KTH
VETENSKAP
OCH KONST

AI for Research in Biology

EQ2461 seminar course — Feb 5th, 2025

Antoine Honoré (Postdoc)

Advisor: Prof. Ming Xiao

Outline

Introduction

Research in Biology

Cell Functions

Variant effect prediction

Proteins

Evolutionary constrain assumption

Data

Predictors

More predictors, based on Transformers

Future perspectives

Conclusions

Introduction: Research in Biology

- Study of living organisms and their interactions with the environment
- Aim: understanding of the principles and processes that govern life at different levels, e.g. molecular, or generally at organism level:
  - (Molecular biology) What is the molecular basis for cell function ?
  - (Physiology) How did specific chemical/physical functions emerge in a biological system ?
- Research in biology can be either
  - Basic: Discovering the determinants of functions in living organisms
  - Applied: Leveraging known processes to achieve something

Introduction: Research in Biology

- Study of living organisms and their interactions with the environment
- Aim: understanding of the principles and processes that govern life at different levels, e.g. molecular, or generally at organism level:
  - (Molecular biology) What is the molecular basis for cell function ?
  - (Physiology) How did specific chemical/physical functions emerge in a biological system ?
- Research in biology can be either
  - Basic: Discovering the determinants of functions in living organisms
  - Applied: Leveraging known processes to achieve something

Introduction: Research in Biology

- Study of living organisms and their interactions with the environment
- Aim: understanding of the principles and processes that govern life at different levels, e.g. molecular, or generally at organism level:
  - (Molecular biology) What is the molecular basis for cell function ?
  - (Physiology) How did specific chemical/physical functions emerge in a biological system ?
- Research in biology can be either
  - Basic: Discovering the determinants of functions in living organisms
  - Applied: Leveraging known processes to achieve something

Introduction: Research in Biology

- Study of living organisms and their interactions with the environment
- Aim: understanding of the principles and processes that govern life at different levels, e.g. molecular, or generally at organism level:
  - (Molecular biology) What is the molecular basis for cell function ?
  - (Physiology) How did specific chemical/physical functions emerge in a biological system ?
- Research in biology can be either
  - Basic: Discovering the determinants of functions in living organisms
  - Applied: Leveraging known processes to achieve something

Introduction: Research in Biology

- Study of living organisms and their interactions with the environment
- Aim: understanding of the principles and processes that govern life at different levels, e.g. molecular, or generally at organism level:
  - (Molecular biology) What is the molecular basis for cell function ?
  - (Physiology) How did specific chemical/physical functions emerge in a biological system ?
- Research in biology can be either
  - Basic: Discovering the determinants of functions in living organisms
  - Applied: Leveraging known processes to achieve something

Introduction: Research in Biology

- Study of living organisms and their interactions with the environment
- Aim: understanding of the principles and processes that govern life at different levels, e.g. molecular, or generally at organism level:
  - (Molecular biology) What is the molecular basis for cell function ?
  - (Physiology) How did specific chemical/physical functions emerge in a biological system ?
- Research in biology can be either
  - Basic: Discovering the determinants of functions in living organisms
  - Applied: Leveraging known processes to achieve something

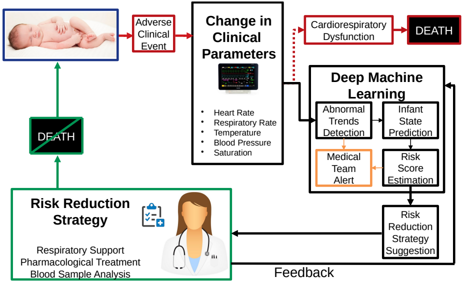
Introduction: Research in Biology

- Study of living organisms and their interactions with the environment
- Aim: understanding of the principles and processes that govern life at different levels, e.g. molecular, or generally at organism level:
  - (Molecular biology) What is the molecular basis for cell function ?
  - (Physiology) How did specific chemical/physical functions emerge in a biological system ?
- Research in biology can be either
  - Basic: Discovering the determinants of functions in living organisms
  - Applied: Leveraging known processes to achieve something

Introduction: Research in Biology

Human physiology

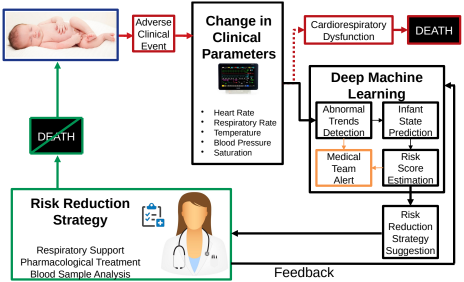
- Basic research: How do infections affect the internal stability of the body ? [1]
- Applied research: Is it possible to leverage the observation of physiological patterns to predict the onset of an infection ? [2]



Introduction: Research in Biology

Human physiology

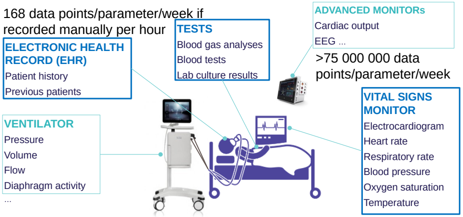
- Basic research: How do infections affect the internal stability of the body ? [1]
- Applied research: Is it possible to leverage the observation of physiological patterns to predict the onset of an infection ? [2]



Introduction: Research in Biology

Human physiology

- Research: conclusions are limited in part by the ability to collect sufficiently large datasets to account for all sources of variance in the problem
- Recent techniques allow to collect (and store) large databases, e.g. :
  - Human physiology : observed with a multitude of high-throughput sensors
  - Cell biology : variant-cell function maps are sampled abundantly & efficiently with deep mutational scanning experiments



168 data points/parameter/week if recorded manually per hour

ELECTRONIC HEALTH RECORD (EHR)
Patient history
Previous patients

TESTS
Blood gas analyses
Blood tests
Lab culture results

ADVANCED MONITORs
Cardiac output
EEG ...

>75 000 000 data points/parameter/week

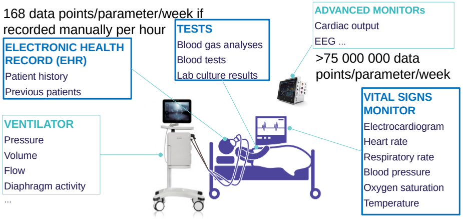
VENTILATOR
Pressure
Volume
Flow
Diaphragm activity
...

VITAL SIGNS MONITOR
Electrocardiogram
Heart rate
Respiratory rate
Blood pressure
Oxygen saturation
Temperature

Introduction: Research in Biology

Human physiology

- Research: conclusions are limited in part by the ability to collect sufficiently large datasets to account for all sources of variance in the problem
- Recent techniques allow to collect (and store) large databases, e.g. :
  - Human physiology : observed with a multitude of high-throughput sensors
  - Cell biology : variant-cell function maps are sampled abundantly & efficiently with deep mutational scanning experiments



168 data points/parameter/week if recorded manually per hour

ELECTRONIC HEALTH RECORD (EHR)
Patient history
Previous patients

TESTS
Blood gas analyses
Blood tests
Lab culture results

ADVANCED MONITORs
Cardiac output
EEG ...

>75 000 000 data points/parameter/week

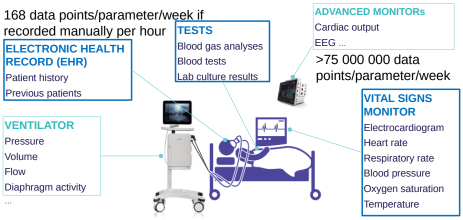
VITAL SIGNS MONITOR
Electrocardiogram
Heart rate
Respiratory rate
Blood pressure
Oxygen saturation
Temperature

VENTILATOR
Pressure
Volume
Flow
Diaphragm activity
...

Introduction: Research in Biology

Human physiology

- Research: conclusions are limited in part by the ability to collect sufficiently large datasets to account for all sources of variance in the problem
- Recent techniques allow to collect (and store) large databases, e.g. :
  - Human physiology : observed with a multitude of high-throughput sensors
  - Cell biology : variant-cell function maps are sampled abundantly & efficiently with deep mutational scanning experiments



168 data points/parameter/week if recorded manually per hour

ELECTRONIC HEALTH RECORD (EHR)
Patient history
Previous patients

TESTS
Blood gas analyses
Blood tests
Lab culture results

ADVANCED MONITORs
Cardiac output
EEG ...

>75 000 000 data points/parameter/week

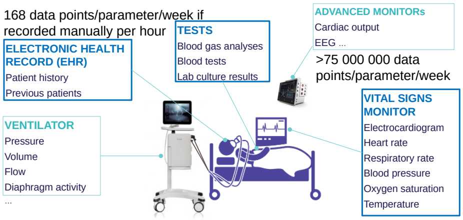
VENTILATOR
Pressure
Volume
Flow
Diaphragm activity
...

