Big Data Applications for High-Impact/Emerging Topics: Case Studies on Molecular Biology and Its Applications to COVID-19

Ka-Chun Wong, Jiecong Lin, Shixiong Zhang, Xiangtao Li
Outline

• (30 minutes) Basic Concepts in Molecular Biology
• (60 minutes) Data Application Case Studies for Molecular Biology
• (30 minutes) Data Application Case Studies for COVID-19
Basic Concepts in Molecular Biology

(30 minutes)
Our body
• Our body consists of a number of organs
• Each organ composes of a number of tissues
• Each tissue composes of cells
Cell

• Cell performs two type of functions:
  – Perform chemical reactions necessary to maintain our life
  – Pass the information for maintaining life to the next generation

• Actors:
  – DNA stores and passes information
  – RNA is the intermediate between DNA and proteins
  – Protein performs chemical reactions
Central Dogma

- Central Dogma tells us how we get the protein from a gene. This process is called **gene expression**.

- The expression of gene consists of steps
  - Transcription: DNA $\rightarrow$ mRNA
  - Translation: mRNA $\rightarrow$ Protein
  - Post-translation Modification: Protein $\rightarrow$ Modified protein
DNA

- DNA stores the instructions needed by the cell to perform daily life function.
- It consists of two strands which are interwoven together and forms a double helix.
- Each strand is a chain of some small molecules called nucleotides.
- There are 4 different nucleotides of DNA:
  - adenine (A), cytosine (C), guanine (G), thymine (T).
Nucleotide for DNA

• DNA nucleotide consists of three parts:
  – Deoxyribose
  – Phosphate (bound to the 5’ carbon)
  – Base (bound to the 1’ carbon)
Double stranded DNA

- Normally, DNA is double-stranded within a cell. The two strands are antiparallel. One strand is the reverse complement of another one.
- The double strands are interwoven together and form a double helix.
- One conjectured reason for the double-stranded structure is that it eases DNA replications.
Chromosome

• Usually, DNA is tightly wound around histone proteins and forms a chromosome.

• The total information stored in all chromosomes constitute a genome.

• Example:
  – Human Genome: has 3G base pairs, organized in 23 pairs of chromosomes
Gene

• A gene is a sequence of DNA that encodes for a protein.

• In human genome, it is expected there are 20,000 – 30,000 genes.

• For the gene that encodes for a protein,
  – In Prokaryotic genome, one gene corresponds to one protein.
  – In Eukaryotic genome, one gene can correspond to more than one protein because of the process “alternative splicing”.
Gene

https://www.nature.com/articles/d41586-018-05462-w
More on Gene Structure

- Each Gene has 4 regions
  - **Regulatory region** contains a promoter (and enhancers) which regulate gene transcription.
  - **Coding region** contains the codons for protein. It is also called open reading frame. Its length is a multiple of 3. It must begin with start codon, end with end codon, and the rest of its codons are not end codon.
  - **mRNA transcript** contains 5’ untranslated region + coding region + 3’ untranslated region
Organism Complexity v.s. Genome Size

• E. coli Genome: 5M base pairs

• Human Genome: 3G base pairs

• Amoeba dubia (a single cell organism): 670G base pairs

• Genome size appears not to have any relationship with the complexity of the organism

https://www.waterfall-d-mannose.com/ecoli-uropathogen.html

© Dr. Ka-Chun Wong (CityU HK)
© Dr. Wing-Kin Sung (NUS)
Number of Genes vs. Genome Size

- **Prokaryotic genome:** e.g. E. coli
  - Number of base pairs: 5M
  - Number of genes: 4k
  - Average length of a gene: 1000 bp
- **Eukaryotic genome:** e.g. Human
  - Number of base pairs: 3G
  - Estimated number of genes: 20k – 30k
  - Estimated average length of a gene: 1000-2000 bp

- 90% of the E. coli genome consists of coding regions.
- Less than 3% of the human genome is believed to be coding regions. The rest is called non-coding regions (junk DNA?).
- Genome size appears not to have any relationship with the number of genes!

Note that, before 2001, scientists thought we had 100k genes.
<table>
<thead>
<tr>
<th>Organism</th>
<th># of protein-coding genes</th>
<th># of genes naive estimate: (genome size /1000)</th>
<th>BNID</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV 1</td>
<td>9</td>
<td>10</td>
<td>105769</td>
</tr>
<tr>
<td><em>Influenza A virus</em></td>
<td>10-11</td>
<td>14</td>
<td>105767</td>
</tr>
<tr>
<td>Bacteriophage λ</td>
<td>66</td>
<td>49</td>
<td>105770</td>
</tr>
<tr>
<td>Epstein Barr virus</td>
<td>80</td>
<td>170</td>
<td>103246</td>
</tr>
<tr>
<td>Buchnera sp.</td>
<td>610</td>
<td>640</td>
<td>105757</td>
</tr>
<tr>
<td><em>T. maritima</em></td>
<td>1,900</td>
<td>1,900</td>
<td>105766</td>
</tr>
<tr>
<td><em>S. aureus</em></td>
<td>2,700</td>
<td>2,900</td>
<td>105500</td>
</tr>
<tr>
<td><em>V. cholerae</em></td>
<td>3,900</td>
<td>4,000</td>
<td>105760</td>
</tr>
<tr>
<td><em>B. subtilis</em></td>
<td>4,400</td>
<td>4,200</td>
<td>111448</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>4,300</td>
<td>4,600</td>
<td>105443</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>6,600</td>
<td>12,000</td>
<td>105444</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>20,000</td>
<td>100,000</td>
<td>101364</td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>27,000</td>
<td>140,000</td>
<td>111380</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>14,000</td>
<td>140,000</td>
<td>111379</td>
</tr>
<tr>
<td><em>F. rubripes</em></td>
<td>19,000</td>
<td>400,000</td>
<td>111375</td>
</tr>
<tr>
<td><em>Z. mays</em></td>
<td>33,000</td>
<td>2,300,000</td>
<td>110565</td>
</tr>
<tr>
<td><em>M. musculus</em></td>
<td>20,000</td>
<td>2,800,000</td>
<td>100308</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>21,000</td>
<td>3,200,000</td>
<td>100399, 111378</td>
</tr>
<tr>
<td><em>T. aestivum</em> (hexaploid)</td>
<td>95,000</td>
<td>16,800,000</td>
<td>105448, 102713</td>
</tr>
</tbody>
</table>

RNA

• RNA has both the properties of DNA and protein; for instance,
  – Similar to DNA, it can store and transfer information.
  – Similar to protein, it can form complex 3-dimensional structure and perform some functions.
  – There are 4 different nucleotides for RNA:
    – adenine(A), cytosine(C), guanine(G), uracil(U)
Nucleotide for RNA

- RNA nucleotide consists of three parts:
  - Ribose Sugar (has an extra OH group at 2’)
  - Phosphate (bound to the 5’ carbon)
  - Base (bound to the 1’ carbon)
RNA vs DNA

- RNA is single-stranded.
- The nucleotides of RNA are quite similar to that of DNA, except that it has an extra OH at position 2’. (see the previous slide)
  - Due to this extra OH, it can form more hydrogen bonds than DNA. Thus, RNA can form a complex 3-dimensional structure.
- RNA uses the base U instead of T.
  - U is chemically similar to T.
  - U is also complementary to A.
Protein

• Protein is a molecular sequence with an alphabet of 20 amino acids.
  – The typical length is from 20 to 5000 amino acids.
  – Average protein contains around 350 amino acids.

• Protein folds into three-dimensional shapes, which form the building blocks and perform most of the chemical reactions within a cell.
Amino acid

- Each amino acid consists of
  - Amino group
  - Carboxyl group
  - R group

\[ \text{Amino group} \quad \text{NH}_2 \quad \text{C} \quad \text{R} \quad (\text{the central carbon}) \]

\[ \text{Carboxyl group} \quad \text{H} \quad \text{O} \quad \text{C} \quad \text{OH} \]

© Dr. Ka-Chun Wong (CityU HK)
© Dr. Wing-Kin Sung (NUS)
Polypeptide

- Protein or polypeptide chain is formed by joining the amino acids together via a peptide bond.

