In biology, a **gene** is a sequence of nucleotides in DNA or RNA that codes for a molecule that has a function. The **genotype** is the part of the genetic makeup of a cell, and therefore of any individual, which determines one of its characteristics (phenotype). The **genotype–phenotype distinction** is drawn in genetics. "Genotype" is an organism's full hereditary information. "Phenotype" is an organism's actual observed properties, such as morphology, development, or behavior. An **allele** is a variant form of a given gene. A **heterozygous** individual is someone who has two different alleles at a locus. A **homozygous** individual has two identical alleles at a locus.

1 Support Vector Machines and Kernels for Computational Biology

RNA splicing, in molecular biology, is a form of RNA processing in which a newly made precursor messenger RNA (pre-mRNA) transcript is transformed into a mature messenger RNA (mRNA).

Kernel trick: Scalar product in feature space can be computed in input space. Common kernels: polynomial, sigmoid, RBF, normalization... Kernels allow to encode application-specific knowledge. Many kernels for different applications available.

String kernel SVMs capable of efficiently dealing with large k-mers k ¿ 10.

Spectrum kernel: position-independent motifs. **Spectrum Kernel with Mismatches**: Do not enforce strictly exact matches. **Weighted-degree kernel**: position-dependent motifs. As weighting use $\beta_k = 2\frac{d-k+1}{d(d+1)}$, where d is the maximal match length taken into account. This way the longer matches are weighted less, but they imply many shorter matches. **Weighted Degree Kernel with Shifts**: partially position-dependent motifs

SVM scoring function: SVM decision function is α-weighting of training points, but we are interested in weights of features. We can explicitly compute w and use it to rank importance. Explicit representation of w allows (some) interpretation. SVM-w does not reflect the score for a motif as substrings and overlapping strings contribute, too!

Positional Oligomer Importance Matrices (POIMs)

- Given k-mer z at position j in the sequence, compute expected score $\mathbb{E}[s(x) \mid x[j] = z]$ (for small k)
- Normalize with expected score over all sequences
- For large k use differential POIM

A **sequence logo** consists of a stack of letters at each position. The relative sizes of the letters indicate their frequency in the sequences. The total height of the letters depicts the information content of the position, in bits. The lowest order POIM (k=1) essentially conveys the same information as is represented in a sequence logo. However, unlike sequence logos, POIMs naturally generalize to higher order nucleotide patterns.

A position weight matrix (PWM), also known as a position-specific weight matrix (PSWM) or **position-specific scoring matrix (PSSM)**, is a commonly used representation of motifs (patterns) in biological sequences. PWMs are often derived from a set of aligned sequences that are thought to be functionally related and have become an important part of many software tools for computational motif discovery.

A PWM has one row for each symbol of the alphabet: 4 rows for nucleotides in DNA sequences or 20 rows for amino acids in protein sequences. It also has one column for each position in the pattern. In the first step in constructing a PWM, a basic position frequency matrix (PFM) is created by counting the occurrences of each nucleotide at each position. From the PFM, a position probability matrix (PPM) can now be created by dividing that former nucleotide count at each position by the number of sequences, thereby normalising the values. Most often the elements in PWMs are calculated as log likelihoods.

2 Biomedical Natural Language Processing

Term frequency (TF): the raw count of a term in a document: the number of times that term t occurs in document d. TF suffers from a critical problem: all terms are considered equally important. In fact certain terms have little or no discriminating power in determining relevance. Basic formulation is $f_{t,d}/\sum_{t'} f_{t',d}$

Document Frequency (DF): the number of documents in the collection that contain a term t. Basic formulation of IDF is $\log N/n_t$

In information retrieval, tf–idf or TFIDF, short for **term frequency–inverse document frequency**, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling.

Brill tagger (Transformation-Based Learning): a transformation-based process, in the sense that a tag is assigned to each word and changed using a set of predefined rules. In the transformation process, if the word is

known, it first assigns the most frequent tag, or if the word is unknown, it naively assigns the tag "noun" to it. Applying over and over these rules, changing the incorrect tags, a quite high accuracy is achieved.

Why use embeddings? Reduce dimensionality of representation. Encodes similarity information, useful for other tasks. Learn representations of entities (words) as well as relationships between them.