VITAL SIGNS MONITOR
Electrocardiogram
Heart rate
Respiratory rate
Blood pressure
Oxygen saturation
Temperature

Introduction: Research in Biology

Human physiology

- Research: conclusions are limited in part by the ability to collect sufficiently large datasets to account for all sources of variance in the problem
- Recent techniques allow to collect (and store) large databases, e.g. :
  - Human physiology : observed with a multitude of high-throughput sensors
  - Cell biology : variant-cell function maps are sampled abundantly & efficiently with deep mutational scanning experiments



168 data points/parameter/week if recorded manually per hour

ELECTRONIC HEALTH RECORD (EHR)
Patient history
Previous patients

TESTS
Blood gas analyses
Blood tests
Lab culture results

ADVANCED MONITORs
Cardiac output
EEG ...

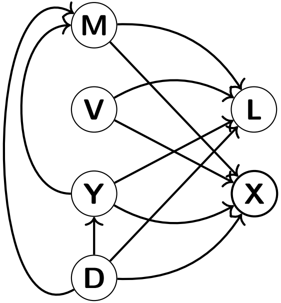
>75 000 000 data points/parameter/week

VENTILATOR
Pressure
Volume
Flow
Diaphragm activity
...

VITAL SIGNS MONITOR
Electrocardiogram
Heart rate
Respiratory rate
Blood pressure
Oxygen saturation
Temperature

Introduction: Research in Biology

- AI systems (= models learnt on data) can be used to model large datasets with many interacting variables
- This allows to make inference in networks such as:



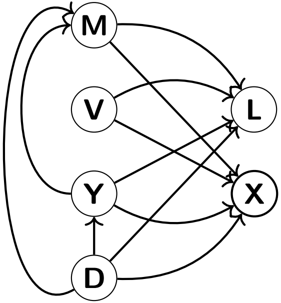
Legend

- **M**: Medication
- **L**: Laboratory values (Blood gas analysis)
- **D**: Demographics (including age)
- **V**: Mechanical Ventilation
- **Y**: Clinical condition
- **X**: Vital signs
- **G**: Genetic background

- Instead of being limited to simpler networks such as : **Y** → **X**.

Introduction: Research in Biology

- AI systems (= models learnt on data) can be used to model large datasets with many interacting variables
- This allows to make inference in networks such as:



Legend

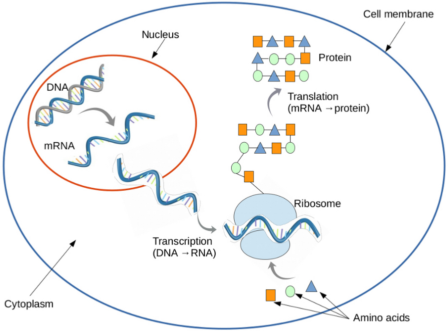
- **M**: Medication
- **L**: Laboratory values (Blood gas analysis)
- **D**: Demographics (including age)
- **V**: Mechanical Ventilation
- **Y**: Clinical condition
- **X**: Vital signs
- **G**: Genetic background

- Instead of being limited to simpler networks such as : **Y** → **X**.

Introduction: Research in Biology

Cell Biology

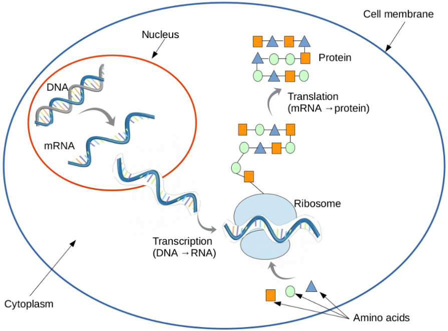
- Basic research: How do variations (mutations) in the genome affect the function of certain proteins in cells ? [3]
- Applied research: From a variant → cell function prediction tool, can we predict the effect of a chemotherapy drug on a cancer patient ? [4]



Introduction: Research in Biology

Cell Biology

- Basic research: How do variations (mutations) in the genome affect the function of certain proteins in cells ? [3]
- Applied research: From a variant → cell function prediction tool, can we predict the effect of a chemotherapy drug on a cancer patient ? [4]



Introduction: Cell Functions

Introduction: Cell Functions

- Random genetic variants occur when cells duplicate.
- Variants may or may not alter the function of the cell.
- The functions in cells are determined by proteins:

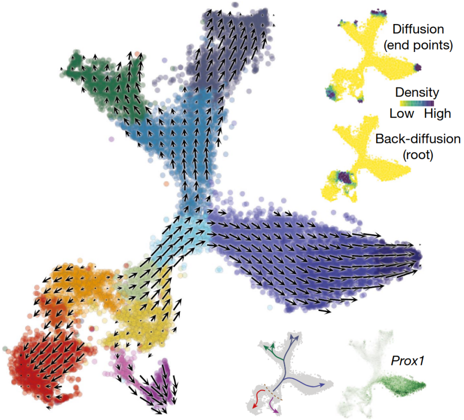


Figure: Each point is related to the 2d-projection of the proteome of a cell for an organism. Each color represents a group of cells with similar function.

Introduction: Cell Functions

- Random genetic variants occur when cells duplicate.
- Variants may or may not alter the function of the cell.
- The functions in cells are determined by proteins:

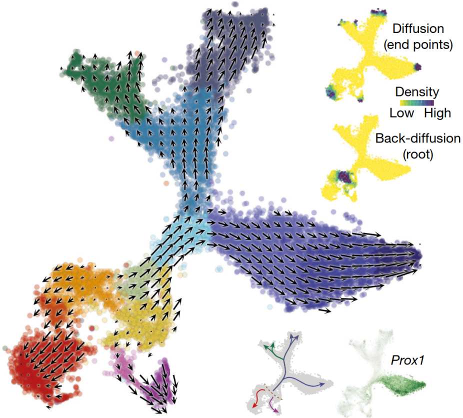
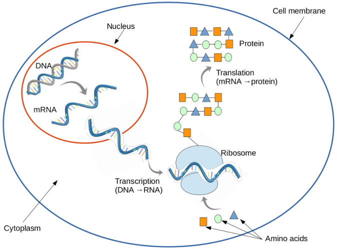


Figure: Each point is related to the 2d-projection of the proteome of a cell for an organism. Each color represents a group of cells with similar function.

Introduction: Cell Functions

- Random genetic variants occur when cells duplicate.
- Variants may or may not alter the function of the cell.
- The functions in cells are determined by proteins:



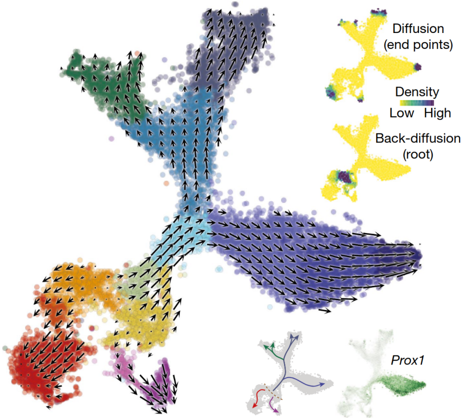


Figure: Each point is related to the 2d-projection of the proteome of a cell for an organism. Each color represents a group of cells with similar function.

Variant effect prediction: Proteins

Variant effect prediction: Proteins

In cells

- Sequences of elementary molecules: amino-acids (AA)
  - There are $d = 20$ amino-acids (in humans), each with p specific bio-physical properties, e.g. mass, volume.
  - The length of a protein is typically $L = 300$ AAs (can vary: 10-2000)
- Determined by their sequence of amino-acid:

  `PPGPTPLPVIGNILQIGIK...`
- Fold into different 3D-structures because of the bio-physical properties of AAs.

Variant effect prediction: Proteins

In cells

- Sequences of elementary molecules: amino-acids (AA)
  - There are $d = 20$ amino-acids (in humans), each with p specific bio-physical properties, e.g. mass, volume.
  - The length of a protein is typically $L = 300$ AAs (can vary: 10-2000)
- Determined by their sequence of amino-acid:

```
PPGPTPLPVIGNILQIGIK...
```

- Fold into different 3D-structures because of the bio-physical properties of AAs.