- One end of the polypeptide is the amino group, which is called N-terminus. The other end of the polypeptide is the carboxyl group, which is called C-terminus.

\[
\begin{align*}
\text{NH}_2 - \text{C} - \text{C} - \text{OH} & \quad + \quad \text{NH}_2 - \text{C} - \text{C} - \text{OH} \\
\text{R} & \quad \text{R}'
\end{align*}
\]
Classification of amino acids (I)

- 20 common amino acids can be classified into 4 types.

- **Positively charged (basic) amino acids:**
  - Arginine (Arg, R)
  - Histidine (His, H)
  - Lysine (Lys, K)

- **Negatively charged (acidic) amino acids:**
  - Aspartic acid (Asp, D)
  - Glutamic acid (Glu, E)
Classification of amino acids (II)

• Polar amino acids:
  – Overall uncharged, but uneven charge distribution. Can form hydrogen bonds with water. They are called hydrophilic. Often found on the outer surface of a folded protein.
  – Asparagine (Asn, N)
  – Cysteine (Cys, C)
  – Glutamine (Gln, Q)
  – Glycine (Gly, G)
  – Serine (Ser, S)
  – Threonine (Thr, T)
  – Tyrosine (Tyr, Y)
Classification of amino acids (III)

• Non-polar amino acids:
  – Overall uncharged and uniform charge distribution. Cannot form hydrogen bonds with water. They are called hydrophobic. Tend to appear on the inside surface of a folded protein.
  – Alanine (Ala, A)
  – Isoleucine (Ile, I)
  – Leucine (Leu, L)
  – Methionine (Met, M)
  – Phenylalanine (Phe, F)
  – Proline (Pro, P)
  – Tryptophan (Trp, W)
  – Valine (Val, V)
# Amino acid properties

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>3-Letter</th>
<th>1-Letter</th>
<th>Side-chain class</th>
<th>Side-chain polarity</th>
<th>Side-chain charge (pH 7.4)</th>
<th>Hydropathy index</th>
<th>Absorbance λ&lt;sub&gt;max&lt;/sub&gt; (nm)</th>
<th>ε (m&lt;sup&gt;-1&lt;/sup&gt; cm&lt;sup&gt;-1&lt;/sup&gt; M&lt;sup&gt;-1&lt;/sup&gt;)</th>
<th>MW (Weight)</th>
<th>Occurrence in proteins (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
<td>aliphatic</td>
<td>nonpolar</td>
<td>neutral</td>
<td>1.8</td>
<td>257, 206, 188</td>
<td>0.2, 9.3, 60.0</td>
<td>89.094</td>
<td>8.75</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
<td>basic</td>
<td>basic polar</td>
<td>positive</td>
<td>-4.5</td>
<td>174.203</td>
<td>5.79</td>
<td>132.119</td>
<td>3.93</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
<td>amide</td>
<td>polar</td>
<td>neutral</td>
<td>-3.5</td>
<td>133.104</td>
<td>5.49</td>
<td>121.154</td>
<td>1.38</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>Asp</td>
<td>D</td>
<td>acid</td>
<td>acidic polar</td>
<td>negative</td>
<td>-3.5</td>
<td>147.131</td>
<td>6.32</td>
<td>146.164</td>
<td>3.9</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
<td>sulfur-containing</td>
<td>nonpolar</td>
<td>neutral</td>
<td>2.5</td>
<td>250</td>
<td>0.3</td>
<td>75.067</td>
<td>7.03</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu</td>
<td>E</td>
<td>acid</td>
<td>acidic polar</td>
<td>negative</td>
<td>-3.5</td>
<td>146.189</td>
<td>5.19</td>
<td>211.192</td>
<td>2.32</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
<td>amide</td>
<td>polar</td>
<td>neutral</td>
<td>-3.5</td>
<td>146.132</td>
<td>3.9</td>
<td>144.028</td>
<td>3.8</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
<td>aliphatic</td>
<td>nonpolar</td>
<td>neutral</td>
<td>-0.4</td>
<td>70.027</td>
<td>7.03</td>
<td>115.132</td>
<td>5.02</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
<td>basic</td>
<td>aromatic</td>
<td>basic polar</td>
<td>-3.2</td>
<td>211</td>
<td>5.9</td>
<td>155.156</td>
<td>2.26</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
<td>aliphatic</td>
<td>nonpolar</td>
<td>neutral</td>
<td>4.5</td>
<td>131.175</td>
<td>5.49</td>
<td>131.175</td>
<td>9.68</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
<td>aliphatic</td>
<td>nonpolar</td>
<td>neutral</td>
<td>3.8</td>
<td>131.175</td>
<td>9.68</td>
<td>146.189</td>
<td>5.19</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
<td>basic</td>
<td>basic polar</td>
<td>positive</td>
<td>-3.9</td>
<td>146.192</td>
<td>3.87</td>
<td>149.208</td>
<td>2.32</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
<td>sulfur-containing</td>
<td>nonpolar</td>
<td>neutral</td>
<td>1.9</td>
<td>165.192</td>
<td>3.87</td>
<td>115.132</td>
<td>5.02</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
<td>aromatic</td>
<td>nonpolar</td>
<td>neutral</td>
<td>2.8</td>
<td>257, 206, 188</td>
<td>0.2, 9.3, 60.0</td>
<td>105.093</td>
<td>7.14</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
<td>cyclic</td>
<td>nonpolar</td>
<td>neutral</td>
<td>-1.6</td>
<td>280, 219</td>
<td>5.6, 47.0</td>
<td>204.228</td>
<td>1.25</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
<td>hydroxyl-containing</td>
<td>polar</td>
<td>neutral</td>
<td>-0.8</td>
<td>274, 222, 193</td>
<td>1.4, 8.0, 48.0</td>
<td>181.191</td>
<td>2.91</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
<td>hydroxyl-containing</td>
<td>polar</td>
<td>neutral</td>
<td>-0.7</td>
<td>280, 219</td>
<td>5.6, 47.0</td>
<td>181.191</td>
<td>2.91</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
<td>aromatic</td>
<td>nonpolar</td>
<td>neutral</td>
<td>-0.9</td>
<td>274, 222, 193</td>
<td>1.4, 8.0, 48.0</td>
<td>181.191</td>
<td>2.91</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
<td>aromatic</td>
<td>polar</td>
<td>neutral</td>
<td>-1.3</td>
<td>274, 222, 193</td>
<td>1.4, 8.0, 48.0</td>
<td>117.148</td>
<td>6.73</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
<td>aliphatic</td>
<td>nonpolar</td>
<td>neutral</td>
<td>4.2</td>
<td>274, 222, 193</td>
<td>1.4, 8.0, 48.0</td>
<td>117.148</td>
<td>6.73</td>
</tr>
</tbody>
</table>

https://en.wikipedia.org/wiki/Amino_acid
Summary of the amino acid properties

http://www.jalview.org/help/html/misc/properties.gif
Protein Structure Levels

• **Primary structures**
  – Amino acid sequences.

• **Secondary structures**
  – Local 3D structures formed by hydrogen bonding (α-helices and β-sheets) or others (random coils).

• **Tertiary structures**
  – 3D structures of single protein based on the interactions of local 3D structures (i.e. secondary structures) due to the hydrophobic effect.

• **Quaternary structures**
  – 3D structures of multiple proteins, forming a protein complex based on different forces.
Central Dogma

• Central Dogma tells us how we get proteins from the genes. This process is called gene expression.

• The expression of each typical gene consists of steps
  – Transcription: DNA $\rightarrow$ mRNA
  – Translation: mRNA $\rightarrow$ Protein
  – Post-translation Modification: Protein $\rightarrow$ Modified protein
Transcription

• Synthesize a piece of RNA (messenger RNA, mRNA) from one strand of the DNA gene.
  1. An enzyme RNA polymerase temporarily separates the double-stranded DNA
  2. It begins the transcription at the transcription start site.
  3. A → A, C→C, G→G, and T→U
  4. Once the RNA polymerase reaches the transcription terminator, transcription stop.