Word2Vec: Train a classifier on a binary prediction task of words occurring in the neighbourhoods of other words, take the learned classifier weights as the word embeddings. Faster and can easily incorporate a new sentence/document or add a word to the vocabulary. **Continuous Bag-of-Words (CBoW) model**: predict center word from sum of surrounding word vectors. **Skip-gram model**: predicting surrounding single words from center word (see NLU notes). Take the target word and a neighboring context word as positive examples, randomly sample other words in the lexicon to get negative samples. Normalized dot-product gives **cosine similarity**. This means we maximise the overlap (via dot product) between a word and the context it appeared in. By transitivity, any other word with a similar context will have a large overlap with the original word. For example, jumps $\sim$ leaps because their context vectors are similar.

Combining embeddings with prior knowledge: from analogical reasoning, abstract relationships were translations in the embedded space. Take this idea and extend the concept of context to include "appears in a relationship with" alongside "appears in a sentence with" and represent these new context-relationships as arbitrary affine transformations (basically, matrices).

Enforcing similarity: define an energy function $\mathcal{E}(S, R, T)$, energy is low if S is related to T through R is true (R is often non-symmetric). An example energy function is $\mathcal{E}(S, R, T \mid \theta) = -\frac{\mathbf{v}_T \cdot G_R \mathbf{c}_S}{||\mathbf{v}_T|| ||G_R \mathbf{c}_S||}$

"Off-task" data helps due to shared semantic information.

3 Time Series Analysis

Auto-regressive model are suited for stationary time series. A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary (i.e., "stationarized") through the use of mathematical transformations.

Classification of AR models $AR(p)$: observed, continuous, RNN: hidden, continuous, Markov chain: observed, discrete, HMM: hidden, discrete

4 Survival Analysis

The **log-rank test** compares the survival times of two or more groups. The null hypothesis for a log-rank test is that the groups have the same survival.

Censoring:

- Left censoring – a data point is below a certain value but it is unknown by how much.
- Interval censoring – a data point is somewhere on an interval between two values.
- Right censoring – a data point is above a certain value but it is unknown by how much.
- Type I censoring occurs if an experiment has a set number of subjects or items and stops the experiment at a predetermined time, at which point any subjects remaining are right-censored.
- Type II censoring occurs if an experiment has a set number of subjects or items and stops the experiment when a predetermined number are observed to have failed; the remaining subjects are then right-censored.
- Random (or non-informative) censoring is when each subject has a censoring time that is statistically independent of their failure time. The observed value is the minimum of the censoring and failure times; subjects whose failure time is greater than their censoring time are right-censored.

- The lifetime distribution function, conventionally denoted F, is defined as the complement of the survival function: $F(t) = P(T \leq t) = 1 - S(t)$
- If F is differentiable then the derivative, which is the density function of the lifetime distribution, is conventionally denoted f: $f(t) = F'(t) = \frac{d}{dt}F(t)$
- The function f is sometimes called the event density; it is the rate of death or failure events per unit time.

The hazard function, $h(t)$, is the instantaneous rate at which events occur, given no previous events. (The hazard function, conventionally denoted λ , is defined as the event rate at time t conditional on survival until time t or later (that is, $T \geq t$).)

Proportional hazards models are a class of survival models in statistics. In a proportional hazards model, the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate. For example, taking a drug may halve one's hazard rate for a stroke occurring, or, changing the material from which a manufactured component is constructed may double its hazard rate for failure.

Kaplan-Meier curves and log-rank tests are most useful when the predictor variable is categorical (e.g., drug vs. placebo), or takes a small number of values (e.g., drug doses 0, 20, 50, and 100 mg/day) that can be treated as categorical. The log-rank test and KM curves don't work easily with quantitative predictors such as gene expression, white blood count, or age. For quantitative predictor variables, an alternative method is **Cox proportional hazards regression analysis**.

5 Privacy Preserving Methods for ML in Healthcare

Anonymization refers to irreversibly severing a data set from the identity of the data contributor in a study to prevent any future re-identification, even by the study organizers under any condition. There's no re-identification of anonymized records, because the links back to the subjects are irreversibly broken. De-identification is also a severing of a data set from the identity of the data contributor, but may include preserving identifying information which can only be re-linked by a trusted party in certain situations.

Why generate synthetic data?

