
G	GUU, GUC, GUA, GUG } Val	GCU, GCC, GCA, GCG } Ala	GAU, GAC } Asp GAA, GAG } Glu	GGU, GGC, GGA, GGG } Gly	U C A G

Sequencing

Taking DNA and turning it into bits

Steps

- Prepare the analyte
- Shred the DNA into 200bp long segments (called *reads*)
- Sequence all the reads separately
- Find overlapping reads (*assembly*)
- Find where the reads belong by comparing to a reference (*alignment*)
- Optional: compare against another genome and output the results (*variant calling*)

The $1000 genome

Cancer

- Not one disease - at least 10 000 diseases
- Organ of origin less interesting than molecular make up
- Cancer is modifications of DNA in various ways
  - Stops apoptosis
  - Enhances G cell cycle (growth)
  - Removes error correcting mechanisms
- Through genetic modifications of various kinds
- Driver mutations vs passenger mutations
- Lots of noise

The treatment problem

- Standard of care is based on organ
- Ovarian cancer has ca 3 first level chemo's
  - If one doesn't work, try the next
- But they're expensive: $100 000 for a round
  - And 3 months of time
  - And severe pain and damage to the body
- The information is out there
  - In research papers
  - In clinical trial data

Some numbers

Base pairs in a human: ~ 3 000 000 000

Germline mutations per person: ~ 5 000 000

Proteins in a human: ~ 100 000

Genes in a human: ~ 21 000

The size of a genome after sequencing: 0.5Tb

Our solution

Our solution

- Suck in data from lots of resources
  - Unify and normalize
  - Types of data
    - Patient
    - Reference
    - Experience
- Put everything in a graph
- Model biology
- Enhance raw information with deduced information
- Connect up treatments in relationships with biomarkers

Reference: DrugBank


```
<drug type="biotech" created="2005-06-13 07:24:05 -0600" updated="2011-07-31 23:04:49 -0600"
version="3.0">
  <drugbank-id>DB00002</drugbank-id>
  <name>Cetuximab</name>
  <description>Epidermal growth factor receptor binding FAB. Cetuximab is composed of the Fv
(variable; antigen-binding) regions of the 225 murine EGFr monoclonal antibody specific for
the N-terminal portion of human EGFr with human IgG1 heavy and kappa light chain constant
(framework) regions.</description>
  <cas-number>205923-56-4</cas-number>
  <general-references></general-references>
  <synthesis-reference></synthesis-reference>
  <indication>For treatment of EGFR-expressing metastatic colorectal cancer in patients who
are refractory to other irinotecan-based chemotherapy regimens. Cetuximab is also indicated
for treatment of squamous cell carcinoma of the head and neck in conjucntion with radiation
therapy.</indication>
  <pharmacodynamics>Used in the treatment of colorectal cancer, cetuximab binds specifically
to the epidermal growth factor receptor (EGFr, HER1, c-ErbB-1) on both normal and tumor cells.
EGFr is over-expressed in many colorectal cancers. Cetuximab competitively inhibits the
binding of epidermal growth factor (EGF) and other ligands, such as transforming growth
factor&#x2013;alpha. Binding of cetuximab to the EGFr blocks phosphorylation and activation of
receptor-associated kinases, resulting in inhibition of cell growth, induction of apoptosis,
decreased matrix metalloproteinase secretion and reduced vascular endothelial growth factor
production.</pharmacodynamics>
```