Translation

- Translation synthesizes a protein from a mRNA.
- In fact, amino acids are encoded by consecutive sequences of 3 nucleotides, called codon.
- The decoding table from codon to amino acid is called Codon Table.
- Note:
  - There are $4^3 = 64$ different codons. Thus, the codons are not one-to-one correspondence to the 20 amino acids.
  - All organisms use the same decoding table!
  - The codons that encode the same amino acid tend to have the same first and second nucleotide.
  - Recall that amino acids can be classified into 4 groups. A single base change in a codon is usually not sufficient to cause a codon to code for an amino acid in different group.
### Codon Table

- **Start codon:** ATG (also code for M)
- **Stop codon:** TAA, TAG, TGA

![Codon Table Diagram](http://bioinfo.bisr.res.in/project/crat/pictures/codon.jpg)
Brief History of Genomics (Bioinformatics)

• 1985: Complete sequencing of genomes of various organisms
• 1990: Launch of Human Genome Project (HGP)
• 2000: By shotgun sequencing, Craig Venter and Francis Collins jointly announced the publication of the first draft of the human genome.

• 2010:
  – ENCODE Project:
    – Annotation of the whole genome including non-coding regions
  – 1000 Human Genomes Project
    – Study of the human population genomes across different regions

• Now:
  – Genotype-Tissue Expression Project (GTEx)
    – To study the relationship between gene expression and genetic variation in human tissue
  – Personal Medicine based on Human Genome
    – To develop customized medical solutions based on genome data
  – The GenomeAsia 100K Project
Data Application Case Studies for Molecular Biology

(60 minutes)
Case Studies

- Case for synthesizing Heterodimeric DNA Motifs [1]
- Case for predicting CRISPR-Cas9 Off-targets [2,3]
- Case for screening Early Cancers from Blood [4]

References


Central Dogma of Molecular Biology:
“DNA makes RNA, RNA makes Protein”
Background

Protein-DNA Binding

• The binding between proteins (e.g. Transcription Factors, TFs) and DNA (e.g. Transcription Factor Binding Sites, TFBSs) play an important role.

• TFs bind in a sequence-specific manner to TFBSs to regulate gene transcription.
Background

• To fully understand a gene’s function, it is essential to identify the TFs that regulate the gene and the corresponding TF-binding sites (also known as DNA Motifs).

• DNA motifs are relatively short (10–20 bp) and highly degenerate sequence motifs, which make their effective identification a challenging task.

A motif logo example
DNA Motif Example

For example, given the following DNA sequences:

<table>
<thead>
<tr>
<th>DNA Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>GAGGTAARC</td>
</tr>
<tr>
<td>TCGTAACT</td>
</tr>
<tr>
<td>CAGGTGGA</td>
</tr>
<tr>
<td>ACAGTCACT</td>
</tr>
<tr>
<td>TAGGTCATT</td>
</tr>
<tr>
<td>TAGTACTG</td>
</tr>
<tr>
<td>ATGTAACCT</td>
</tr>
<tr>
<td>CAGGTATAC</td>
</tr>
<tr>
<td>TGGGTGACT</td>
</tr>
<tr>
<td>AAGGTAACCT</td>
</tr>
</tbody>
</table>

the PWM using relative frequencies is:

\[
M = \begin{bmatrix}
A & 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\
C & 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\
G & 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\
T & 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \\
\end{bmatrix}
\]
DNA Motif Example at 3D Level
### Traditional Research

#### Finding DNA Motifs on DNA Sequences

<table>
<thead>
<tr>
<th>Sequence (with primer)</th>
<th>Normalized Signal Intensity</th>
<th>Protein Binding Microarray</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTAACGATGCAGCGTGTGTTG</td>
<td>3021.785264</td>
<td></td>
</tr>
<tr>
<td>GCGGACGATGACGTCAGTTG</td>
<td>1942.543654</td>
<td></td>
</tr>
<tr>
<td>CTGGGATGGATGACGTCAGTTG</td>
<td>759.8673106</td>
<td></td>
</tr>
<tr>
<td>TGGGGATGGATGACGTCAGTTG</td>
<td>868.2056512</td>
<td></td>
</tr>
<tr>
<td>CTGGGGATGGATGACGTCAGTTG</td>
<td>597.3772212</td>
<td></td>
</tr>
<tr>
<td>AGTTGACGATGACGTCAGTTG</td>
<td>3156.546346</td>
<td></td>
</tr>
<tr>
<td>TGTTGACGATGACGTCAGTTG</td>
<td>982.9229197</td>
<td></td>
</tr>
<tr>
<td>CTGGGGATGGATGACGTCAGTTG</td>
<td>491.8523141</td>
<td></td>
</tr>
<tr>
<td>CTGGGGATGGATGACGTCAGTTG</td>
<td>673.6856787</td>
<td></td>
</tr>
</tbody>
</table>

**Segments of a DeBruijn Sequence**

```
> 3000 sequences
```

**Primer sequence**

```
……………
……………
……………
```

**Normalized Signal Intensity**

```
> 3000 sequences
```

**Find DNA Motifs?**

![Find DNA Motifs](image)

**Protein Binding Microarray**

```
1421.285654
719.1041392
1762.812037
465.8865101
588.0584476
582.1760333
710.6338075
934.8380349
741.7662929
18507.18435
759.8673106
868.2056512
597.3772212
1156.546346
982.9229197
491.8523141
673.6856787
```
Motif finding problem is NP-hard!
Motif discovery by over-representation
- MEME (Expectation Maximization)
- Gibbs Sampling
- SP-STAR
- Winnower
- Random projection
- Mitra
- Weeder
- SPACE
- GALF
Gibbs Sampling: a technique to do random sampling

Random Initialization

Random Deletion

Selection based on $g(\{Z, x'\}/x, \Theta)$

Repeat

Repeat until the selection becomes stable (i.e. converged)

Given a set of sequences $S$, our problem is to find a motif model $\Theta$ and the set of corresponding motif instances $Z$ which maximizes the log likelihood ratio score $g(Z, \Theta)$

$$g(Z, \Theta) = \log \frac{\Pr(Z|\Theta)}{\Pr(Z|\Theta_0)}$$

where $\Theta_0$ is a given background model.
DNA-dependent formation of transcription factor pairs alters their binding specificity

Arttu Jolma, Yimeng Yin, Kazuhiro R. Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova & Jussi Taipale

*Nature* 527, 384–388(2015) | Cite this article

5406 Accesses | 220 Citations | 130 Altmetric | Metrics
Most Updated Research on DNA Motif [1]

Case for synthesizing Heterodimeric DNA Motifs [1]

AI for synthesizing Heterodimeric DNA Motifs [1]

AI for synthesizing Heterodimeric DNA Motifs [1]


https://doi.org/10.1109/DSAA.2015.7344851
We propose to exploit the sequence information and make the Markov assumptions to infer HD from (M1, M2) using input-output hidden Markov models (IOHMMs) (28). Briefly, IOHMM is the generalized variant of generic hidden Markov model (HMM). The difference between IOHMM and HMM is that IOHMM takes into account inputs and outputs with hidden states while HMM only considers outputs with hidden states. Therefore, IOHMMs are chosen in this study since we have to consider the contributions from the input DNA motif model matrices (M1, M2). The IOHMM mathematical modeling $\phi_{IOHMM} = (A, B, C, f, g)$ with hidden states $k \in \{1, 2, ..., K\}$ is defined as follows:

$$
HD[i, j] = A_{s_j [i, j]} M1[i, j] + B_{s_j [i, j]} M2[i, j] + C_{s_j [i, j]}
$$

$$
P(s_j = k) = f(s_{j-1}, s_j, M1[:,j-1], M2[:,j-1], j-1)
$$

$$
P(s_1 = k) = g(s_1, M1[:,1], M2[:,1])
$$

s.t. $\sum_{t=1}^{N} HD[i, j] = 1$

$\forall i \leq 4, j \leq N, k \leq K \in \mathbb{N}$

where $s_j$ denotes the hidden state variable at the $j$-th position; $A_k$ is the regression coefficient matrix for $M1$ at state $k$ in the set $A$; $B_k$ is the regression coefficient matrix for $M2$ at state $k$ in the set $B$; $C_k$ is the bias coefficient matrix at state $k$ in the set $C$; the functions $f$ and $g$ denote the hidden state transition function and initial state estimation function respectively. In this study, given the data availability, we have adopted the classic logistic regression functions as $f$ and $g$ for computational efficiency (28).
AI for synthesizing Heterodimeric DNA Motifs [1]

AI for synthesizing Heterodimeric DNA Motifs [1]

AI for synthesizing Heterodimeric DNA Motifs [1]

ETS_Homeo_ERF_DLX2

AI for synthesizing Heterodimeric DNA Motifs [1]

AI for synthesizing Heterodimeric DNA Motifs [1]


https://doi.org/10.1109/DSAA.2015.7344851
Lastly, we apply the proposed approach to infer previously uncharacterized heterodimeric motifs. Their motif instances are supported by DNase accessibility, gene ontology, protein-protein interactions, ChIP-seq peaks, and even structural data from PDB.