- Data could be shared and published without privacy concerns (e.g. scientific reproducibility)
- Data can be used to augment or enrich similar datasets
- Represents an alternative approach to build predictive systems
- Can benefit medical community for use in medical training simulator

TRTS is not as interesting as the TSTR case as it cannot diagnose mode collapse. Evaluating synthetic datasets: Mechanical Turks when no domain knowledge is needed, Inception score for images

Differential privacy addresses the case when a trusted data curator wants to release some statistic over its data without revealing information about a particular value itself. It is a constraint on the algorithms used to publish aggregate information about a statistical database which limits the privacy impact on individuals whose information is in the database. Roughly, an algorithm is differentially private if an observer seeing its output cannot tell if a particular individual's information was used in the computation. The most general mechanism is known as the Laplace mechanism, which adds Laplace noise to data so that everything an adversary receives becomes noisy and imprecise, and so it is much more difficult to breach privacy (if it is feasible at all). Challenges in DP: The more information you intend to "ask" of your database, the more noise has to be injected in order to minimize the privacy leakage. Once data has been leaked, it's gone. The total allowed leakage is often referred to as a "privacy budget", and it determines how many queries will be allowed (and how accurate the results will be). "Estimation from repeated queries" is also one of the fundamental limitations of differential privacy.

6 Interpretability of ML Models

Random forests tries to improve on bagging by "de- correlating" the trees.

Sensitivity Analysis of Individual Variables: It is a global explanation method. We examine what impact each feature has on the model's prediction. Possible transformations that can be done during analysis are sampling uniformly from the feature distribution, permutation of the feature values, replacing the values by mean or zero.

Mean Decrease in Impurity (MDI), (also called as Gini Importance) is defined as the total decrease in node impurity (weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that node)) averaged over all trees of the ensemble.

Standard Encoder-Decoder framework: q, f, g nonlinear functions, s_t is the hidden state of the decoder RNN. Here, the context vector c is the same for $\forall y_t$.

Encoder-Decoder with Bahdanau attention: a is so-called alignment model, jointly trained with all other components. Unlike existing encoder-decoder models, probability in g is conditioned on a distinct context vector c_i for each target word y_i. Probability α_{ij} (or energy e_{ij} for that reason) reflects the importance of the annotation h_j w.r.t. previous hidden state s_{i-1} in deciding the next state s_i and generating y_i.

Show, attend and tell (two mechanism for obtaining context vectors from annotation vectors):

- Hard (stochastic) attention: returns a sample from every point in time, based upon a categorical distribution (of locations) parametrized by α

- Soft attention: takes the expectation of the context vector directly

7 Appendix

- sensitivity, recall, hit rate, or true positive rate (**TPR**)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

- specificity, selectivity or true negative rate (**TNR**)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{N} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

- precision, positive predictive value (**PPV**)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$$

- negative predictive value (**NPV**)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = 1 - \mathrm{FOR}$$

- miss rate, false negative rate (**FNR**)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{P} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

- fall-out, false positive rate (**FPR**)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

- false discovery rate (**FDR**)

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$

- false omission rate (**FOR**)

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$

- accuracy (**ACC**)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

- **F1 score**

$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

It can be more flexible to predict probabilities of an observation belonging to each class in a classification problem rather than predicting classes directly. This is required when using models where the cost of one error outweighs the cost of other types of errors. For example, in a smog prediction system, we may be far more concerned with having low false negatives than low false positives. A false negative would mean not warning about a smog day when in fact it is a high smog day, leading to health issues in the public that are unable to take precautions. A false positive means the public would take precautionary measures when they didn't need to.

The **ROC curve** is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease. Area under ROC Curve (AUROC) is robust to imbalanced classes (for example, mortality has 2% positive examples).

Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. ROC curves are appropriate when the observations are balanced between each class, whereas **precision-recall curves** are appropriate for imbalanced datasets. Key to the calculation of precision and recall is that the calculations do not make use of the true negatives. It is only concerned with the correct prediction of the minority class, class 1. Area under Precision-Recall Curve (AUPRC) quantifies the tradeoff between sensitivity and false discovery, which is relevant in a clinical setting.

One reason to use the **logarithmic scale** is to respond to skewness towards large values; i.e., cases in which one or a few points are much larger than the bulk of the data. log scales allow a large range to be displayed without small values being compressed down into bottom of the graph. Another reason is to show percent change or multiplicative factors. In linear scale, even if the performance in percentage terms has been fairly constant a graph of the funds will appear to have grown most rapidly at the right hand end. With a logarithmic scale a constant percentage change is seen as a constant vertical distance so a constant growth rate is seen as a straight line. That is often a substantial advantage. In short, a logarithmic axis linearizes compound interest and exponential growth. A logarithmic axis is useful for plotting ratios. Ratios are intrinsically asymmetrical, but ratios are symmetrical on a log scale.

Digital phenotyping is a multidisciplinary field of science, defined as the "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices", in particular smartphones.