Extremely high-dimensional medical prediction

Kevin Bleakley

institutCurie
Together, let's beat cancer.

MINES
ParisTech

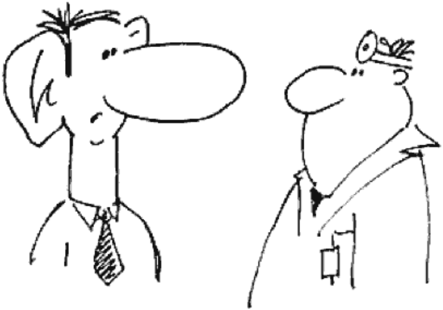
Instituts thématiques

Inserm
Institut national
de la santé et de la recherche médicale

Journées MAS, August 2010

Goals

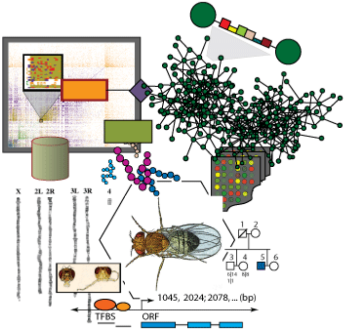
- Diagnosis (esp. when good treatments are available)
- Prognosis (treatment selection, e.g., metastasis or not)
- Drug selection (efficacity, side-effects)



"Just what kind of specialist did you have in mind?"

Common link / Framework

- Biological data ⟶ Prediction ⟶
  - diagnosis
  - prognosis
  - drug selection
  - etc. . .

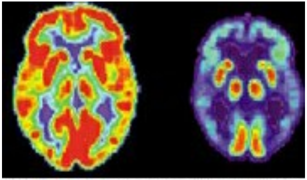


In reality!

- Prediction needs:
  - Extremely high accuracy $\sim 100\%$
  - Very low false negative rate
  - Fairly low false positive rate

Diagnosis

- Case study: Alzheimer's
  - usually only confirmed at autopsy
  - thought to start decade before symptoms
  - when obvious symptoms, maybe too late to save brain
  - goal: find people who are getting it, use them in drug studies
  - not something you want to get wrong!



PET scans of normal brain (left) and an Alzheimer's brain. Photo: U.S. National Institute on Aging

Published last week

- De Meyer et al, "*Diagnosis-Independent Alzheimer Disease Biomarker Signature in Cognitively Normal Elderly People*."
- patients in their 70's
  - 114 normal memories
  - 200 with memory problems
  - 102 with Alzheimer's
- spinal fluid analysed for:
  - amyloid beta (protein fragment that forms plaques in brain)
  - tau (protein accumulates in dead/dying nerve cells in brain)
- researchers 'didn't know' the clinical status of subjects (?!)
- used a 2-component mixture model

Results and remarks

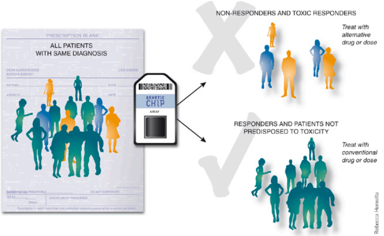
- Results:
  - nearly all with Alzheimer's had 'characteristic' spinal fluid protein levels
  - nearly 3/4 with 'mild' had the signal, all got Alzheimer's within 5 years
  - 1/3 with 'normal' had the signal: suspected future cases?
- Remarks:
  - test already available. Needle in spine!
  - co-author: 'how early do you want to label people?'
  - low-dimensional prediction!
  - infinite number of biological markers they could have chosen
  - vast dimension-reduction using prior biological knowledge

Prognosis. Case study: breast cancer

- Diagnosis ⟶ Clinical Variables ⟶
  - Surgery?
    - breast conserving?
    - masectomy?
    - lymph node dissection?
  - Chemotherapy? (violent)
  - Radiation therapy? (less violent)
  - Hormonal therapy?
  - Targeted therapies? (eg. Herceptin)

What we want to do

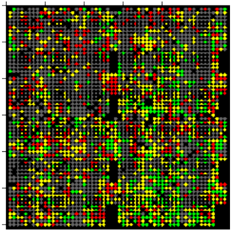
- Predict future of person using data at time t
- Personalise treatment based on this:
  - e.g. breast conserving instead of masectomy
  - less violent (e.g., less chemo) if low risk of recurrence
- Using clinical variables, already ‘personalised’ a bit:
  - tumour grade
  - HER2 status
  - age, etc.
- This is prediction using tens of variables.