Case for predicting CRISPR-Cas9 Off-targets [2,3]


CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes

ABSTRACT

Genome editing tools such as the clustered regularly interspaced short palindromic repeat (CRISPR)-associated system (Cas) have been widely used to modify genes in model systems including animal zygotes and human cells, and hold tremendous promise for both basic research and clinical applications. To date, a serious knowledge gap remains in our
Pattern Recognition on CRISPR-Cas9 Off-targets [3]

The 2020 Nobel Prize in Chemistry was awarded jointly to Emmanuelle Charpentier and Jennifer A. Doudna for the development of a method for genome editing: the CRISPR/Cas9 system. This technological revolution propelled advances not only in basic research in the life sciences, but also in other fields such as medicine, materials science, and biotechnology. To honor the achievements of the laureates, and to show how this discovery has affected progress in various fields, *ChemBioChem, Angewandte Chemie, Small, Small*
CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences

Yanni Lin¹, Thomas J. Cradick¹, Matthew T. Brown¹, Harshavardhan Deshmukh¹, Piyush Ranjan², Neha Sarode², Brian M. Wile¹, Paula M. Vertino³, Frank J. Stewart² and Gang Bao¹,*

¹Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA, ²School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA and ³Department of Radiation Oncology, Emory University School of Medicine, Atlanta, GA 30322, USA
CRISPR-Cas9 Off-targets

Motivation

Although specific fragments of DNA are aimed, sgRNA can sometimes influence other regions and incur off-target mutations (Chen et al. 2017). CRISPR-Cas9 can tolerate mismatches in sgRNA-DNA at different positions in a sequence-dependent manner, sensitive to the number, position, and distribution of mismatch (Hsu et al. 2013 and Zhang et al. 2015). Since sgRNA is able to endure some mismatches across several nucleotide positions, many off-target sites could be found on the target genome (Kim et al. 2015).

Off-targets can lead to genomic instability and disturb the normal gene function, which is still a major problem when applying CRISPR-Cas9 gene editing to clinical applications. Consequently, we still need accurate off-target prediction methods for complementary purposes.
Motivation

Most of the existing off-target prediction methods just calculate scores based on the positions of the mismatches to the guide sequence. The score of each base pair in sgRNA-DNA is derived using classic bioinformatics statistics on the mismatch effects based on previous gene editing experiments. For example, CFD (Cutting Frequently Determination) score is derived by infecting a large number of sgRNAs with single-bp replacement, deletion or insertion corresponding to the validated sgRNAs in MOLM13 cells; it calculates the percentage activity rates of different mutation sites based on LFC (Log Fold Change) value (Doench et al. 2016).

In light of the above, their performance are vulnerable to experimental variation. Most importantly, the existing methods can not take advantage of the growing CRISPR-Cas9 data for constant self-learning. In addition, most of the existing methods do not consider the potential relationships between mismatched and matched sites, which may be relevant to the off-target activity in CRISPR-Cas9 gene editing (Xu et al. 2017).
Proposed Approach (Deep Neural Network)

<table>
<thead>
<tr>
<th>A</th>
<th>G</th>
<th>C</th>
<th>-</th>
<th>C</th>
<th>G</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>.</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

One-hot Code

<table>
<thead>
<tr>
<th>One-hot Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0 0 0 0 0 0</td>
</tr>
<tr>
<td>0 1 0 0 1 1 0</td>
</tr>
<tr>
<td>0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>0 1 0 0 0 0 0</td>
</tr>
</tbody>
</table>

OR operation on Mismatches

Target DNA

Complementary guide RNA

guide RNA

Fig. 5. An example on how to encode a sgRNA-DNA sequence pair. The table with thick borders in the middle of the figure shows the final matrix code of a sgRNA-DNA sequence pair, which can be used as the input for convolutional neural network modelling.

**Fig. 2.** The architecture of standard deep convolutional neural network (CNN_std) for off-target prediction. The input of this deep neural network is the encoded sgRNA-DNA sequence with length 23, which are converted into hidden features by one convolutional layer. The convolutional layer consists of 40 filters including 10 for each of the sizes $4 \times 1$, $4 \times 2$, $4 \times 3$ and $4 \times 5$. The BN layer is used to normalise the output of the convolutional layer in order to make the training process faster and avoid over-fitting. The global max-pooling layer applies a filter of window size 5 to the previous layers. The output of max-pooling layers are joined together into one vector by flattening. Each neurons in the flattened layer is fully connected to the first dense layer. Two dense layer consists of 100 and 23 neurons respectively. The second dense layer with a drop-out layer is fully connected to two output neurons to predict the probability of binary class. The neurons in output layer and dense layers use softmax function as the activation function, while all the neurons in other layers use ReLU as the activation function.
Performance Comparison on CRISPOR data

**Fig. 3.** ROC curves of two deep learning models and five current prediction methods under stratified 5-fold cross-validation on CRISPOR dataset.

**Fig. 4.** ROC curves of two deep learning models (i.e., FNN_3layer and DNN_std) and three traditional machine learning models including logistic regression (LR), random forest (RF) and gradient boosting trees (GBT). The ROC curve and AUC value of CFD score were regarded as the state-of-arts benchmark.
Performance Comparison on GUIDE-seq data

Fig. 5. ROC curves of deep learning models, CFD score and three traditional machine learning models on GUIDE-seq dataset.
Fig. 6. 15 off-targets with the highest score predicted by final convolutional neural network and CFD score respectively on GUIDE-seq dataset. The sgRNA-DNA sequence marked with star is the true off-target.
Pattern Recognition on CRISPR-Cas9 Off-targets [2]

- We presented that deep convolutional neural networks are able to accurately predict the off-targets of CRISPR-Cas9 gene editing. To our knowledge, this is the first time that deep neural networks are designed and implemented for off-target predictions.

- Our final convolutional neural network obtained the best performance on both CRISPOR dataset and GUIDE-seq dataset, outperforming the current state-of-art off-target prediction methods and three traditional machine learning algorithms including logistic regression, random forest, and gradient boosting trees. We discussed and attributed its performance successes to the neural network layer designs which are general enough to self-learn and capture sequence features by itself.