Variant effect prediction: Proteins

In cells

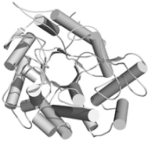
- Sequences of elementary molecules: amino-acids (AA)
  - There are $d = 20$ amino-acids (in humans), each with p specific bio-physical properties, e.g. mass, volume.
  - The length of a protein is typically $L = 300$ AAs (can vary: 10-2000)
- Determined by their sequence of amino-acid:

  `PPGPTPLPVIGNILQIGIK...`

- Fold into different 3D-structures because of the bio-physical properties of AAs.

avGFP

Bgl3



GB1

Variant effect prediction: Proteins

Encoding

Proteins can be encoded as

① sequences of one-hot vectors in $\{0,1\}^d$, i.e. $\underline{\mathbf{x}} \in \{0,1\}^{L\times d}$, (using the alphabet of AAs).

② sequences of vectors in $\mathbb{R}^p$, i.e. $\underline{\mathbf{x}} \in \mathbb{R}^{L\times p}$, (using p bio-physical properties).

③ graph signals: $(\underline{\mathbf{x}}, \mathcal{E})$, where there is an edge (collected in the set $\mathcal{E}$) between two amino-acids (the nodes) if the amino-acids are close in 3D space.

Variant effect prediction: Proteins

Encoding

Proteins can be encoded as

① sequences of one-hot vectors in $\{0,1\}^d$, i.e. $\underline{\mathbf{x}} \in \{0,1\}^{L\times d}$, (using the alphabet of AAs).

② sequences of vectors in $\mathbb{R}^p$, i.e. $\underline{\mathbf{x}} \in \mathbb{R}^{L\times p}$, (using p bio-physical properties).

③ graph signals: $(\underline{\mathbf{x}}, \mathcal{E})$, where there is an edge (collected in the set $\mathcal{E}$) between two amino-acids (the nodes) if the amino-acids are close in 3D space.

Variant effect prediction: Proteins

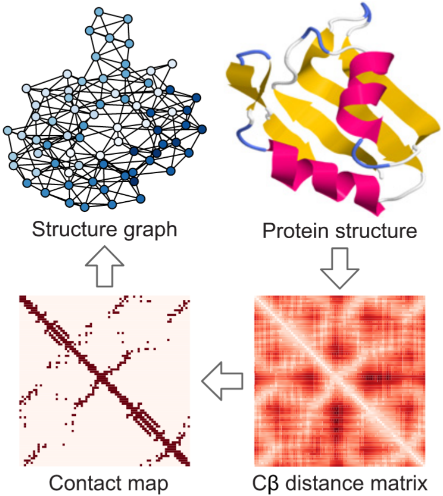
Encoding

Proteins can be encoded as

① sequences of one-hot vectors in $\{0,1\}^d$, i.e. $\underline{\mathbf{x}} \in \{0,1\}^{L\times d}$, (using the alphabet of AAs).

② sequences of vectors in $\mathbb{R}^p$, i.e. $\underline{\mathbf{x}} \in \mathbb{R}^{L\times p}$, (using p bio-physical properties).

③ graph signals: $(\underline{\mathbf{x}}, \mathcal{E})$, where there is an edge (collected in the set $\mathcal{E}$) between two amino-acids (the nodes) if the amino-acids are close in 3D space.



Variant effect prediction: Proteins

Why predicting protein variant effects ?

- In our project we focus on drug transporter proteins:
  - Membrane proteins involved in drug absorption, distribution, metabolism and excretion (ADME) [5].
- Loss-of-function variants are directly related to an individual's ability to respond to a drug treatment, e.g. chemotherapy.
- Being able to predict the effect of variants on these proteins can help select personalized patient treatments.

Variant effect prediction: Proteins

Why predicting protein variant effects ?

- In our project we focus on drug transporter proteins:
  - Membrane proteins involved in drug absorption, distribution, metabolism and excretion (ADME) [5].
- Loss-of-function variants are directly related to an individual's ability to respond to a drug treatment, e.g. chemotherapy.
- Being able to predict the effect of variants on these proteins can help select personalized patient treatments.

Variant effect prediction: Proteins

Why predicting protein variant effects ?

- In our project we focus on drug transporter proteins:
  - Membrane proteins involved in drug absorption, distribution, metabolism and excretion (ADME) [5].
- Loss-of-function variants are directly related to an individual's ability to respond to a drug treatment, e.g. chemotherapy.
- Being able to predict the effect of variants on these proteins can help select personalized patient treatments.

Variant effect prediction: Proteins

Why predicting protein variant effects ?

- In our project we focus on drug transporter proteins:
  - Membrane proteins involved in drug absorption, distribution, metabolism and excretion (ADME) [5].
- Loss-of-function variants are directly related to an individual's ability to respond to a drug treatment, e.g. chemotherapy.
- Being able to predict the effect of variants on these proteins can help select personalized patient treatments.

Variant effect prediction: Evolutionary constrain assumption

Variant effect prediction: Evolutionary constrain assumption

Formulated in terms of positional information content

- Given N observations of a certain protein, e.g. across species:
  - Positions where many observations have the same amino-acids are functionally important
    - Why? because preserved by evolution.
  - Variants at these positions are considered deleterious
    - Quantified by the amount of self-information at these positions [6]

Variant effect prediction: Evolutionary constrain assumption

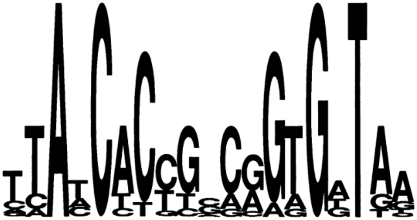
Formulated in terms of positional information content

- Given N observations of a certain protein, e.g. across species:
  - Positions where many observations have the same amino-acids are functionally important
    - Why? because preserved by evolution.
  - Variants at these positions are considered deleterious
    - Quantified by the amount of self-information at these positions [6]

Variant effect prediction: Evolutionary constrain assumption

Formulated in terms of positional information content

- Given N observations of a certain protein, e.g. across species:
  - Positions where many observations have the same amino-acids are functionally important
    - Why? because preserved by evolution.
  - Variants at these positions are considered deleterious
    - Quantified by the amount of self-information at these positions [6]



A sequence logo. From [7].
x-axis: Position in a sequence of AA

y-axis:

- At each position i, $f_{i,j}$ the frequency of letter j
  - The entropy: $H_i = -\sum_{j=1}^{4} f_{i,j} \log_2 f_{i,j}$
  - Maximum height = maximum self-information

$$I_{max} = -\log_2(1/4) = 2 \text{ bits}$$

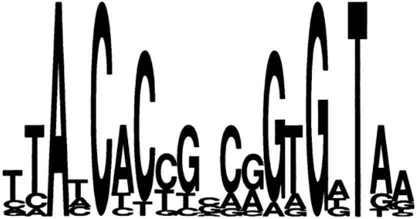
  - Letter j height

$$(I_{max} - H_i)\, f_{i,j}$$

Variant effect prediction: Evolutionary constrain assumption

Formulated in terms of positional information content

- Given N observations of a certain protein, e.g. across species:
  - Positions where many observations have the same amino-acids are functionally important
    - Why? because preserved by evolution.
  - Variants at these positions are considered deleterious
    - Quantified by the amount of self-information at these positions [6]



A sequence logo. From [7].
x-axis: Position in a sequence of AA

y-axis:

- At each position i, $f_{i,j}$ the frequency of letter j
  - The entropy: $H_i = -\sum_{j=1}^{4} f_{i,j} \log_2 f_{i,j}$
  - Maximum height = maximum self-information

$$I_{max} = -\log_2(1/4) = 2 \text{ bits}$$

  - Letter j height

$$(I_{max} - H_i)\, f_{i,j}$$

Variant effect prediction: Evolutionary constrain assumption

Formulated in terms of distributions over sequence space

- Another way to phrase it, found in [8]:
  1. The distribution of protein sequences observed in nature is the result of billions of evolutionary experiments
  2. Unfit variants were selected out
  3. By learning a distribution over these sequences we implicitly capture the biochemical constraints that characterize fit variants

The degree of fitness of a variant can be quantified as a log-likelihood ratio:

$$\ln \frac{p(\underline{\mathbf{x}}^{(v)}|\theta)}{p(\underline{\mathbf{x}}^{(wt)}|\theta)}, \qquad (1)$$

- $\underline{\mathbf{x}}^{(v)}$ (resp. $\underline{\mathbf{x}}^{(wt)}$): the encoded sequence with (resp. without) the variant.
- θ: parameter set, learnt from sequences found in nature only !

Variant effect prediction: Evolutionary constrain assumption

Formulated in terms of distributions over sequence space

- Another way to phrase it, found in [8]:
  1. The distribution of protein sequences observed in nature is the result of billions of evolutionary experiments
  2. Unfit variants were selected out
  3. By learning a distribution over these sequences we implicitly capture the biochemical constraints that characterize fit variants

The degree of fitness of a variant can be quantified as a log-likelihood ratio:

$$\ln \frac{p(\underline{\mathbf{x}}^{(v)}|\theta)}{p(\underline{\mathbf{x}}^{(wt)}|\theta)}, \tag{1}$$

- $\underline{\mathbf{x}}^{(v)}$ (resp. $\underline{\mathbf{x}}^{(wt)}$): the encoded sequence with (resp. without) the variant.
- θ: parameter set, learnt from sequences found in nature only !