Reference: DrugBank


```
    <partner id="6">
      <name>Coagulation factor XIII A chain</name>
      <general-function>Involved in protein-glutamine gamma-glutamyltransferase activity</
general-function>
      <specific-function>Factor XIII is activated by thrombin and calcium ion to a
transglutaminase that catalyzes the formation of gamma-glutamyl- epsilon-lysine cross-links
between fibrin chains, thus stabilizing the fibrin clot. Also cross-link alpha-2-plasmin
inhibitor, or fibronectin, to the alpha chains of fibrin</specific-function>
      <gene-name>F13A1</gene-name>
      <locus>6p25.3-p24.3</locus>
      <reaction>protein glutamine + alkylamine = protein N5-alkylglutamine + NH3</reaction>
      <signals>None</signals>
      <cellular-location>Cytoplasm. Secreted protein. Secreted into the blood plasma.
Cytoplasmic in most tissues, but also s</cellular-location>
      <transmembrane-regions>None</transmembrane-regions>
      <theoretical-pi>5.95</theoretical-pi>
      <molecular-weight>83137</molecular-weight>
      <chromosome></chromosome>
      <essentiality>Non-Essential</essentiality>
```

Reference: NCBI 36.3

```
9606 12   25011328 25011705 -   NT_009714.16 17879035 17879412 -   BRI3P2   GeneID:441630
    GENE reference    -   protein;;
9606 12   25037625 25041640 -   NT_009714.16 17905332 17909347 -   LOC196415    GeneID:
196415   GENE reference    -   best RefSeq;identical;N
9606 12   25048543 25069167 +   NT_009714.16 17916250 17936874 +   LOC645177    GeneID:
645177   GENE reference    -   mRNA;identical;N
9606 12   25096508 25152536 +   NT_009714.16 17964215 18020243 +   LRMP GeneID:4033  GENE
    reference    -   best RefSeq;identical;N
9606 12   25152490 25239361 -   NT_009714.16 18020197 18107068 -   CASC1    GeneID:55259 GENE
    reference    -   best RefSeq;identical;N
9606 12   25239417 25249216 +   NT_009714.16 18107124 18116923 +   LYRM5    GeneID:144363
    GENE reference    -   best RefSeq;identical;N
9606 12   25249447 25295121 -   NT_009714.16 18117154 18162828 -   KRAS GeneID:3845  GENE
    reference    -   best RefSeq;identical;N
9606 12   25453475 25453646 -   NT_009714.16 18321182 18321353 -   LOC100133222 GeneID:
100133222    GENE reference    -   mRNA;identical;N
9606 12   25520283 25597445 -   NT_009714.16 18387990 18465152 -   IFLTD1   GeneID:160492
    GENE reference    -   best RefSeq;mismatch;N
```

Patient data: MAF

```
AAAS 8086 broad.mit.edu 36 12 51987696 51987696 + Silent SNP G G A novel
    none TCGA-13-2060-01A-01W-0799-08 TCGA-13-2060-10A-01W-0799-08 G G A A G
    G Unknown Valid Somatic Phase_I Capture Illumina GAIIx
AACS 65985 broad.mit.edu 36 12 124184519 124184519 + Missense_MutationSNP
    G G T novel none TCGA-25-2393-01A-01W-0799-08 TCGA-25-2393-10A-01W-0799-08
    G G T T G G Unknown Valid Somatic Phase_I Illumina
GAIIx
AACS 65985 broad.mit.edu 36 12 124184519 124184519 + Missense_MutationSNP
    G G T novel none TCGA-25-2393-01A-01W-0799-08 TCGA-25-2393-10A-01W-0799-08
    G G T T G G Unknown Valid Somatic Phase_I Capture Illumina
GAIIx
AADACL4 343066 broad.mit.edu 36 1 12648632 12648632 + Missense_MutationSNP A
    A C novel none TCGA-13-0913-01A-01W-0420-08 TCGA-13-0913-10A-01D-0399-08 A
    A C C A A Unknown Valid Somatic Phase_I Illumina GAIIx
AADACL4 343066 broad.mit.edu 36 1 12648632 12648632 + Missense_MutationSNP A
    A C novel none TCGA-13-0913-01A-01W-0420-08 TCGA-13-0913-10A-01D-0399-08 A
    A C C A A Unknown Valid Somatic Phase_I Capture Illumina
GAIIx
    C C G novel none TCGA-25-2392-01A-01W-0799-08 TCGA-25-2392-10A-01W-0799-08
    C C G G C C Unknown Valid Somatic Phase_I Capture Illumina
GAIIx
ABCA1 19 broad.mit.edu 36 9 106634791 106634791 + Missense_MutationSNP
    G G A novel none TCGA-24-1471-01A-01W-0551-08 TCGA-24-1471-10A-01W-0551-08
    G G A A G G Unknown Valid Somatic Phase_I Illumina
GAIIx
```