- We believe that such intelligent approaches can contribute to CRISPR-Cas9 off-target predictions or other similar problems in a rigorous manner.
Abstract

The off-target effects induced by guide RNAs in the CRISPR/Cas9 gene-editing system have raised substantial concerns in recent years. Many in silico predictive models have been developed for predicting the off-target activities; however, few are capable of predicting the off-target activities with insertions or deletions between guide RNA and target DNA sequence pair. In order to fill this gap, a recurrent convolutional network named CRISPR-Net is developed for scoring the gRNA-target pairs with mismatches and indels; and a machine-learning based model named CRISPR-Net-Aggregate is also developed for aggregating the scores as the consensus off-target score for each potential guide RNA. It is demonstrated that CRISPR-Net achieves competitive performance on CIRCLE-Seq and GUIDE-seq datasets with indels and mismatches, outperforming the state-of-the-art off-target prediction methods on two independent mismatch-only datasets. The CRISPR-Net-Aggregate also surpasses a competing method on the aggregation task. Moreover, a two-stage sensitivity analysis is introduced to visualize the CRISPR-Net prediction on the gRNA-target pair of interest, demonstrating how implicit knowledge encoded in CRISPR-Net contributes to the accurate off-target activity quantification. Finally, the source code is made available at the Code Ocean repository (https://codeocean.com/capsule/9553651/tree/v1).
Case for screening Early Cancers from Blood [4]

Case for screening Early Cancers from Blood [4]

Motivation

GLOBAL

BURDEN OF CANCER

Total population (2019)
7,676,965,500

Total # cancer cases (2018)
18,078,957

Total # cancer deaths (2018)
9,555,027

Premature deaths from NCDs (2016)
15,179,108

Cancer as % of NCD premature deaths (2016)
29.7%

Most common cancer cases (2018)

- Lung
- Breast
- Colorectum
- Prostate
- Stomach
- Liver
- Oesophagus
- Cervix uteri
- Thyroid
- Bladder

Incidence
Mortality

PAFs
(population attributable fractions)

- Tobacco (2017)*
- Alcohol (2016)*
- Infections (2012)*
- Obesity (2012)*
- UV (2012)*
- Occupational risk (2017)*

* PAF, cancer deaths
* PAF, cancer cases
* PAF, melanoma cases
Motivation

https://www3.ha.org.hk/cancereg/top10incidence.html
Background

Detection and localization of surgically resectable cancers with a multi-analyte blood test

Joshua D. Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova...

Science, 23 Feb 2018; Vol. 359, Issue 6378, pp. 926-930
DOI: 10.1126/science.aar3247
Background

Performance of CancerSEEK [1]
(A) ROC curve for CancerSEEK. The red point on the curve indicates the test’s average performance (62%) at >99% specificity. Error bars represent 95% confidence intervals for sensitivity and specificity at this particular point. The median performance among the eight cancer types assessed was 70%.
(B) Sensitivity of CancerSEEK by stage. Bars represent the median sensitivity of the eight cancer types, and error bars represent standard errors of the median.
(C) Sensitivity of CancerSEEK by tumor type. Error bars represent 95% confidence intervals.

Background

However, we note three limitations of CancerSEEK [1]:

1. Its front-line cancer detection component is based on logistic regression, whereby linear assumption on different markers is hardly realistic.

2. Its second-line cancer type localization component is based on random forest, a modeling known to be difficult for interpretations.

3. From the user perspective, its lack of public Web service also limits its potential impacts.

Data Collection

- We have collected the multianalyte blood test data from Cohen et al. (2018). Those data have been processed according to the supplementary guideline provided, resulting in two datasets.

  - The first dataset has 1,817 patient blood test records, which are designed and adopted to build models to detect cancers as the front-line detector in a binary manner (i.e., cancer or normal). Therefore, to be scalable and economical, it has the minimal number of input feature information involving eight circulating protein marker concentrations and one cell-free DNA mutation score (OmegaScore) as listed in the following table.

  - The second dataset has 626 patient blood test records, which are designed and adopted to build models to localize cancer types as the second-line diagnosis (i.e., Breast, Colorectum, Upper GI, Liver, Lung, Ovary, or Pancreas). Therefore, its input feature set covers the previous nine features and includes additional 31 protein markers and patient gender as listed in the later slides.
First Dataset

Feature Ranking for Binary Cancer Detection

<table>
<thead>
<tr>
<th>InfoG</th>
<th>Input Features</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6897</td>
<td>CA19-9 (U/ml)</td>
<td>Circulating Cancer Antigen 19-9 Concentration in U/ml</td>
</tr>
<tr>
<td>0.5119</td>
<td>CA-125 (U/ml)</td>
<td>Circulating Cancer Antigen 125 Concentration in U/ml</td>
</tr>
<tr>
<td>0.5001</td>
<td>HGF (pg/ml)</td>
<td>Circulating Hepatocyte Growth Factor Concentration in pg/ml</td>
</tr>
<tr>
<td>0.2779</td>
<td>OPN (pg/ml)</td>
<td>Circulating Osteopontin Concentration in pg/ml</td>
</tr>
<tr>
<td>0.2208</td>
<td>OmegaScore</td>
<td>Omega Score for Mutations in Circulating Cell-Free DNA</td>
</tr>
<tr>
<td>0.1826</td>
<td>Prolactin (pg/ml)</td>
<td>Circulating Prolactin Concentration in pg/ml</td>
</tr>
<tr>
<td>0.1518</td>
<td>CEA (pg/ml)</td>
<td>Circulating CarcinoEmbryonic Antigen Concentration in pg/ml</td>
</tr>
<tr>
<td>0.0989</td>
<td>Myeloperoxidase (ng/ml)</td>
<td>Circulating Myeloperoxidase Concentration in ng/ml</td>
</tr>
<tr>
<td>0.0916</td>
<td>TIMP-1 (pg/ml)</td>
<td>Circulating Tissue Inhibitor of MetalloProteinases 1 Concentration in pg/ml</td>
</tr>
</tbody>
</table>

Table S1: Feature List for Cancer Detection ranked by Information Gain (InfoG), related to Figure 1

First Dataset

Linear Discriminant Analysis
Dimensional Reduction

First Dataset

(a) t-distributed Stochastic Neighbor Embedding

(b) Principal Component Analysis

(c) Nonnegative Matrix Factorization

(d) Spectral Embedding
Proposed Approach – A1DE

First Dataset

WEKA Packages

IMPORTANT: make sure there are no old versions of Weka (<3.7.2) in your CLASSPATH before starting Weka

Installation of Packages

A GUI package manager is available from the “Tools” menu of the GUI Chooser

```
java -jar weka.jar
```

For a command line package manager type:

```
java weka.core.WekaPackageManager -h
```

Running packaged algorithms from the command line

```
java weka.Run [algorithm name]
```

Substring matching is also supported. E.g. try:

```
java weka.Run Bayes
```

Available Packages (206)

- AffectiveTweets
- Text classification
- Text Filters for Analyzing Sentiment and Emotions of Tweets
- Averaged N-Dependence Estimators (includes A1DE and A2DE)

We Have Tried All 206 Packages

It is the best for this task.
First Dataset

Proposed Approach – A1DE

Hence, given that each blood marker sample can be represented by a vector $x = \langle x_1, x_2, ..., x_n \rangle$ where $x_i$ is a marker attribute value, an A1DE model can be trained and assigned cancer detection label $y$ based on its posterior probability:

$$P(y|x) = \frac{P(y, x)}{P(x)} \propto P(y, x)$$  \hspace{1cm} (1)

By aggregating all possible 1-dependence classifiers, $P(y, x)$ can be written as:

$$P(y, x) = \frac{\sum_{1 \leq i \leq n \land F(x_i) \geq m} P(y, x_i) P(x | y, x_i)}{|\{1 \leq i \leq n \land F(x_i) \geq m\}|}$$  \hspace{1cm} (2)

Therefore, the label assignment (cancer detection label $y$) can be derived as follows:

$$\arg \max_y \sum_{1 \leq i \leq n \land F(x_i) \geq m} \hat{P}(y, x_i) \prod_{j=1}^{n} \hat{P}(x_j | y, x_i)$$  \hspace{1cm} (3)

where $\hat{P}$ denotes the probability estimate. From the above, we can see that, if none of the parent attributes $x_i$ have its $F(x_i)$ count greater than $m$, the A1DE is identical to a traditional NB classifier. On the other hand, the posterior of classes can be derived as follows:

$$\hat{P}(y|x) = \frac{\sum_{1 \leq i \leq n \land F(x_i) \geq m} \hat{P}(y, x_i) \prod_{j=1}^{n} \hat{P}(x_j | y, x_i)}{\sum_{y' \in Y} \sum_{1 \leq i \leq n \land F(x_i) \geq m} \hat{P}(y', x_i) \prod_{j=1}^{n} \hat{P}(x_j | y', x_i)}$$  \hspace{1cm} (4)

where the above formula is derived from the Bayes rule $P(y|x) = P(y, x)/P(x)$. The

*with Minimum Description Length (MDL) feature discretization
ROC Curves for Binary Cancer Detection

Sensitivities (Recalls) for Binary Cancer Detection

Figure 2. Proportion of Detected Cancers with Different Stages at the 99% Specificity Level. Each color represents a method, and the horizontal axis has been ordered by cancer stages. Each bar represents the median sensitivity of each method on each cancer stage with standard errors.