Variant effect prediction: Evolutionary constrain assumption

Formulated in terms of distributions over sequence space

- Another way to phrase it, found in [8]:
  1. The distribution of protein sequences observed in nature is the result of billions of evolutionary experiments
  2. Unfit variants were selected out
  3. By learning a distribution over these sequences we implicitly capture the biochemical constraints that characterize fit variants

The degree of fitness of a variant can be quantified as a log-likelihood ratio:

$$\ln \frac{p(\underline{\mathbf{x}}^{(v)}|\theta)}{p(\underline{\mathbf{x}}^{(wt)}|\theta)}, \tag{1}$$

- $\underline{\mathbf{x}}^{(v)}$ (resp. $\underline{\mathbf{x}}^{(wt)}$): the encoded sequence with (resp. without) the variant.
- θ: parameter set, learnt from sequences found in nature only !

Variant effect prediction: Evolutionary constrain assumption

Formulated in terms of distributions over sequence space

- Another way to phrase it, found in [8]:
  1. The distribution of protein sequences observed in nature is the result of billions of evolutionary experiments
  2. Unfit variants were selected out
  3. By learning a distribution over these sequences we implicitly capture the biochemical constraints that characterize fit variants

The degree of fitness of a variant can be quantified as a log-likelihood ratio:

$$\ln \frac{p(\underline{\mathbf{x}}^{(v)}|\theta)}{p(\underline{\mathbf{x}}^{(wt)}|\theta)}, \tag{1}$$

- $\underline{\mathbf{x}}^{(v)}$ (resp. $\underline{\mathbf{x}}^{(wt)}$): the encoded sequence with (resp. without) the variant.
- θ: parameter set, learnt from sequences found in nature only !

Variant effect prediction: Evolutionary constrain assumption

Formulated in terms of distributions over sequence space

- Another way to phrase it, found in [8]:
  1. The distribution of protein sequences observed in nature is the result of billions of evolutionary experiments
  2. Unfit variants were selected out
  3. By learning a distribution over these sequences we implicitly capture the biochemical constraints that characterize fit variants

The degree of fitness of a variant can be quantified as a log-likelihood ratio:

$$\ln \frac{p(\underline{\mathbf{x}}^{(v)}|\theta)}{p(\underline{\mathbf{x}}^{(wt)}|\theta)}, \tag{1}$$

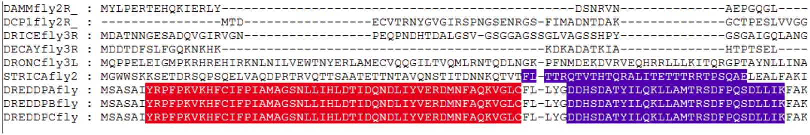
- $\underline{\mathbf{x}}^{(v)}$ (resp. $\underline{\mathbf{x}}^{(wt)}$): the encoded sequence with (resp. without) the variant.
- θ: parameter set, learnt from sequences found in nature only !

Variant effect prediction: Data

Multiple Sequence Alignment (MSA)

Observations of a protein in different organisms

- Aligned sequencing outputs for different observations of a protein:



```
DAMMfly2R_ : MYLPERTEHQKIERLY----------------------------------------------DSNRVN---------------AEPGQGL----
DCP1fly2R_ : ----------------MTD-----------------ECVTRNYGVGIRSPNGSENRGS-FIMADNTDAK--------------GCTPESLVVGG
DRICEfly3R : MDATNNGESADQVGIRVGN-----------------PEQPNDHTDALGSV-GSGGAGSSGLVAGSSHPY--------------GSGAIGQLANG
DECAYfly3R : MDDTDFSLFGQKNKHK-------------------------------------------KDKADATKIA--------------HTPTSEL----
DRONCfly3L : MQPPELEIGMPKRHREHIRKNLNILVEWTNYERLAMECVQQGILTVQMLRNTQDLNGK-PFNMDEKDVRVEQHRRLLLKITQRGPTAYNLLINA
STRICAfly2 : MGWWSKKSETDRSQPSQELVAQDPRTRVQTTSAATETTNTAVQNSTITDNNKQTVTFL-TTRQTVTHTQRALITETTTRRTPSQAELEALFAKI
DREDDPAfly : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPBfly : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPCfly : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
```

- Large databases contain curated MSAs:
  - Wild-type: A reference sequence, derived from a MSA
  - Variants: All other sequences.
- Some variants are indexed in other databases,
  - associated with clinical phenotypes: needed to score variant effect predictors.

Output

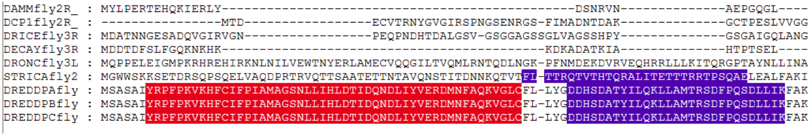
$\{X\}$ set of sequences to train both positional information-based, and distribution-based models

Variant effect prediction: Data

Multiple Sequence Alignment (MSA)

Observations of a protein in different organisms

- Aligned sequencing outputs for different observations of a protein:



```
DAMMfly2R_  : MYLPERTEHQKIERLY-----------------------------------------------DSNRVN--------------AEPGQGL----
DCP1fly2R_  : ----------------MTD-----------------ECVTRNYGVGIRSPNGSENRGS-FIMADNTDAK--------------GCTPESLVVGG
DRICEfly3R  : MDATNNGESADQVGIRVGN-----------------PEQPNDHTDALGSV-GSGGAGSSGLVAGSSHPY--------------GSGAIGQLANG
DECAYfly3R  : MDDTDFSLFGQKNKHK-------------------------------------------KDKADATKIA--------------HTPTSEL----
DRONCfly3L  : MQPPELEIGMPKRHREHIRKNLNILVEWTNYERLAMECVQQGILTVQMLRNTQDLNGK-PFNMDEKDVRVEQHRRLLLKITQRGPTAYNLLINA
STRICAfly2  : MGWWSKKSETDRSQPSQELVAQDPRTRVQTTSAATETTNTAVQNSTITDNNKQTVTFL-TTRQTVTHTQRALITETTTRRTPSQAELEALFAKI
DREDDPAfly  : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPBfly  : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPCfly  : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
```

- Large databases contain curated MSAs:
  - Wild-type: A reference sequence, derived from a MSA
  - Variants: All other sequences.
- Some variants are indexed in other databases,
  - associated with clinical phenotypes: needed to score variant effect predictors.

Output

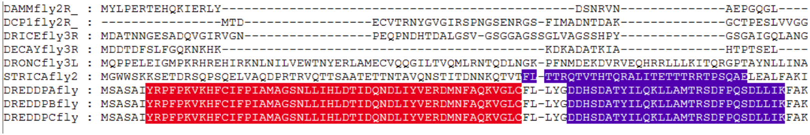
$\{X\}$ set of sequences to train both positional information-based, and distribution-based models

Variant effect prediction: Data

Multiple Sequence Alignment (MSA)

Observations of a protein in different organisms

- Aligned sequencing outputs for different observations of a protein:



```
DAMMfly2R_  : MYLPERTEHQKIERLY-----------------------------------------------DSNRVN--------------AEPGQGL----
DCP1fly2R_  : ----------------MTD-----------------ECVTRNYGVGIRSPNGSENRGS-FIMADNTDAK--------------GCTPESLVVGG
DRICEfly3R  : MDATNNGESADQVGIRVGN-----------------PEQPNDHTDALGSV-GSGGAGSSGLVAGSSHPY--------------GSGAIGQLANG
DECAYfly3R  : MDDTDFSLFGQKNKHK-------------------------------------------KDKADATKIA--------------HTPTSEL----
DRONCfly3L  : MQPPELEIGMPKRHREHIRKNLNILVEWTNYERLAMECVQQGILTVQMLRNTQDLNGK-PFNMDEKDVRVEQHRRLLLKITQRGPTAYNLLINA
STRICAfly2  : MGWWSKKSETDRSQPSQELVAQDPRTRVQTTSAATETTNTAVQNSTITDNNKQTVTFL-TTRQTVTHTQRALITETTTRRTPSQAELEALFAKI
DREDDPAfly  : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPBfly  : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPCfly  : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
```

- Large databases contain curated MSAs:
  - Wild-type: A reference sequence, derived from a MSA
  - Variants: All other sequences.
- Some variants are indexed in other databases,
  - associated with clinical phenotypes: needed to score variant effect predictors.