CGH

Chromosome	Start	End	Probe_Number	Segment_Mean
1	554267	2279815	127	-0.0462
1	2296033	10898771	773	-0.5518
1	10913017	11307517	47	-0.3344
1	11352106	11726937	32	-0.5429
1	11738095	11951932	26	0.0457
1	11958137	13671229	93	-0.4834
1	13676991	14371177	44	-0.331
1	14405510	14502895	5	-0.1132
1	14538345	14712399	7	0.1963
1	14736245	15991065	148	-0.047
1	15995929	16002006	2	1.9722
1	16007963	72533855	5031	-0.0296
1	72550247	72568008	2	2.279
1	72602596	150839753	3961	5e-04
1	150844443	150848508	2	0.8553
1	150857069	194978217	3721	-0.013
1	195005519	195067763	7	-0.5762
1	195091757	200613453	478	-0.0112
1	200622655	200786281	12	-0.3698

Experience Data

Molecule	Alias (molecule)	DNA/mRNA/Protein	State (molecule)	Modifier	Alias (modifier)	Relationship	Drug (Therapy)	Alias (drug)	Model	H	Cases	Reference
BRAF	B-Raf	DNA mut V600E	sensitivity to	17-AAG	Tanespimycin (17-allylamino-17-demethoxygeldanamycin)	3	1					Da Rocha Dias S, Cancer Res 2005, 65:10686-91
GSK3B		mRNA	downregulated by	GSK3B siRNA		sensitivity to	Sorafenib	Nexavar (R)	3	1		Panka DJ, J Biol Chem 2008, 283:726-32
TYRO3	Sky, TYRO3 protein tyrosine kinase	mRNA	downregulated by	TYRO3 siRNA		sensitivity to	Cisplatin	CDDP	3	1		Zhu S, Proc Natl Acad Sci USA 2009, 106:17025-30
CXCR1		mRNA	expressed			sensitivity to	CXCR1 siRNA		4	1		Singh S, Int J Cancer 2010, 126:328-36
KIT	c-KIT	DNA mut V560A (exon 11)	sensitivity to	Sorafenib	Nexavar (R)	5	1	1				Quintas-Cardama A, Nat Clin Pract Oncol 2008, 5:737-40
KIT	c-KIT	DNA mut K642E (exon 13)	sensitivity to	Imatinib	Gleevec (R)	5	1	1				Lutzky J, Pigment Cell Melanoma Res 2008, 21:492-3
ERBB3	HER3	mRNA	downregulated by	HER3 siRNA		sensitivity to	Dacarbazine	DTIC	3	1		Reschke M, Clin Cancer Res 2008, 14:5188-97
HERG	KCNH2	mRNA	expressed			sensitivity to	HERG siRNA		3	1		Afrasiabi E, Cell Signal 2010, 22:57-64
E2F1	E2F1 transcription factor	mRNA	expressed			sensitivity to	E2F1 siRNA		4	1		Alla V, J Natl Cancer Inst 2010, 102:127-33