Figure 3. Detected Proportions of Different Cancer Types at the 99% Specificity Level
Different colors represent different methods. The horizontal axis is ordered by cancer types. Each bar represents the sensitivity of each method on each cancer type with 95% confidence intervals.

<table>
<thead>
<tr>
<th>InfoG</th>
<th>Input Features</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0389</td>
<td>TGFα (pg/ml)</td>
<td>Circulating Transforming Growth Factor Alpha Concentration in pg/ml</td>
</tr>
<tr>
<td>0.8301</td>
<td>HE4 (pg/ml)</td>
<td>Circulating Human Epididymis Protein 4 Concentration in pg/ml</td>
</tr>
<tr>
<td>0.6135</td>
<td>sFlt (pg/ml)</td>
<td>Circulating soluble Fas Cell Surface Death Receptor Concentration in pg/ml</td>
</tr>
<tr>
<td>0.5272</td>
<td>Thrombospondin-2 (pg/ml)</td>
<td>Circulating Thrombospondin-2 Concentration in pg/ml</td>
</tr>
<tr>
<td>0.5073</td>
<td>APP (pg/ml)</td>
<td>Circulating Amyloid Precursor Protein Concentration in pg/ml</td>
</tr>
<tr>
<td>0.3759</td>
<td>G-CSF (pg/ml)</td>
<td>Circulating Granulocyte-Colony Stimulating Factor Concentration in pg/ml</td>
</tr>
<tr>
<td>0.3836</td>
<td>IL-6 (pg/ml)</td>
<td>Circulating Interleukin-6 Concentration in pg/ml</td>
</tr>
<tr>
<td>0.3597</td>
<td>CA-125 (U/ml)</td>
<td>Circulating Cancer Antigen 125 Concentration in U/ml</td>
</tr>
<tr>
<td>0.2568</td>
<td>SAA</td>
<td>Serum Amyloid A Concentration in mg/ml</td>
</tr>
<tr>
<td>0.2312</td>
<td>HGF</td>
<td>Hepatocyte Growth Factor Concentration in pg/ml</td>
</tr>
<tr>
<td>0.2289</td>
<td>CD44 (ng/ml)</td>
<td>Circulating CD44 Concentration in pg/ml</td>
</tr>
<tr>
<td>0.1853</td>
<td>CA125-9 (U/ml)</td>
<td>Circulating Cancer Antigen 125-9 Concentration in U/ml</td>
</tr>
<tr>
<td>0.1805</td>
<td>IL-8 (pg/ml)</td>
<td>Circulating Interleukin-8 Concentration in pg/ml</td>
</tr>
<tr>
<td>0.1616</td>
<td>CA-15-3 (U/ml)</td>
<td>Circulating Cancer Antigen 15-3 Concentration in U/ml</td>
</tr>
<tr>
<td>0.1448</td>
<td>RBP</td>
<td>Retinol Binding Protein Concentration in ng/ml</td>
</tr>
<tr>
<td>0.1431</td>
<td>GDF15 (ng/ml)</td>
<td>Circulating Growth Differentiation Factor 15 Concentration in ng/ml</td>
</tr>
<tr>
<td>0.1340</td>
<td>Leptin (pg/ml)</td>
<td>Circulating Leptin Concentration in pg/ml</td>
</tr>
<tr>
<td>0.1271</td>
<td>Myeloperoxidase (ug/ml)</td>
<td>Circulating Myeloperoxidase Concentration in ug/ml</td>
</tr>
<tr>
<td>0.1125</td>
<td>Kallikrein-6 (pg/ml)</td>
<td>Circulating Kallikrein-6 Concentration in pg/ml</td>
</tr>
<tr>
<td>0.1173</td>
<td>TMP-1 (ng/ml)</td>
<td>Circulating Tissue Inhibitor of Metalloproteinases 1 Concentration in ng/ml</td>
</tr>
<tr>
<td>0.1122</td>
<td>Midkine (pg/ml)</td>
<td>Circulating Midkine Concentration in pg/ml</td>
</tr>
<tr>
<td>0.1080</td>
<td>Proctolin (ng/ml)</td>
<td>Circulating Proctolin Concentration in ng/ml</td>
</tr>
<tr>
<td>0.1058</td>
<td>Mesoethin (ng/ml)</td>
<td>Circulating Mesoethin Concentration in ng/ml</td>
</tr>
<tr>
<td>0.1010</td>
<td>Galectin-3 (mg/ml)</td>
<td>Circulating Galectin-3 Concentration in mg/ml</td>
</tr>
<tr>
<td>0.1096</td>
<td>OPN (pg/ml)</td>
<td>Circulating Osteopontin Concentration in pg/ml</td>
</tr>
<tr>
<td>0.0958</td>
<td>NSE (ng/ml)</td>
<td>Circulating Neuron-Specific Enolase Concentration in ng/ml</td>
</tr>
<tr>
<td>0.0901</td>
<td>sEGRF (pg/ml)</td>
<td>Circulating soluble Epidermal Growth Factor Receptor Concentration in pg/ml</td>
</tr>
<tr>
<td>0.0911</td>
<td>CEA (pg/ml)</td>
<td>Circulating Carcinoembryonic Antigen Concentration in pg/ml</td>
</tr>
<tr>
<td>0.0830</td>
<td>AXI (pp/ml)</td>
<td>Circulating AXI Receptor Tyrosine Kinase Concentration in pp/ml</td>
</tr>
<tr>
<td>0.0771</td>
<td>APECAM-1 (pg/ml)</td>
<td>Circulating soluble Platelet and Endothelial Cell Adhesion Molecule 1 Concentration in pg/ml</td>
</tr>
<tr>
<td>0.0637</td>
<td>SHBG (nM)</td>
<td>Circulating Sex Hormone-Binding Globulin Concentration in nM</td>
</tr>
<tr>
<td>0.0635</td>
<td>OmegaScore</td>
<td>Omega Score for Mutations in Circulating Cell-Free DNA</td>
</tr>
</tbody>
</table>

Table 1: Feature List for Cancer Type Localization ranked by Information Gain (InfoG)
Figure 4. Localized Proportions of Different Cancer Types using the Top One Prediction Approach

Different colors represent different methods. The horizontal axis is ordered by cancer types. Each bar represents the sensitivity of each method on each cancer type with 95% confidence intervals.

Figure 5. Feature Importance Heatmap for Cancer Type Localization under One-Class-versus-Others Setting

The feature rankings are measured based on the Learning Vector Quantization (LVQ) building under Python caret package (Bischl et al., 2016). Ten-fold cross-validations are run to compute the feature importance values. After that, the function “heatmap.2” in R language is adopted with the default setting to cluster and visualize the feature importance values. Further details can be found in Figure S14.
http://cancer.cs.cityu.edu.hk/

CANCER DETECTION

Do You Want to Know Your Probability to Get a Cancer? Just Have a Try!

