Machine learning for CRISPR gene editing

Nicolo Fusi
Microsoft Research
Acknowledgements

Broad Institute of MIT and Harvard

John Doench

Meagan Sullender
Mudra Hegde
Emma W. Vaimberg
Katherine Donovan
Ian Smith
David Root

Microsoft Research

Jennifer Listgarten

Washington University School of Medicine

Zuzana Tothova

Dana Farber Cancer Institute

Craig Wilen
Robert Orchard
Herbert W. Virgin
A Powerful New Way to Edit DNA

By Andrew Pollack  March 3, 2014

Could the DNA-editing CRISPR revolutionize medicine?

By Carina Storrs, Special to CNN
Updated 12:22 PM ET, Wed August 12, 2015
No hunger.
No pollution.
No disease.
And the end of life as we know it.
The Genesis Engine.

Editing DNA is now as easy as cut and paste. Welcome to the post-natural world.

A Powerful New Way

By ANDREW POLLACK  MARCH 3, 2014
Stephen B Montgomery @sbmontgom · 2h
With CRISPRs my lab is picking and choosing what X-men they want to be
@wired #crisprfacts @dgmacarthur

smarf dos @smarfdoc · Aug 20
CRISPR can turn you into a baby ALL OVER AGAIN #crisprfacts #late

Matthew Cobb @matthewcobb
CRISPR is both gold AND blue #crisprfacts

Chris Dwan @fdmts
@dgmacarthur @EricTopol CRISPR cannot be overhyped.
CRISPR proves P = NP. #crisprfacts

Henry Scowcroft @oh_henry
If you genetically edit the lettuce genome, you can make it CRISPR
#crisprfacts

Terry D. Johnson @terryjohnson
Peter Jackson worked with CRISPR to edit The Lord of the Rings.
CRISPR was unavailable for The Hobbit. #crisprfacts
Promising results for translational medicine

Proof of principle in stem cells/model organisms:

• Remove CCR5 receptor used by HIV.¹
• Correct a CFTR defect associated with cystic fibrosis.²
• Corrected muscular dystrophy gene to produce cured mice.³

¹ Mandal et al, Cell Stem Cell 2014
² Schwank et al, Cell Stem Cell 2013
³ Long et al, Science 2014
Not quite ready for prime time

Want

Have
Not quite ready for prime time

Two problems and two solutions:
2. Elimination/reduction of “off-target” effects: *Elevation*.

Solution paths:
• Smarter/improved lab protocols.
• Machine learning.
A short intro to CRISPR for gene editing

CRISPR = Clustered Regularly Interspaced Short Palindromic Repeats

crística de la genómica CRISPR para edición genética

cristica de la genómica CRISPR para edición genética

naturally-occurring bacterial immune defense mechanism

co-opted for generic editing in any organism

human disease

drug development

agriculture
Originates from two-part bacterial defense mechanism.

1. **Viral scrapbook**
   - Scrapbook of viral memories (CRISPR)

2. **Cut & paste mechanism**
Gene editing using CRISPR

any organism of choice

synthetic viral memory (aka RNA guide)

DNA to be edited

Cas9
Two problems and two solutions:

2. Elimination/reduction of “off-target” effects: Elevation.

Solution paths:
• Smarter/improved lab protocols.
• Machine learning.
Machine learning predictive modelling for CRISPR

any organism of choice

synthetic viral memory (aka RNA guide)

DNA to be edited

RNA guide

\[ f(\tilde{x}) \]

[\( x_1, x_2, \ldots, x_M \)]

\( y = \text{not effective} \)

[\( x_1, x_2, \ldots, x_M \)]

\( y = \text{effective} \)

[\( x_1, x_2, \ldots, x_M \)]

\( y = \text{very effective} \)

[\( x_1, x_2, \ldots, x_M \)]

\( y = \text{not effective} \)
*In silico* prediction of guide efficiency

Input features
(e.g. guide sequence, GC content of target gene)

Measured guide efficacy
(e.g. “working” vs “not working”)

\[ f(\hat{x}) \]

Model
(e.g. Logistic Regression)

\[ y = \text{effective} \]
Azimuth: our state-of-the-art approach

• Investigate and use richer features of the RNA guide.
• Removed information bottlenecks to the supervised signal.
• Investigate richer model classes.
Azimuth: our state-of-the-art approach

- Investigate and use richer features of the RNA guide.
- Removed information bottlenecks to the supervised signal.
- Investigate richer model classes.
Featurization of a guide

20mer guide

TGGAGGCTGCTTTACCCGCTGTGGGGCGC

NGG PAM

3mer extra context

4mer extra context

\[ \hat{x} = [x_1, x_2, \ldots, x_M] = [0, 1, 1, 0, \ldots, 3.4, 0, 1, 0, 0, 0, 9.8, 0, 0, 0.1] \]
J.A.J.
Just Ask John
Melting temperatures

temperature at which half of the DNA strands are in the random coil or single-stranded (ssDNA) state.