KIT	c-KIT	protein	expressed			sensitivity to	Imatinib	Gleevec (R)	3	1		All-Ericsson C, Invest Ophthalmol Vis Sci 2004, 45:2075-82
KIT	c-KIT	DNA mut D820Y				sensitivity to	Sunitinib	Sutent (R)	3	1		Ashida A, Int J Cancer 2009, 124:862-8
KIT	c-KIT	DNA mut D820Y				no relationship with	Imatinib	Gleevec (R)	3	0		Ashida A, Int J Cancer 2009, 124:862-8
MGMT	O-6-methylguanine-DNA methyltransferase		expressed			resistance to	Temozolomide	TMZ	3	-1		Augustine CK, Clin Cancer Res 2009, 15:502-10
MGMT	O-6-methylguanine-DNA methyltransferase	protein	active (high activity)			resistance to	Temozolomide	TMZ	3	-1		Augustine CK, Clin Cancer Res 2009, 15:502-10
MGMT	O-6-methylguanine-DNA methyltransferase	DNA methylated				no relationship with	Temozolomide	TMZ	3	0		Augustine CK, Clin Cancer Res 2009, 15:502-10
EPAC		mRNA	expressed			sensitivity to	EPAC siRNA		3	1		Baljinnyam E, Am J Physiol Cell Physiol 2009, 297:C802-13
Mcl-1			expressed			resistance to	ARC		3	-1		Bhat UG, Cell Cycle 2008, 7:1851-5
Mcl-1			expressed			resistance to	Siomycin A		3	-1		Bhat UG, Cell Cycle 2008, 7:1851-5
KIT	c-KIT	DNA amplified (gene)				sensitivity to	Imatinib	Gleevec (R)	5	1	1	Carvajal RD, J Clin Oncol 27:15s, 2009 (suppl; abstr 9001)
KIT	c-KIT	DNA mut ? (exon 11)				sensitivity to	Imatinib	Gleevec (R)	5	1	3	Carvajal RD, J Clin Oncol 27:15s, 2009 (suppl; abstr 9001)
BRAF	B-Raf	DNA mut V600E				no relationship with	Sorafenib	Nexavar (R)	5	0	34	Eisen T, Br J Cancer 2006, 95:581-6
BRAF	B-Raf	DNA mut V600E				no relationship with	Sorafenib	Nexavar (R)	5	0	37	Min CJ, J Clin Oncol 2008, 26: abstract 9072
BRAF	B-Raf	DNA mut V600E				sensitivity to	RAF-265		4	1		Fecher LA, Pigment Cell Melanoma Res 2008, 21:410-1
BRAF	B-Raf	DNA mut V600E				resistance to	Sorafenib	Nexavar (R)	3	-1		McDermott U, Proc Natl Acad Sci USA 2007, 104:19936-41
BRAF	B-Raf	DNA mut V600E				sensitivity to	AZ628		3	1		McDermott U, Proc Natl Acad Sci USA 2007, 104:19936-41
BRAF	B-Raf	DNA mut V600E				sensitivity to	PLX4032	RG7204	3	1		Sala E, Mol Cancer Res 2008, 6:751-9