Output

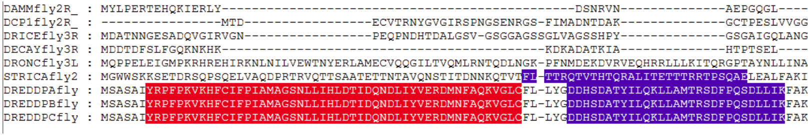
$\{X\}$ set of sequences to train both positional information-based, and distribution-based models

Variant effect prediction: Data

Multiple Sequence Alignment (MSA)

Observations of a protein in different organisms

- Aligned sequencing outputs for different observations of a protein:



```
DAMMfly2R_ : MYLPERTEHQKIERLY-----------------------------------------------DSNRVN--------------AEPGQGL----
DCP1fly2R_ : ----------------MTD-----------------ECVTRNYGVGIRSPNGSENRGS-FIMADNTDAK--------------GCTPESLVVGG
DRICEfly3R : MDATNNGESADQVGIRVGN-----------------PEQPNDHTDALGSV-GSGGAGSSGLVAGSSHPY--------------GSGAIGQLANG
DECAYfly3R : MDDTDFSLFGQKNKHK-------------------------------------------KDKADATKIA--------------HTPTSEL----
DRONCfly3L : MQPPELEIGMPKRHREHIRKNLNILVEWTNYERLAMECVQQGILTVQMLRNTQDLNGK-PFNMDEKDVRVEQHRRLLLKITQRGPTAYNLLINA
STRICAfly2 : MGWWSKKSETDRSQPSQELVAQDPRTRVQTTSAATETTNTAVQNSTITDNNKQTVTFL-TTRQTVTHTQRALITETTTRRTPSQAELEALFAKI
DREDDPAfly : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPBfly : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPCfly : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
```

- Large databases contain curated MSAs:
    - Wild-type: A reference sequence, derived from a MSA
    - Variants: All other sequences.
- Some variants are indexed in other databases,
    - associated with clinical phenotypes: needed to score variant effect predictors.

Output

$\{X\}$ set of sequences to train both positional information-based, and distribution-based models

Variant effect prediction: Data

Deep Mutational Scanning

Hard-coding mutations in cells

① Create a collection of cells with many single variants

- Sequence the collection of cells to get a count of the initial variants (✤)

② Expose the cells to an environment

- Sequence again and count the remaining variants (✦)

③ Compare ✦ and ✤ to get a cell variant ”survival score”

Output

$\{(X, y)\}$: X a varied sequence and $y \in \mathbb{R}$ a regression target !

Variant effect prediction: Data

Deep Mutational Scanning

Hard-coding mutations in cells

① Create a collection of cells with many single variants

- ▶ Sequence the collection of cells to get a count of the initial variants (✤)

② Expose the cells to an environment

- ▶ Sequence again and count the remaining variants (✦)

③ Compare ✦ and ✤ to get a cell variant "survival score"

Output

$\{(X, y)\}$: X a varied sequence and $y \in \mathbb{R}$ a regression target !

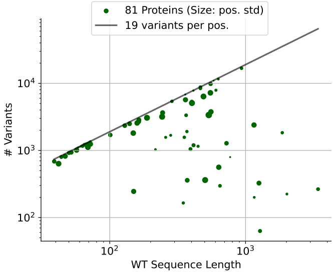


Figure: Position coverage of many experiments

Variant effect prediction: Data

Deep Mutational Scanning

Hard-coding mutations in cells

① Create a collection of cells with many single variants

- ▶ Sequence the collection of cells to get a count of the initial variants (✤)

② Expose the cells to an environment

- ▶ Sequence again and count the remaining variants (✦)

③ Compare ✦ and ✤ to get a cell variant "survival score"

Output

$\{(X, y)\}$: X a varied sequence and $y \in \mathbb{R}$ a regression target !

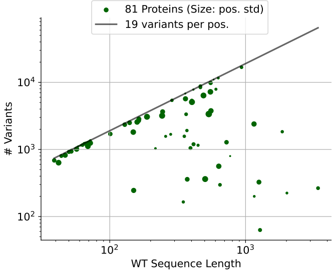


Figure: Position coverage of many experiments

Variant effect prediction: Data

Deep Mutational Scanning

Hard-coding mutations in cells

① Create a collection of cells with many single variants

- ▶ Sequence the collection of cells to get a count of the initial variants (✤)

② Expose the cells to an environment

- ▶ Sequence again and count the remaining variants (✦)

③ Compare ✦ and ✤ to get a cell variant "survival score"

Output

$\{(X, y)\}$: X a varied sequence and $y \in \mathbb{R}$ a regression target !

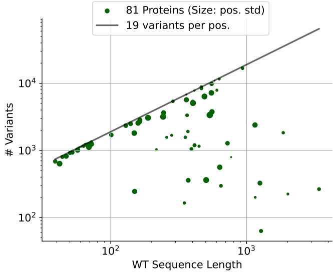


Figure: Position coverage of many experiments

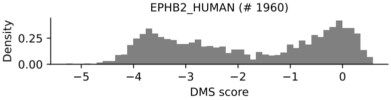


Figure: Density of output scores for 1 experiment

Variant effect prediction: Data

Deep Mutational Scanning

Hard-coding mutations in cells

① Create a collection of cells with many single variants

- ▶ Sequence the collection of cells to get a count of the initial variants (✤)

② Expose the cells to an environment

- ▶ Sequence again and count the remaining variants (✦)

③ Compare ✦ and ✤ to get a cell variant ”survival score”

Output

$\{(X, y)\}$: X a varied sequence and $y \in \mathbb{R}$ a regression target !

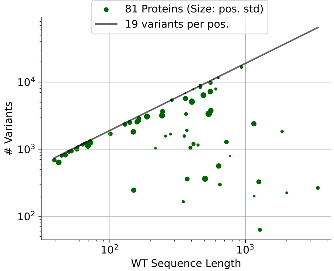


Figure: Position coverage of many experiments

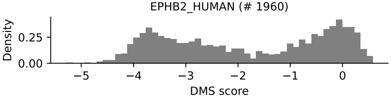


Figure: Density of output scores for 1 experiment

Variant effect prediction: Predictors

Benchmarking

Models are evaluated on variants used in a DMS experiment, and referenced in a clinical database.

1. x-axis: AUC (Area under the receiver-operating curve)
   - Computed against clinical observation
2. y-axis: Correlation between the rank predicted and the rank in the DMS experiment.

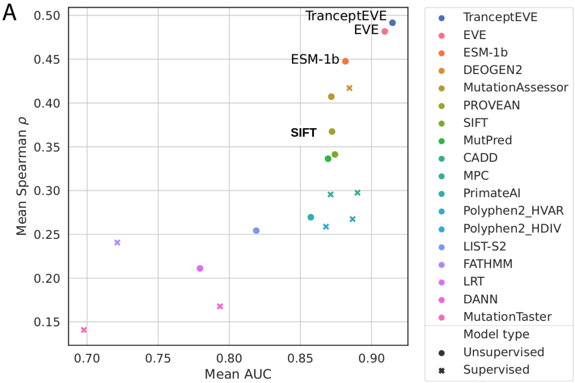


Figure from [9].

Variant effect prediction: Predictors

Benchmarking

Models are evaluated on variants used in a DMS experiment, and referenced in a clinical database.

① x-axis: AUC (Area under the receiver-operating curve)

- Computed against clinical observation

② y-axis: Correlation between the rank predicted and the rank in the DMS experiment.

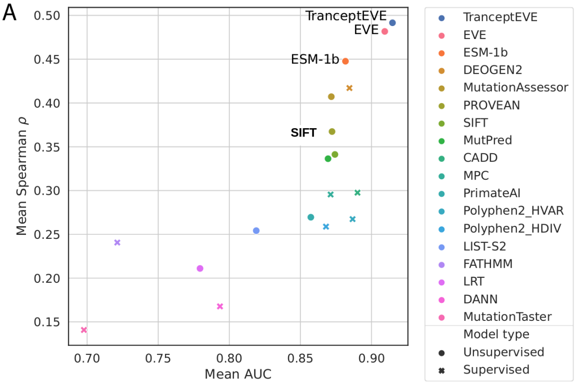


Figure from [9].

Variant effect prediction: Predictors

Benchmarking

Models are evaluated on variants used in a DMS experiment, and referenced in a clinical database.

① x-axis: AUC (Area under the receiver-operating curve)
   - ▶ Computed against clinical observation

② y-axis: Correlation between the rank predicted and the rank in the DMS experiment.