TGGAGGCTGCTTTACCCGCTGTGGGGGCGC

- 30mer
- 5mer proximal to PAM
- 8mer in position 8-15 of 20mer guide
- 5mer in position 3-7 of 20mer guide

[credit: McGovern Institute for Brain Research at MIT]
Additional features improve performance

-** original + new features
-**** original features

evaluated using L1 regression
Azimuth: our state-of-the-art approach

• Investigate and use richer features of the RNA guide.

• Removed information bottlenecks to the supervised signal.

• Investigate richer model classes.
Non-linear modelling

• Simple linear models are incapable of representing or capturing complex interactions between the variables.

• For the final model we use **Gradient-Boosted Regression Trees (GBRTs)**

• An ensemble of weak predictors (regression trees).

• Each RT is trained on the residuals of the previous one.

• GBRTs can easily handle non-homogeneous data (mix of categorical and continuous).
Systematic comparison of models
Impact of our Azimuth model

• *Nature Biotechnology* 2016.
• Recommended by independent studies (Haeussler et al. 2016).
• **Adoption** by two startups and academics/researchers worldwide.
• Azure ML service ~**1000 requests/day**, doubling every 3 months
• Web service ~**300 requests/day**.
• Over **1000 open-source software downloads**.

Not quite ready for prime time

Two problems and two solutions:
2. Elimination/reduction of “off-target” effects: Elevation.

Solution paths:
• Smarter/improved lab protocols.
• Machine learning.
Elevation: prediction of off-target effects

Much more challenging than on-target:

• For just one single guide need to check for imperfect matches genome-wide.

• Combinatorial explosion of mismatches, hard to get enough training data.

intended target

GGCTGCTTTACCCGCTGTGGG

...CTATAACTGGCAGCTCTACCCGGTGTGGGACAAG...

whole genome—potential off-targets
Combinatorial explosion (for 1 guide in 1 gene)

1 mismatch: 69 sites
2 mismatches: 2277 sites
3 mismatches: 47,817 sites
4 mismatches: 717,255 sites
5 mismatches: 8,176,707 sites

very sparsely sampled across different genes

1 full example
Previous state-of-the-art approach: CFD (Doench et al 2016)

intended target

\[
\text{GGCTGCTTTACCCGCTGTGGG}
\]

\[
\text{CTATAACTGGCAGCTCTACCCGGTGTGGGACAAG...}
\]

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

featurization

\[
T:C,8
\]

categorical (i.e. one-hot) encoding of single mismatch and position
Previous state-of-the-art approach: $CFD$ (Doench et al 2016)

$$CFD \approx \prod_i P(Y = 1|X_i = 1)$$

- Measured off-target activities (on a continuous scale) are discretized in present $1$ vs not present $0$.
- CFD computes probability of off-target given mismatch.
- Probabilities are aggregated assuming conditional independencies.
Elevation: generalizations of CFD

\[ CFD \approx \prod_i P(Y = 1|X_i = 1) \]

1. Change from classification to regression for \( P(Y = 1|X_i = 1) \).
2. Augment the feature space from \( T:C,8 \).
3. Use non-linear regression model for \( P(Y = 1|X_i = 1) \), in particular Boosted Regression trees.
4. Refine predictions with a second model layer using the multi-mismatch data.
Goal 1: make better use of the better-sampled 1 mismatch data

\[ CFD \approx \prod_{i} P(Y = 1|X_i = 1) \]

1. Change from classification to regression for \( P(Y = 1|X_i = 1) \).
2. Augment the feature space from \( T:C,8 \).
3. Use non-linear regression model for \( P(Y = 1|X_i = 1) \), in particular Boosted Regression trees.
4. Refine predictions with a second model layer using the multi-mismatch data.
Goal 2: relax independence and other assumptions using sparsely-sampled data

1. Change from classification to regression for $P(Y = 1 | X_i = 1)$.
2. Augment the feature space from $T:C,8$.
3. Use non-linear regression model for $P(Y = 1 | X_i = 1)$, in particular Boosted Regression trees.
4. Refine predictions with a second model layer using the multi-mismatch data.
Cascading from single mismatch to multi-mismatch

1. **Non-linear** regression model trained on **1-mismatch data**.
   - Complex model capturing interactions
   - Can only compute predictions for 1 mismatch at a time

2. **Linear** model trained on scarce **multi-mismatch data**
   - Relatively simple model
   - Trained on individual and aggregated predictions (e.g. product, sum) from layer 1
Elevation outperforms CFD by 64%

- Elevation spearman $\rho = 0.59$
- CFD spearman $\rho = 0.36$
- 64% improvement ($p = 5.5 \times 10^{-5}$)
*Elevation* performs best on 4/5 other data sets
Mitigation of assumptions

Quantitative correction from the full-assumptions model
Putting it all together

• Elevation cloud prediction server.
• Open source code.
• Framework to efficiently search genome-wide for mismatches and call Azimuth & Elevation.
Acknowledgements

Broad Institute of MIT and Harvard

Meagan Sullender
Mudra Hegde
Emma W. Vaimberg
Katherine Donovan
Ian Smith
David Root

John Doench

Microsoft Research

Jennifer Listgarten

Washington University School of Medicine

Zuzana Tothova

Dana Farber Cancer Institute

Craig Wilen
Robert Orchard
Herbert W. Virgin