Experience Data

Molecule	Alias (molecule)	DNA/mRNA/Protein	State (molecule)	Modifier	Alias (modifier)	Relationship	Drug (Therapy)	Alias (drug)	Model	H	Cases	Reference
BRAF	B-Raf	DNA mut V600E	sensitivity to	17-AAG	Tanespimycin (17-allylamino-17-demethoxygeldanamycin)	3	1					Da Rocha Dias S, Cancer Res 2005, 65:10686-91
GSK3B		mRNA	downregulated by	GSK3B siRNA		sensitivity to	Sorafenib	Nexavar (R)	3	1		Panka DJ, J Biol Chem 2008, 283:726-32
TYRO3	Sky, TYRO3 protein tyrosine kinase	mRNA	downregulated by	TYRO3 siRNA		sensitivity to	Cisplatin	CDDP	3	1		Zhu S, Proc Natl Acad Sci USA 2009, 106:17025-30
CXCR1		mRNA	expressed	sensitivity to	CXCR1 siRNA		4	1				Singh S, Int J Cancer 2010, 126:328-36
KIT	c-KIT	DNA mut V559A (exon 11)	sensitivity to	Sorafenib	Nexavar (R)	5	1	1				Quintas-Cardama A, Nat Clin Pract Oncol 2008, 5:737-40
KIT	c-KIT	DNA mut K642E (exon 13)	sensitivity to	Imatinib	Gleevec (R)	5	1	1				Lutzky J, Pigment Cell Melanoma Res 2008, 21:492-3
ERBB3	HER3	mRNA	downregulated by	HER3 siRNA		sensitivity to	Dacarbazine	DTIC	3	1		Reschke M, Clin Cancer Res 2008, 14:5188-97
HER2	c-ERB2	mRNA	expressed	sensitivity to	HER2 siRNA		3	1				Afrasiabi E, Cell Signal 2010, 22:57-64
E2F1	E2F1 transcription factor	mRNA	expressed	sensitivity to	E2F1 siRNA		4	1				Alla V, J Natl Cancer Inst 2010, 102:127-33
KIT	c-KIT	protein	expressed	sensitivity to	Imatinib	Gleevec (R)	3	1				All-Ericsson C, Invest Ophthalmol Vis Sci 2004, 45:2075-82
KIT	c-KIT	DNA mut D820Y	sensitivity to	Sunitinib	Sutent (R)	3	1					Ashida A, Int J Cancer 2009, 124:862-8
KIT	c-KIT	DNA mut D820Y	no relationship with	Imatinib	Gleevec (R)	3	0					Ashida A, Int J Cancer 2009, 124:862-8
MGMT	O-6-methylguanine-DNA methyltransferase		expressed	resistance to	Temozolomide	TMZ	3	-1				Augustine CK, Clin Cancer Res 2009, 15:502-10
MGMT	O-6-methylguanine-DNA methyltransferase	protein	active (high activity)	resistance to	Temozolomide	TMZ	3	-1				Augustine CK, Clin Cancer Res 2009, 15:502-10
MGMT	O-6-methylguanine-DNA methyltransferase	DNA methylated	no relationship with	Temozolomide	TMZ	3	0					Augustine CK, Clin Cancer Res 2009, 15:502-10
EPAC		mRNA	expressed	sensitivity to	EPAC siRNA		3	1				Baljinnyam E, Am J Physiol Cell Physiol 2009, 297:C802-13
Mcl-1			expressed	resistance to	ARC	3	-1					Bhat UG, Cell Cycle 2008, 7:1851-5
Mcl-1			expressed	resistance to	Siomycin A	3	-1					Bhat UG, Cell Cycle 2008, 7:1851-5
KIT	c-KIT	DNA amplified (gene)	sensitivity to	Imatinib	Gleevec (R)	5	1	1				Carvajal RD, J Clin Oncol 27:15s, 2009 (suppl; abstr 9001)
KIT	c-KIT	DNA mut ? (exon 11)	sensitivity to	Imatinib	Gleevec (R)	5	1	3				Carvajal RD, J Clin Oncol 27:15s, 2009 (suppl; abstr 9001)
BRAF	B-Raf	DNA mut V600E	no relationship with	Sorafenib	Nexavar (R)	5	0	34				Eisen T, Br J Cancer 2006, 95:581-6
BRAF	B-Raf	DNA mut V600E	no relationship with	Sorafenib	Nexavar (R)	5	0	37				Min CJ, J Clin Oncol 2008, 26: abstract 9072
BRAF	B-Raf	DNA mut V600E	sensitivity to	RAF-265		4	1					Fecher LA, Pigment Cell Melanoma Res 2008, 21:410-1
BRAF	B-Raf	DNA mut V600E	resistance to	Sorafenib	Nexavar (R)	3	-1					McDermott U, Proc Natl Acad Sci USA 2007, 104:19936-41
BRAF	B-Raf	DNA mut V600E	sensitivity to	AZ628		3	1					McDermott U, Proc Natl Acad Sci USA 2007, 104:19936-41
BRAF	B-Raf	DNA mut V600E	sensitivity to	PLX4032	RG7204	3	1					Sala E, Mol Cancer Res 2008, 6:751-9