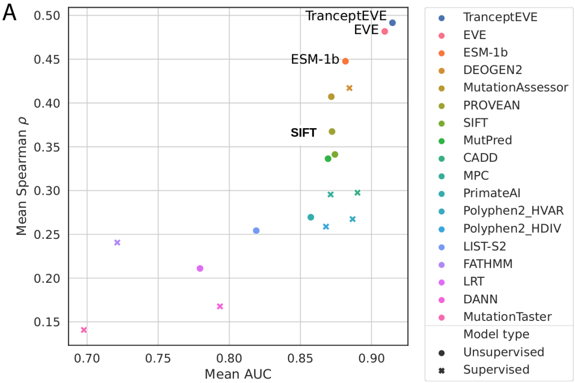


Figure from [9].

Predictors: SIFT

SIFT: Sorting Intolerant From Tolerant

① Given an input WT sequence and a variant

② Query databases to find similar aligned sequences

③ Computes Shannon entropy at each position [10]

④ Sum to high entropy $\implies$ likely benign variant.

Limitations

① The prediction depends upon the similar sequences that were queried

② Only the position of the variant is taken into account, not the impact on the relation to neighbor AA

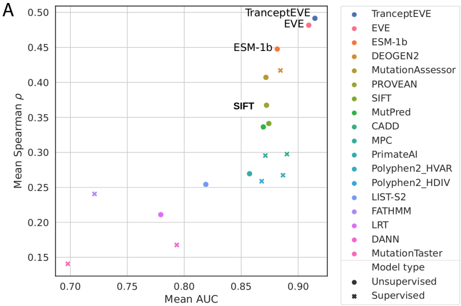


Figure from [9].

Predictors: SIFT

SIFT: Sorting Intolerant From Tolerant

1. Given an input WT sequence and a variant
2. Query databases to find similar aligned sequences
3. Computes Shannon entropy at each position [10]
4. Sum to high entropy $\implies$ likely benign variant.

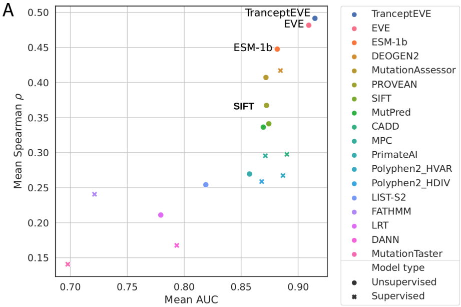


Figure from [9].

Limitations

1. The prediction depends upon the similar sequences that were queried
2. Only the position of the variant is taken into account, not the impact on the relation to neighbor AA

Predictors: SIFT

SIFT: Sorting Intolerant From Tolerant

① Given an input WT sequence and a variant

② Query databases to find similar aligned sequences

③ Computes Shannon entropy at each position [10]

④ Sum to high entropy $\implies$ likely benign variant.

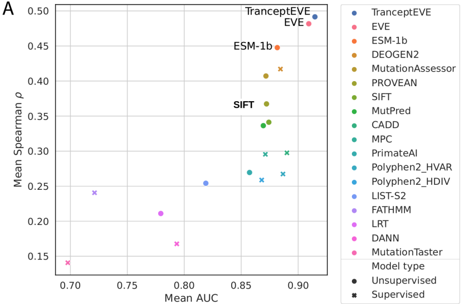


Figure from [9].

Limitations

① The prediction depends upon the similar sequences that were queried

② Only the position of the variant is taken into account, not the impact on the relation to neighbor AA

Predictors: SIFT

SIFT: Sorting Intolerant From Tolerant

1. Given an input WT sequence and a variant
2. Query databases to find similar aligned sequences
3. Computes Shannon entropy at each position [10]
4. Sum to high entropy $\implies$ likely benign variant.

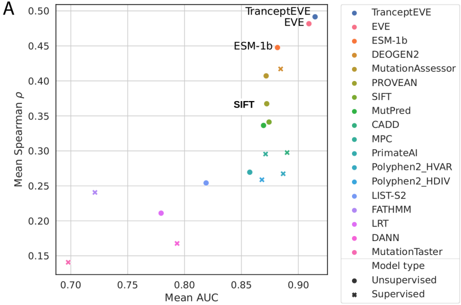


Figure from [9].

Limitations

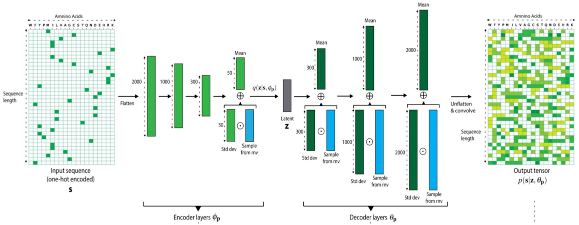
1. The prediction depends upon the similar sequences that were queried
2. Only the position of the variant is taken into account, not the impact on the relation to neighbor AA

Predictors: Deep Sequence

Deep Sequence

A variational auto-encoder (VAE) [3]

- Learn the parameters of the log-likelihood by maximizing a lower bound



Limitations

① Vectorized input !

- Needs retraining on new proteins

② Unimodal distribution in the latent space

- Limit expressivity of latent vectors

③ For inference, the ratio of likelihoods is replaced with a ratio of lower bounds:

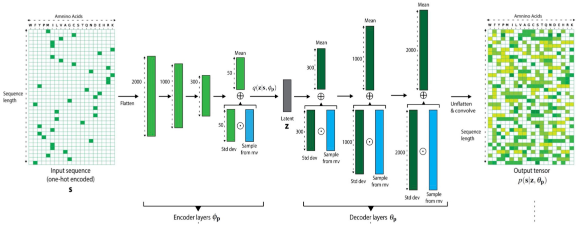
$$\ln \frac{p(\underline{\mathbf{x}}^{v}|\theta)}{p(\underline{\mathbf{x}}^{wt}|\theta)} \approx \ln \frac{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{v}|\theta)}{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{wt}|\theta)} \qquad (2)$$

Predictors: Deep Sequence

Deep Sequence

A variational auto-encoder (VAE) [3]

- Learn the parameters of the log-likelihood by maximizing a lower bound



Limitations

① Vectorized input !

- Needs retraining on new proteins

② Unimodal distribution in the latent space

- Limit expressivity of latent vectors

③ For inference, the ratio of likelihoods is replaced with a ratio of lower bounds:

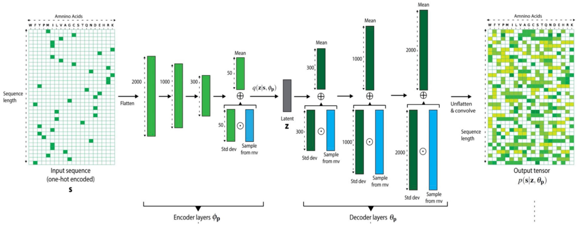
$$\ln \frac{p(\underline{\mathbf{x}}^{v}|\theta)}{p(\underline{\mathbf{x}}^{wt}|\theta)} \approx \ln \frac{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{v}|\theta)}{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{wt}|\theta)} \tag{2}$$

Predictors: Deep Sequence

Deep Sequence

A variational auto-encoder (VAE) [3]

- Learn the parameters of the log-likelihood by maximizing a lower bound



Limitations

① Vectorized input !

- Needs retraining on new proteins

② Unimodal distribution in the latent space

- Limit expressivity of latent vectors

③ For inference, the ratio of likelihoods is replaced with a ratio of lower bounds:

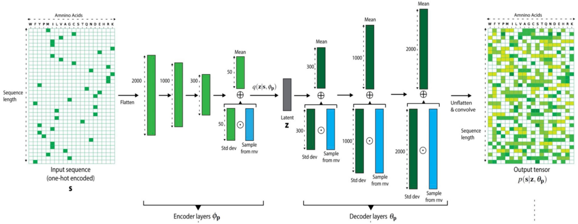
$$\ln \frac{p(\underline{\mathbf{x}}^{v}|\theta)}{p(\underline{\mathbf{x}}^{wt}|\theta)} \approx \ln \frac{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{v}|\theta)}{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{wt}|\theta)} \qquad (2)$$

Predictors: Deep Sequence

Deep Sequence

A variational auto-encoder (VAE) [3]

- Learn the parameters of the log-likelihood by maximizing a lower bound



Limitations

① Vectorized input !