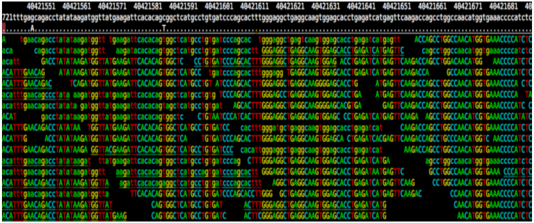


Other data available today

- gene expression $\sim$ 100 Kilo
- SNP data $\sim$ 1 Mega
- Copy number data $\sim$ 1 Mega
- Full genome data $\sim$ 4 Gig

- Prediction: $f(data) \in \{0, 1\}$
  - e.g. metastasis vs no metastasis
  - e.g. drug reaction vs not




40421551 40421561 40421571 40421581 40421591 40421601 40421611 40421621 40421631 40421641 40421651 40421661 40421671 40421681 40
721tttgagcagacctatataagatggttatgaagattcacacagcggctcatgcctgtgatcccagcactttgggaggctgaggcaagtggagcacctgagatcatgagttcaagaccagcctggccaacatggtgaaaccccatctc

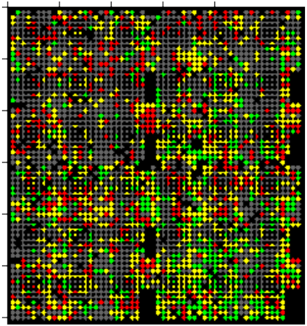
Minor problem

- up to 100's of patients
- up to billions of data dimensions
- binary prediction
- big-time overfitting, statistical problems, it's a mess.
- article: Jelizarow et al. (2010) "*Over-optimism in bioinformatics: an illustration*."

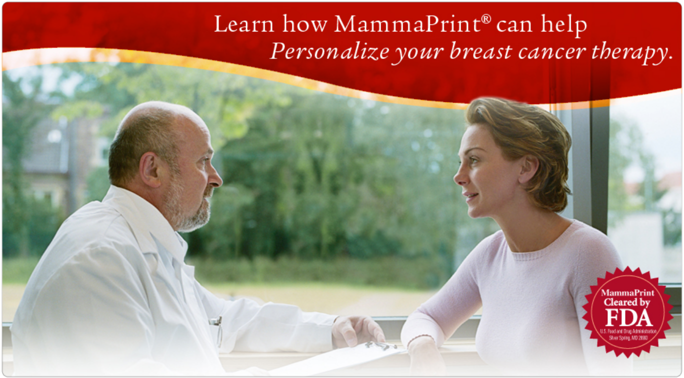
  "We conclude that, if the improvement of a quantitative criterion such as the error rate is the main contribution of a paper, the superiority of new algorithms should always be demonstrated on independent validation data."

Example: gene expression data to predict future metastasis

- every man and his dog has tried to do this
- including me!
- success rates hover around 80 %
- not good enough. or is it . . . ?



MammaPrint


Learn how MammaPrint® can help
Personalize your breast cancer therapy.
MammaPrint
Cleared by
FDA
U.S. Food and Drug Administration
Silver Spring, MD 20993

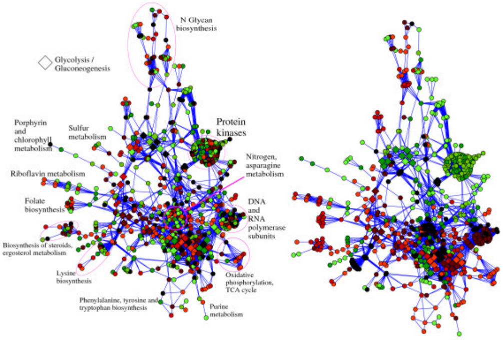
- van 't Veer et al. *Nature* (2002).
- Amsterdam 70-gene breast cancer signature

MammaPrint

- Math: supervised learning on gene expression data of 117 patients.
- Results: "outperforms all currently used clinical parameters in predicting disease outcome."
- Follow-up studies: Van de Vijver et al. *NEJM* (2002).
- Results:
  - 295 patients
  - mean overall 10-year survival rates: 54.6% vs 94.5%
  - probability of remaining free of distant metastases: 50.6% vs 85.2%
- MammaPrint price: $US 4200

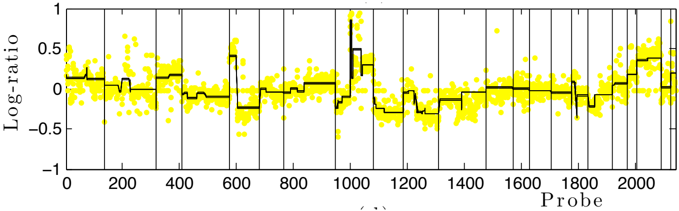
Fundamental question

- black box or. . .
- interpretability/feature selection/stability?
- e.g., Rapaport et al. BMC Bioinformatics (2007).
- a priori connected gene network → supervised classification
- Results: no improvement