BRAF mut V600E
sensitivity to 17-AAG
model:3 h:1
Da Rocha Dias S, Cancer Res 2005, 65:10686-91

TPMT for a specific patient

TPMT for a specific patient

TPMT for a specific patient

{

G460V

A719G

}

TPMT for a specific patient

G460V

A719G

genotype *3A

TPMT for a specific patient

G460V

A719G

genotype *3A

activity low

TPMT for a specific patient

G460V

A719G

genotype *3A

activity low

expression downregulated

Models

1 = Animal, in vitro

2 = Animal, in vivo

3 = Human, in vitro

4 = Human xenograft

5 = Clinical study / non randomized clinical trial

6 = Randomized controlled trial

7 = Meta-analysis

$$\text{weight}(\text{model}) = 2^{\text{model-1}}\ 6$$

```
(defn size-score [cases]
  (condp = cases
    nil 1
    0 1
    (/ cases 10)))

(defn evidence-line-score [{:keys [model cases]}]
  (* (model-score model)
     (size-score cases)))

(defn h-category [{h :h}]
  (condp = h
    0 :null
    1 :positive
    -1 :negative))

(defn merge-scores [last-score line]
  (merge-with + last-score {(h-category line) (evidence-line-score line)}))

(defn evidence-score [evidence-lines]
  (reduce merge-scores
          {:positive 0
           :negative 0
           :null     0}
          evidence-lines))
```

Tech stack

Clojure

Neo4J

JRuby

CoffeeScript

Sinatra

Compojure

Jetty

Kinds of biological data

Not only DNA

RNAseq

Different kinds of variants (structural rearrangements, copy number variants)

Proteins:

- IHC
- PCR

Copies:

- FISH
- CGH

SNP arrays

Some aspects

Clojure and Neo4J

Clojure and Neo4J

Borneo (Kalimantan)

Morph

Indexes

Mutation

Data loading

```
(morph-into ->RawInsightRow
        [patient patient-node]
        :match
          (patient :molecule> igene :molecule>
           rgene [re #{:sensitivity :toxicity :synergism}]
           relationship :drug drug)
          (igene :state s)
          (igene :anchor a)
          (rgene :?at gp :?in chr)
          (relationship :?reference reference)
        :where
          (presence= (state re) (state_name s))
          (not (has (superceded s)))
        :return
          rgene igene relationship drug
          reference chr (state re) s a)
```

Infrastructure

Infrastructure

Go

Puppet

Statsd & Graphite

Fabric & Boto

RSpec for Puppet

RPMs

Recreating boxes on deploy

Why Clojure?

```
(defn cdna->genomic [desc pos]
 (let [strand-positive (= (strand desc) :+)
       exons (if strand-positive (exons desc) (reverse (exons desc)))
       lengths (partition 2 1 (reductions (fn [acc ex] (+ (length ex) acc)) 0 exons))
       {:keys [before target after]} (group-by (before-or-after pos) (map vector exons lengths))
       positions-before (reduce + (map #(length (first %)) before))
       target-exon (first (first target))
       relative (- pos (+ positions-before 1))
       ]
   (when (not target-exon)
    (throw (IllegalArgumentException. (str "cdna position " pos " doesn't point to a coding region"))))
   (if strand-positive
    (+ (start target-exon) (start desc) relative)
    (- (+ (stop target-exon) (start desc)) relative))))
```

```
(defmacro in [gene-name & assertions]
  (let [fforms (map (comp (assert-form gene-name) implication)
                    (remove (comp split? first)
                            (partition-by split? assertions)))]
    `(do
      ~@fforms)))
```

Testing

Conclusions

Molecular Biology is complicated and not well understood

More of you should get into it

Clojure is the only language we could have done this in

This approach is likely the best attack for cancer, short term

Questions?

Ola Bini

ThoughtWorks®

http://olabini.com
obini@thoughtworks.com
@olabini