- Needs retraining on new proteins

② Unimodal distribution in the latent space

- Limit expressivity of latent vectors

③ For inference, the ratio of likelihoods is replaced with a ratio of lower bounds:

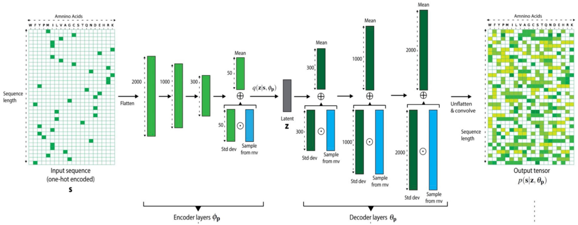
$$\ln \frac{p(\underline{\mathbf{x}}^{v}|\theta)}{p(\underline{\mathbf{x}}^{wt}|\theta)} \approx \ln \frac{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{v}|\theta)}{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{wt}|\theta)} \tag{2}$$

Predictors: Deep Sequence

Deep Sequence

A variational auto-encoder (VAE) [3]

- Learn the parameters of the log-likelihood by maximizing a lower bound



Limitations

① Vectorized input !

- Needs retraining on new proteins

② Unimodal distribution in the latent space

- Limit expressivity of latent vectors

③ For inference, the ratio of likelihoods is replaced with a ratio of lower bounds:

$$\ln \frac{p(\underline{\mathbf{x}}^{v}|\theta)}{p(\underline{\mathbf{x}}^{wt}|\theta)} \approx \ln \frac{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{v}|\theta)}{\mathcal{L}_{VLB}(\underline{\mathbf{x}}^{wt}|\theta)} \tag{2}$$

Predictors: Deep Sequence

My current work: addressing the limitations of DeepSequence

- Allow multimodality in the latent space:
  - Mixture of Gaussian, instead of Gaussian ?
  - Did not work, the mixture was not used by the model.
- Better idea:
  - Latent space = probability simplex
  - Replace D_{KL} with entropy of latent vector in the lower bound

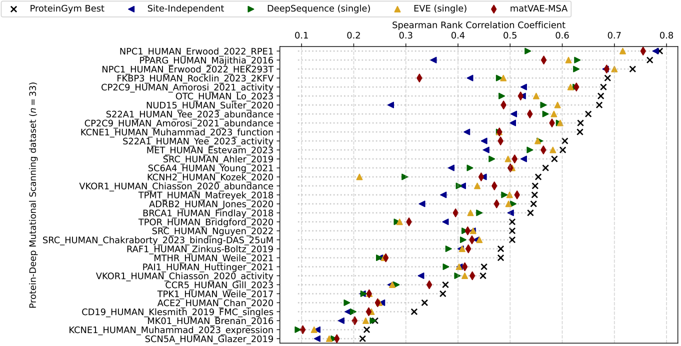
Recall: $\ln p(\mathbf{x}) = D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})\right) - D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln p_\theta(\mathbf{x}|\mathbf{z})\right]$

Predictors: Deep Sequence

My current work: addressing the limitations of DeepSequence

- Less parameters (8-9M) (no need for Bayesian decoder)
- More interpretability,
- Better/similar results!

Model name	Spearmanr
Best Benchmark	0.529 ± 0.151
EVE (single)	0.432 ± 0.161
matVAE-MSA	0.428 ± 0.15
DeepSequence (single)	0.412 ± 0.149
Site-Independent	0.39 ± 0.145



More predictors, based on Transformers: ESM

ESM: Evolutionary Scale Modeling

A protein model with $650M$ parameters [11]

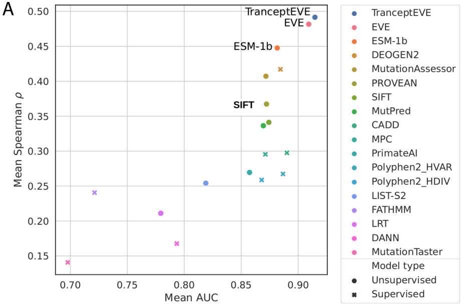
① Given an input varied sequence $\underline{\mathbf{x}}^v = (\mathbf{x}_1, \ldots, \mathbf{x}_L)^T$, the statistical model is similar to a large language model :

$$p(\underline{\mathbf{x}}^v|\theta) = \prod_{i=1}^{L} p(\mathbf{x}_i|\mathbf{x}_{N(i)}, \theta), \qquad (3)$$

with $N(i)$ a randomly masked neighborhood for i.

② The prediction score is the ratio

$$\ln \frac{p(\underline{\mathbf{x}}^v|\theta)}{p(\underline{\mathbf{x}}^{wt}|\theta)} \qquad (4)$$



More predictors, based on Transformers: ESM

ESM: Evolutionary Scale Modeling

A protein model with 650M parameters [11]

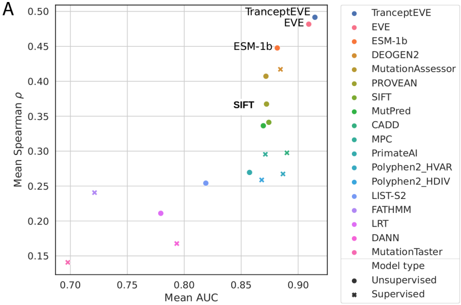
① Given an input varied sequence $\underline{\mathbf{x}}^v = (\mathbf{x}_1, \ldots, \mathbf{x}_L)^T$, the statistical model is similar to a large language model :

$$p(\underline{\mathbf{x}}^v|\theta) = \prod_{i=1}^{L} p(\mathbf{x}_i|\mathbf{x}_{N(i)}, \theta), \tag{3}$$

with $N(i)$ a randomly masked neighborhood for i.

② The prediction score is the ratio

$$\ln \frac{p(\underline{\mathbf{x}}^v|\theta)}{p(\underline{\mathbf{x}}^{wt}|\theta)} \tag{4}$$



More predictors, based on Transformers: ESM

ESM: Evolutionary Scale Modeling

A protein model with $650M$ parameters [11]

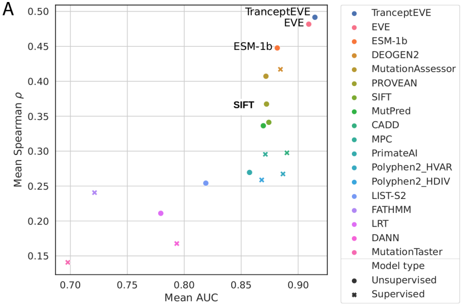
① Given an input varied sequence $\underline{\mathbf{x}}^v = (\mathbf{x}_1, \ldots, \mathbf{x}_L)^T$, the statistical model is similar to a large language model :

$$p(\underline{\mathbf{x}}^v|\theta) = \prod_{i=1}^{L} p(\mathbf{x}_i|\mathbf{x}_{N(i)}, \theta), \tag{3}$$

with $N(i)$ a randomly masked neighborhood for i.

② The prediction score is the ratio

$$\ln \frac{p(\underline{\mathbf{x}}^v|\theta)}{p(\underline{\mathbf{x}}^{wt}|\theta)} \tag{4}$$



More predictors, based on Transformers: TranceptEVE

TranceptEVE: Tranception + EVE

An auto-regressive protein language model combined with other models [3]

- Tranception [12]: auto-regressive model, i.e. similar to ESM but the neighborhood is all the previous AAs.

$$p_T(\mathbf{x}_1, \dots, \mathbf{x}_L | \theta) = \prod_{i=1}^{L} p_T(\mathbf{x}_i | \mathbf{x}_{i-1}, \dots, \mathbf{x}_1, \theta). \tag{5}$$

- The final log-likelihood is computed as a convex combination of log-likelihood from other models:

$$\log p(\underline{\mathbf{x}} | \theta) \propto \sum_{i=1}^{L} (1-\beta)[(1-\alpha)\log p_T(\mathbf{x}_i | \mathbf{x}_{<i}) + \alpha \log p_R(\mathbf{x}_i)] + \beta \log p_E(\mathbf{x}_i), \tag{6}$$

- p_R is the empirical distribution from the MSA, p_E is the output of EVE, (α, β) are hyper-parameters which favor p_T when the MSA is shallow, i.e. unreliable.

More predictors, based on Transformers: TranceptEVE

TranceptEVE: Tranception + EVE

An auto-regressive protein language model combined with other models [3]

- Tranception [12]: auto-regressive model, i.e. similar to ESM but the neighborhood is all the previous AAs.

$$p_T(\mathbf{x}_1, \ldots, \mathbf{x}_L|\theta) = \prod_{i=1}^{L} p_T(\mathbf{x}_i|\mathbf{x}_{i-1}, \ldots, \mathbf{x}_1, \theta). \tag{5}$$

- The final log-likelihood is computed as a convex combination of log-likelihood from other models:

$$\log p(\underline{\mathbf{x}}|\theta) \propto \sum_{i=1}^{L} (1-\beta)[(1-\alpha)\log p_T(\mathbf{x}_i|\mathbf{x}_{<i}) + \alpha \log p_R(\mathbf{x}_i)] + \beta \log p_E(\mathbf{x}_i), \tag{6}$$

- p_R is the empirical distribution from the MSA, p_E is the output of EVE, (α, β) are hyper-parameters which favor p_T when the MSA is shallow, i.e. unreliable.