N Glycan biosynthesis
Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Protein kinases
Nitrogen, asparagine metabolism
Riboflavin metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative phosphorylation, TCA cycle
Phenylalanine, tyrosine and tryptophan biosynthesis
Purine metabolism

General framework

- high-dim biological data ⟶ dimension reduction using prior biological info simultaneously with classification and/or feature selection
- e.g segmentation of copy-number profiles



Segmentation of copy-number profiles

- prior info: expect piecewise constant signal
- Harchaoui and Lévy-Leduc:

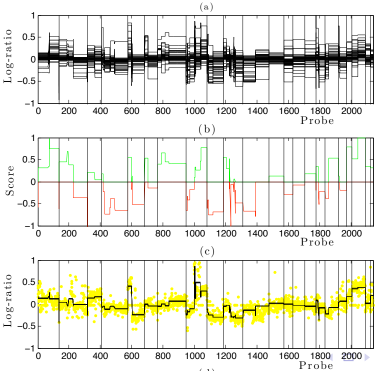
$$\min_{\beta} \|X - \beta\|_2 \qquad \textit{such that} \qquad \sum_{i=2}^{p} |\beta_i - \beta_{i-1}| < \mu$$

- Rapaport et al. (2008) *Bioinformatics*. Fused SVM.

$$\min_{\beta} \sum_{i=1}^{n} \max(0, 1 - y_i \beta^T x_i) \qquad \textit{such that} \ldots$$

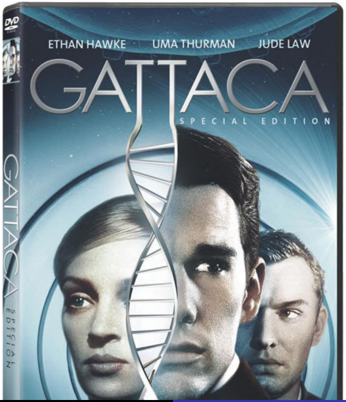
Joint segmentation

- Biological hypothesis: same disease → shared copy number variations.
- J.-P. Vert and K.B.
- test to find regions with common variation
- dimension reduction
- theoretical results



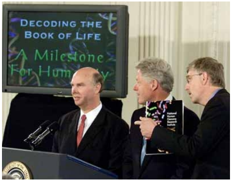
Giga-dimensional biology

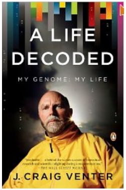
- Revolves around sequencing technology:
  - full genome sequencing
  - CHiP-seq, CNV-seq, methyl-seq, etc.
- ambient dimension of around 4 billion
- how much information is in there? Gattaca anyone?



History

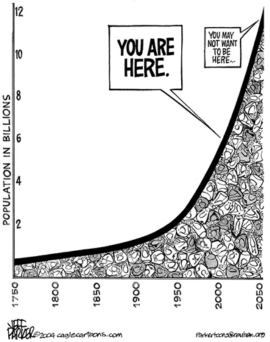
- Watson and Collins (and co.) vs Venter
  - "The genome war"
  - Venter's autobiography
- Speed:
  - $\sim$ ten years for first draft of human genome 1990 – 2001
  - Today: A couple of days for 20x coverage
- Cost:
  - First time: $\sim$ 3 billion \$US
  - Two years ago: 1 million \$US
  - This year: 20 thousand \$US
  - 2015 (or earlier?) 100 \$US in 10 minutes.





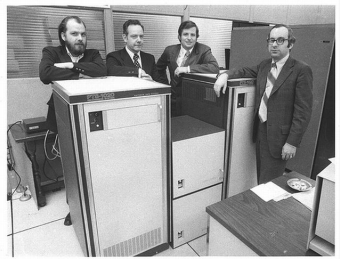
Dimensionality nightmare

- billion-dimensional data
- attempts already started (e.g. SNP studies)
- and. . . world population = 7 billion
- will *you* decide to be sequenced?
- ethics issues. . . but if you are sick, will you decide to be sequenced?

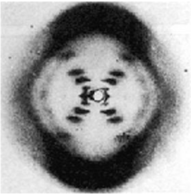


Practical issues

- Data access → contact places with Next-Gen sequencing machines
- Computing hell
- e.g. of computing hell: normalising SNP arrays with 2 million probes
- just getting next-gen sequencing data into a computer network and moving it around is a feat of brilliance (terabytes of image files)



Conclusion



- Give it a try. There's lots to do!
- Good luck!