Step 1. Blood Test Marker Values

<table>
<thead>
<tr>
<th>Marker</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>OmegaScore</td>
<td>2.96</td>
</tr>
<tr>
<td>CA-125 (U/ml)</td>
<td>5.09</td>
</tr>
<tr>
<td>CEA (pg/ml)</td>
<td>540.1</td>
</tr>
<tr>
<td>CA19-9 (U/ml)</td>
<td>16.45</td>
</tr>
<tr>
<td>Prolactin (pg/ml)</td>
<td>11.506.6</td>
</tr>
<tr>
<td>HGF (pg/ml)</td>
<td>377.26</td>
</tr>
<tr>
<td>OPN (pg/ml)</td>
<td>56516.58</td>
</tr>
<tr>
<td>Myeloperoxidase (ng/ml)</td>
<td>14.22</td>
</tr>
<tr>
<td>TIMP-1 (pg/ml)</td>
<td>56428.71</td>
</tr>
</tbody>
</table>

Step 2. Model Selection

| Model | CancerAIDE |

The probability to get a cancer is:
Conclusion

Summary

We explore different supervised learning approaches for multiple cancer type detection and observe significant improvements; for instance, one of our approaches (i.e., CancerA1DE) can double the existing sensitivity from 38% to 77% for the earliest cancer detection (i.e., Stage I) at the 99% specificity level. For Stage II, it can even reach up to about 90% across multiple cancer types. In addition, CancerA1DE can also double the existing sensitivity from 30% to 70% for detecting breast cancers at the 99% specificity level. Data and model analysis are conducted to reveal the underlying reasons. A website is built at http://cancer.cs.cityu.edu.hk/.

Summary - Case Studies

- Case for synthesizing Heterodimeric DNA Motifs [1]
- Case for predicting CRISPR-Cas9 Off-targets [2,3]
- Case for screening Early Cancers from Blood [4]

References


Data Application Case Studies for COVID-19

(30 minutes)
Outline

• Data Application Cases at Molecular Biology Level
  – 1D Sequence Data from COVID-19
  – 3D Molecular Structure Data from COVID-19
  – 3D Molecular Vaccine Design for COVID-19 Treatment

• Data Application Cases at Public Health / Clinical Level
  – Infection Case Data on COVID-19
  – Integrated Data Analysis on COVID-19

• Data Application Cases at Text Level
  – Academic Paper Analysis on COVID-19
  – Social Media Text Analysis on COVID-19
Molecular Biology Level
COVID-19 at Molecular Biology Level

https://www.nature.com/collections/hajgidghjb
COVID-19 at Molecular Biology Level

COVID-19 at Molecular Biology Level

COVID-19 at Molecular Biology Level

COVID-19 at Molecular Biology Level

GISAID

Genomic epidemiology of hCoV-19

Phylogeny

GISAID Clade:

GR

Ghana/35394_549/2020

No nucleotide mutations

Divergence: 8

Date: 2020-05-24

GISAID Clade: GR

Author: Ngpi et al

GISAID EPI ISU: s15064

Click on site to display more info

Geography

© 2005 – 2020 Freunde von GISAID e.V. | Imprint / Privacy | Terms of Use | Contact | a | twitter

https://www.gisaid.org/epiflu-applications/phylo-dynamics/
COVID-19 at Molecular Biology Level

Fig. 1. Structure of SARS-CoV-2. (A) Illustration of the SARS-CoV-2 virion created at the Centers for Disease Control and Prevention (CDC). The spikes on the outer edge of the virus particles look like a crown, giving the disease its characteristic name. (B) Schematic representation of the structure of SARS-CoV-2. It has four structural proteins, S (spike), E (envelope), M (membrane), and N (nucleocapsid) proteins; the N protein holds the RNA genome, and the S, E, and M proteins together create the viral envelope. (C) An electron microscopic image of a thin section of SARS-CoV-2 within the cytoplasm of an infected cell, showing the spherical particles and cross-sections through the viral nucleocapsid (Sohrabi et al., 2020).

Fig. 2. The life cycle of SARS-CoV-2 in human lung cells. Coronavirus is most often transmitted by droplets while sneezing and coughing and its journey begins in the first days after infiltration from the upper respiratory tract. The spike proteins of SARS-CoV-2 binds to ACE2 receptors. The virion then releases RNA genome into the cell and translation of structural and non-structural proteins follows. ORF1a and ORF1ab are translated to produce pp1a and pp1ab polyproteins, which are cleaved by the proteases that are encoded by ORF1a to yield non-structural proteins. This is followed by assembly and budding into the lumen of the ERGIC. Virions are then released from the infected cell through exocytosis (Adnan Shereen et al., 2020). NSP, non-structural proteins; ACE2, Angiotensin-Converting Enzyme 2; Rough ER, Rough Endoplasmic Reticulum; ERGIC, Endoplasmic Reticulum Golgi Intermediate Compartment.
COVID-19 at Molecular Biology Level

https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52&feature=true
COVID-19 at Molecular Biology Level

7KFI
SARS-CoV-2 Main protease immature form - apo structure
To be published
- Released: 2020-10-28
- Method: X-RAY DIFFRACTION 1.6 Å
- Organisms: Severe acute respiratory syndrome coronavirus 2
- Macromolecule: Main protease (protein)
- Unique Ligands: DMS, PEG

7KG3
Crystal structure of CoV-2 Nsp3 Macromain
To be published
- Released: 2020-10-28
- Method: X-RAY DIFFRACTION 1.45 Å
- Organisms: Severe acute respiratory syndrome coronavirus 2
- Macromolecule: Non-structural protein 3 (protein)
- Unique Ligands: GOL, MES, MLI, 504

7KHP
Acyl-enzyme intermediate structure of SARS-CoV-2 Mpro in complex with its C-terminal autoprocessing sequence.
To be published
- Released: 2020-10-28
- Method: X-RAY DIFFRACTION 1.95 Å
- Organisms: Severe acute respiratory syndrome coronavirus 2
- Macromolecule: 3C-like protease (protein)
COVID-19 at Molecular Biology Level

Potential Vaccine Targets
COVID-19 at Molecular Biology Level

### Types of coronavirus vaccine approaches

<table>
<thead>
<tr>
<th>Types of vaccines</th>
<th>DNA and RNA</th>
<th>Live attenuated</th>
<th>Inactivated</th>
<th>Subunit</th>
<th>Viral vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>How it works</td>
<td>This vaccine uses DNA or RNA molecules to teach the immune system to target key viral proteins.</td>
<td>This is a weakened version of the actual virus.</td>
<td>An inactivated vaccine uses the whole virus after it has been killed with heat or chemicals.</td>
<td>This vaccine uses a piece of a virus' surface to boost your immune system on a single target.</td>
<td>This approach takes a harmless virus and uses it to deliver viral genes to build immunity.</td>
</tr>
<tr>
<td>Advantages</td>
<td>Easy and quick to design.</td>
<td>Stimulates a robust immune response without causing serious disease.</td>
<td>Safe because the virus is already dead and is easy to make.</td>
<td>Focuses the immune response on the most important part of the virus for protection and cannot cause infection.</td>
<td>Live viruses tend to elicit stronger immune responses than dead viruses or subunit vaccines.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>May not be safe for those with compromised immune systems.</td>
<td>Not as effective as a live virus. Some previous inactivated vaccines have made the disease worse; safety for the novel coronavirus needs to be shown in clinical trials.</td>
<td>May not stimulate a strong response; other chemicals may need to be added to boost long-term immunity.</td>
<td>Important to pick a viral vector that is truly safe. An immune response to the viral vector could make the vaccine less effective.</td>
<td></td>
</tr>
<tr>
<td>Existing examples</td>
<td>None</td>
<td>Measles, Mumps and Rubella</td>
<td>Pertussis, Hepatitis B, Human papillomavirus (HPV)</td>
<td>Ebola, Veterinary medicine</td>
<td>Examples: Seasonal influenza vaccine COVID-19, INO-4800 in phase 1 clinical trials</td>
</tr>
</tbody>
</table>

### Classical platforms

- **Whole-inactivated virus**
  - Example: Polio vaccine COVID-19
  - Phase 1/2 clinical trials

- **Live-attenuated virus**
  - Example: MMR vaccine COVID-19
  - Not currently licensed

### Next-generation platforms

- **Viral vector**
  - Example: SARS-CoV-2
  - Not currently licensed

- **DNA**
  - Example: Not currently licensed COVID-19

- **RNA**
  - Example: Not currently licensed COVID-19

- **Antigen-presenting cells**
  - Example: Not currently licensed COVID-19

### Group testing for COVID-19

- Moderna (DNA)
- Inovio (DNA)
- Codagenix (vaccinomimetics)
- Synovax (SNOPharm)
- Novavax (AdapVac)
- University of Oxford & AstraZeneca
- Cardiologic Biologics
- Johnson & Johnson