More predictors, based on Transformers: TranceptEVE

TranceptEVE: Tranception + EVE

An auto-regressive protein language model combined with other models [3]

- Tranception [12]: auto-regressive model, i.e. similar to ESM but the neighborhood is all the previous AAs.

$$p_T(\mathbf{x}_1, \ldots, \mathbf{x}_L | \theta) = \prod_{i=1}^{L} p_T(\mathbf{x}_i | \mathbf{x}_{i-1}, \ldots, \mathbf{x}_1, \theta). \tag{5}$$

- The final log-likelihood is computed as a convex combination of log-likelihood from other models:

$$\log p(\underline{\mathbf{x}}|\theta) \propto \sum_{i=1}^{L} (1-\beta)[(1-\alpha)\log p_T(\mathbf{x}_i|\mathbf{x}_{<i}) + \alpha \log p_R(\mathbf{x}_i)] + \beta \log p_E(\mathbf{x}_i), \tag{6}$$

- p_R is the empirical distribution from the MSA, p_E is the output of EVE, (α, β) are hyper-parameters which favor p_T when the MSA is shallow, i.e. unreliable.

More predictors, based on Transformers: TranceptEVE

TranceptEVE: Tranception + EVE

An auto-regressive protein language model combined with other models [3]

- Tranception [12]: auto-regressive model, i.e. similar to ESM but the neighborhood is all the previous AAs.

$$p_T(\mathbf{x}_1, \ldots, \mathbf{x}_L|\theta) = \prod_{i=1}^{L} p_T(\mathbf{x}_i|\mathbf{x}_{i-1}, \ldots, \mathbf{x}_1, \theta). \tag{5}$$

- The final log-likelihood is computed as a convex combination of log-likelihood from other models:

$$\log p(\underline{\mathbf{x}}|\theta) \propto \sum_{i=1}^{L} (1-\beta)[(1-\alpha)\log p_T(\mathbf{x}_i|\mathbf{x}_{<i}) + \alpha \log p_R(\mathbf{x}_i)] + \beta \log p_E(\mathbf{x}_i), \tag{6}$$

- p_R is the empirical distribution from the MSA, p_E is the output of EVE, (α, β) are hyper-parameters which favor p_T when the MSA is shallow, i.e. unreliable.

Future perspectives

1. Existing prediction models remain reliant on evolutionary constrain assumption:
   - Either through the informational or probabilistic formulation, or both.
   - Problem: Drug transporter proteins were not constrained by evolution (similar to nutriment transport in and out of cells)
2. Many models do not take into account the 3d structure of proteins:
   - A few works used graph neural networks, did not show very big improvements over e.g. CNN
3. Deep mutational scanning data are promising
   - Problem:
     - lack of standardization of experimental methods, i.e. output are not always numerically comparable.
   - Advantages:
     - Can provide information on specific phenotype,
     - Train against quantitative functional scores

Future perspectives

1. Existing prediction models remain reliant on evolutionary constrain assumption:
    - Either through the informational or probabilistic formulation, or both.
    - Problem: Drug transporter proteins were not constrained by evolution (similar to nutriment transport in and out of cells)
2. Many models do not take into account the 3d structure of proteins:
    - A few works used graph neural networks, did not show very big improvements over e.g. CNN
3. Deep mutational scanning data are promising
    - Problem:
        - lack of standardization of experimental methods, i.e. output are not always numerically comparable.
    - Advantages:
        - Can provide information on specific phenotype,
        - Train against quantitative functional scores

Future perspectives

① Existing prediction models remain reliant on evolutionary constrain assumption:

- Either through the informational or probabilistic formulation, or both.
- Problem: Drug transporter proteins were not constrained by evolution (similar to nutriment transport in and out of cells)

② Many models do not take into account the 3d structure of proteins:

- A few works used graph neural networks, did not show very big improvements over e.g. CNN

③ Deep mutational scanning data are promising

- Problem:
  - lack of standardization of experimental methods, i.e. output are not always numerically comparable.
- Advantages:
  - Can provide information on specific phenotype,
  - Train against quantitative functional scores

Future perspectives

① Existing prediction models remain reliant on evolutionary constrain assumption:
   - Either through the informational or probabilistic formulation, or both.
   - Problem: Drug transporter proteins were not constrained by evolution (similar to nutriment transport in and out of cells)

② Many models do not take into account the 3d structure of proteins:
   - A few works used graph neural networks, did not show very big improvements over e.g. CNN

③ Deep mutational scanning data are promising
   - Problem:
     - lack of standardization of experimental methods, i.e. output are not always numerically comparable.
   - Advantages:
     - Can provide information on specific phenotype,
     - Train against quantitative functional scores

Future perspectives

① Existing prediction models remain reliant on evolutionary constrain assumption:
- Either through the informational or probabilistic formulation, or both.
- Problem: Drug transporter proteins were not constrained by evolution (similar to nutriment transport in and out of cells)

② Many models do not take into account the 3d structure of proteins:
- A few works used graph neural networks, did not show very big improvements over e.g. CNN

③ Deep mutational scanning data are promising
- Problem:
  - lack of standardization of experimental methods, i.e. output are not always numerically comparable.
- Advantages:
  - Can provide information on specific phenotype,
  - Train against quantitative functional scores

Future perspectives

1. Existing prediction models remain reliant on evolutionary constrain assumption:
    - Either through the informational or probabilistic formulation, or both.
    - Problem: Drug transporter proteins were not constrained by evolution (similar to nutriment transport in and out of cells)
2. Many models do not take into account the 3d structure of proteins:
    - A few works used graph neural networks, did not show very big improvements over e.g. CNN
3. Deep mutational scanning data are promising
    - Problem:
        - lack of standardization of experimental methods, i.e. output are not always numerically comparable.
    - Advantages:
        - Can provide information on specific phenotype,
        - Train against quantitative functional scores

Conclusions

① We have seen what research in biology focuses on: understanding the processes that govern life

② We have seen how AI systems can be used to help research in biology

- i.e. have the potential to provide conclusive answers to difficult questions by integrating the large amount of data generated by technological advances

③ Certain models are at the forefront of research: graph neural networks, transformers, VAEs

④ Engineering efficient analysis pipelines remains important to harness powerful models and large datasets

Thank you !

SciLifeLab

KAROLINSKA INSTITUTET ANNO 1810
Karolinska
Institutet

WASP
WALLENBERG AI,
AUTONOMOUS SYSTEMS
AND SOFTWARE PROGRAM

References I

[1] V. Siljehav, A. M. Hofstetter, K. Leifsdottir, and E. Herlenius, “Prostaglandin E2 Mediates Cardiorespiratory Disturbances during Infection in Neonates,” vol. 167, no. 6, pp. 1207–1213.e3.

[2] A. Honoré, D. Forsberg, K. Adolphson, S. Chatterjee, K. Jost, and E. Herlenius, “Vital sign-based detection of sepsis in neonates using machine learning,” vol. n/a, no. n/a.

[3] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, “Deep generative models of genetic variation capture the effects of mutations,” vol. 15, no. 10, pp. 816–822.

[4] Y. Zhou, R. Tremmel, E. Schaeffeler, M. Schwab, and V. M. Lauschke, “Challenges and opportunities associated with rare-variant pharmacogenomics,” vol. 43, no. 10, pp. 852–865.

[5] A. G. Roberts, “The Structure and Mechanism of Drug Transporters,” vol. 2342, pp. 193–234.

References II

[6] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," vol. 188, no. 3, pp. 415–431.

[7] T. D. Schneider and R. M. Stephens, "Sequence logos: A new way to display consensus sequences," vol. 18, no. 20, pp. 6097–6100.

[8] P. Notin, L. V. Niekerk, A. W. Kollasch, D. Ritter, Y. Gal, and D. S. Marks, "TranceptEVE: Combining Family-specific and Family-agnostic Models of Protein Sequences for Improved Fitness Prediction."

[9] P. Notin, A. W. Kollasch, D. Ritter, L. V. Niekerk, S. Paul, H. Spinner, N. J. Rollins, A. Shaw, R. Orenbuch, R. Weitzman, J. Frazer, M. Dias, D. Franceschi, Y. Gal, and D. S. Marks, "ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design,"

[10] P. C. Ng and S. Henikoff, "Predicting Deleterious Amino Acid Substitutions," vol. 11, no. 5, pp. 863–874.

References III

[11] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," vol. 118, no. 15, p. e2016239118.

[12] P. Notin, M. Dias, J. Frazer, J. M. Hurtado, A. N. Gomez, D. Marks, and Y. Gal, "Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval," in *Proceedings of the 39th International Conference on Machine Learning*, pp. 16990–17017, PMLR.