---

Sources: CDC, NIAID, FDA

Michelle Guerrero and Jonathan Wode


[https://www.nature.com/articles/s41563-020-0746-0/figures/1](https://www.nature.com/articles/s41563-020-0746-0/figures/1)
Public Health Level
Globally, as of 1:49pm CET, 1 November 2020, there have been 45,942,902 confirmed cases of COVID-19, including 1,192,644 deaths, reported to WHO.
### COVID-19 at Public Health Level

[https://covid19.who.int/](https://covid19.who.int/)

#### Situation by Country, Territory & Area

<table>
<thead>
<tr>
<th>Name</th>
<th>Cases - cumulative total</th>
<th>Cases - newly reported in last 24 hours</th>
<th>Deaths - cumulative total</th>
<th>Deaths - newly reported in last 24 hours</th>
<th>Transmission Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>46,942,902</td>
<td>473,311</td>
<td>1,192,644</td>
<td>6,578</td>
<td></td>
</tr>
<tr>
<td>United States of America</td>
<td>8,952,086</td>
<td>99,356</td>
<td>228,185</td>
<td>1,007</td>
<td>Community transmission</td>
</tr>
<tr>
<td>India</td>
<td>8,184,082</td>
<td>46,963</td>
<td>122,111</td>
<td>470</td>
<td>Clusters of cases</td>
</tr>
<tr>
<td>Brazil</td>
<td>5,516,658</td>
<td>22,282</td>
<td>159,477</td>
<td>508</td>
<td>Community transmission</td>
</tr>
<tr>
<td>Russian Federation</td>
<td>1,930,781</td>
<td>18,665</td>
<td>28,235</td>
<td>245</td>
<td>Clusters of cases</td>
</tr>
<tr>
<td>France</td>
<td>1,331,808</td>
<td>32,530</td>
<td>36,473</td>
<td>223</td>
<td>Community transmission</td>
</tr>
<tr>
<td>Spain</td>
<td>1,185,078</td>
<td>0</td>
<td>35,878</td>
<td>0</td>
<td>Community transmission</td>
</tr>
<tr>
<td>Argentina</td>
<td>1,157,179</td>
<td>13,379</td>
<td>30,792</td>
<td>350</td>
<td>Community transmission</td>
</tr>
<tr>
<td>Colombia</td>
<td>1,063,151</td>
<td>10,029</td>
<td>31,135</td>
<td>209</td>
<td>Community transmission</td>
</tr>
<tr>
<td>The United Kingdom</td>
<td>1,011,664</td>
<td>21,915</td>
<td>46,555</td>
<td>326</td>
<td>Community transmission</td>
</tr>
<tr>
<td>Mexico</td>
<td>918,811</td>
<td>6,000</td>
<td>91,289</td>
<td>516</td>
<td>Community transmission</td>
</tr>
</tbody>
</table>
COVID-19 at Public Health Level

https://coronavirus.jhu.edu/map.html
COVID-19 at Public Health Level

https://coronavirus.jhu.edu/covid-19-daily-video
Coronavirus Pandemic (COVID-19) – the data

Research and data: Hannah Ritchie, Esteban Ortiz-Ospina, Diana Buitkei, Edouard Mathieu, Joe Hasell, Bobble Maconald, Charile Giattino, and Max Roser
Web development: Breck Yunits, Ernst van Woerden, Daniel Gavrilov, Matihieu Bergel, Shahid Ahmad, and Jason Crawford

The data on the coronavirus pandemic is updated daily. Last update: November 1, 2020 (10:00, London time).

Our work belongs to everyone

Download the complete Our World In Data COVID-19 dataset

All our code is open source
All our research and visualizations are free for everyone to use for all purposes

The purpose of this page here is simply to lists all our visualizations on the pandemic.
COVID-19 at Public Health Level


Data on COVID-19 (coronavirus) by Our World in Data

Our complete COVID-19 dataset is a collection of the COVID-19 data maintained by Our World in Data. It is updated daily and includes data on confirmed cases, deaths, hospitalizations, and testing, as well as other variables of potential interest.

Download our complete COVID-19 dataset: CSV | XLSX | JSON

We will continue to publish up-to-date data on confirmed cases, deaths, hospitalizations, and testing, throughout the duration of the COVID-19 pandemic.

Our data sources

- **Confirmed cases and deaths**: our data comes from the European Centre for Disease Prevention and Control (ECDC). We discuss how and when the ECDC collects and publishes this data here. The cases & deaths dataset is updated daily. Note: the number of cases or deaths reported by any institution—including the ECDC, the WHO, Johns Hopkins and others—on a given day does not necessarily represent the actual number on that date. This is because of the long reporting chain that exists between a new case/death and its inclusion in statistics. This also means that negative values in cases and deaths can sometimes appear when a country sends a correction to the ECDC, because it had previously overestimated the number of cases/deaths. Alternatively, large changes can sometimes (although rarely) be made to a country’s entire time series of the ECDC decides (and has access to the necessary data) to correct values retrospectively.

- **Hospitalizations and intensive care unit (ICU) admissions**: our data comes from the European Centre for Disease Prevention and Control (ECDC), who provide these statistics only for a select number of European countries, and update it on a weekly basis. Unfortunately, we are unable to provide data on hospitalizations for other countries: there is currently no global, aggregated database on COVID-19 hospitalization, and our team at Our World in Data do not have the capacity to build such a dataset.

- **Testing for COVID-19**: this data is collected by the Our World in Data team from official reports; you can find further details in our post on COVID-19 testing, including our checklist of questions to understand testing data, information on geographical and temporal coverage, and detailed country-by-country source information. The testing dataset is updated around twice a week.

- **Other variables**: this data is collected from a variety of sources (United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government, etc.). More information is available in our codebook.

The complete Our World in Data COVID-19 dataset

Our complete COVID-19 dataset is available in CSV, XLSX, and JSON formats, and includes all of our historical data on the pandemic up to the date of publication.

The CSV and XLSX files follow a format of 1 row per location and date. The JSON version is split by country ISO code, with static variables and an array of daily records.
COVID-19 at Public Health Level

https://ourworldindata.org/mortality-risk-covid

The case fatality rate

Case fatality rate of the ongoing COVID-19 pandemic

The Case Fatality Rate (CFR) is the ratio between confirmed deaths and confirmed cases. During an outbreak of a pandemic the CFR is a poor measure of the mortality risk of the disease. We explain this in detail at OurWorldInData.org/Coronavirus.
COVID-19 at Public Health Level

Text Level
COVID-19 at Text Level

• Text Mining on Academic Papers
  • https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

• Over 200,000 academic papers have been collected.

• Total Data Size: 22 GB  (as of 2nd Nov 2020)
COVID-19 at Text Level

• Text Mining on Academic Papers

• https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Given a query:
Does hypertension increase the risks associated with Covid-19?

“the most common pre-diagnosis comorbidity in 24 patients (25.9%) was hypertension...”

“Hypertension increases the rate of Covid-19 relapse...”

Retrieval relevant articles
Extract supporting evidence

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7251955/
COVID-19 at Text Level

• Text Mining on Tweets

https://ieeecomputing.org/open-access/coronavirus-covid-19-tweets-dataset
COVID-19 at Text Level

• — Number of tweets : 694,080,653 tweets
• — Coverage : Global
• — Language : English (EN)
• — Geo-tagged Version: Coronavirus (COVID-19) Geo-tagged Tweets Dataset (GeoCOV19Tweets Dataset)
• — Dataset updates : Everyday
• — Usage policy : As per Twitter's Developer Policy


https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset
COVID-19 at Text Level

COVID-19 at Text Level


Figure 1. Collected tweets, by language, as of 13 March 2020.
COVID-19 at Text Level

COVID-19 at Text Level

- Google Search Keyword Analysis

The COVID-19 Search Trends symptoms dataset shows aggregated, anonymized trends in Google searches for more than 400 health symptoms, signs, and conditions, such as cough, fever and difficulty breathing. The dataset provides a time series for each region showing the relative volume of searches for each symptom.
Google Search Trend from 2017 to 2020