deepinfer - Bayesian approach for pre-calculating disease probabilities of genetic variants

2024-05-27

deepinfer - Bayesian approach for pre-calculating disease probabilities of genetic variants

To pre-calculate the probability of every possible variant causing any set of diseases using a Bayesian approach. The goal is to integrate known priors and estimated unknowns into a cohesive Bayesian framework.

Strategy Overview

1. **Variant Data Collection and Annotation**:
   - **Variant Identification**: Collect all known nucleotide variants and complex variants.
   - **Variant Annotation**: Use tools like `VariantAnnotation` to annotate each variant with functional consequences.
2. **Probability Estimation**:
   - **Frequency Estimation**: Estimate the probability of each variant based on true frequency in the population using databases such as gnomAD.
   - **Random Variant Estimation**: Estimate the probability of random, novel variants.
3. **Prior Information Incorporation**:
   - **Biological Data Integration**: Integrate biological data (clinical pathogenicity, protein structure, etc.) as priors.
   - **Bayesian Priors**: Utilize priors from databases like ClinVar for pathogenicity and structural data from UniProt.
4. **Bayesian Inference**:
   - **Disease Probability Calculation**: Apply Bayesian methods to update the probability of each variant causing specific diseases.
   - **Posterior Probability Calculation**: Use the Bayes theorem to combine prior knowledge and observed data.
5. **Tools and Packages in R**:
   - **VariantAnnotation**: For annotating genetic variants.
   - **gnomAD**: To get population frequency of variants.
   - **ClinVar**: To get clinical significance of variants.
   - **UniProt**: To get protein structural information.
   - **brms**: For Bayesian regression modeling.
   - **rstan**: For Bayesian inference using Stan.

```
# Load necessary packages
# if (!require("BiocManager", quietly = TRUE))
    # install.packages("BiocManager")
# BiocManager::install(c("VariantAnnotation"))
# BiocManager::install(c("UniProt.ws"))
# install.packages(c("rstan"))
# install.packages(c("brms"))
```

```
library(VariantAnnotation)
library(UniProt.ws)
library(brms)
library(rstan)
# gnomAD data
# ClinVar data
```

Step-by-Step Methodology

1. **Data Collection and Annotation**:

```
# Load variants
vcf <- readVcf("variants.vcf")

# Annotate variants
annotateVariants <- function(vcf) {
  return(vcf)
}
ann <- annotateVariants(vcf)
```

2. **Probability Estimation**:

```
# get frequencies
get_gnomAD_frequencies <- function(ann) {
  # Simulate frequencies for illustration
  freq <- runif(length(ann), min = 0, max = 0.01)
  return(freq)
}
estimate_random_variants <- function(ann) {
  # Simulate random variant probabilities
  random_prob <- runif(length(ann), min = 0, max = 0.001)
  return(random_prob)
}

# Frequency estimation from gnomAD
freq <- get_gnomAD_frequencies(ann)

# Random variant estimation
random_prob <- estimate_random_variants(ann)
```

3. **Incorporate Prior Information**:

```
get_clinvar_significance <- function(ann) {
  # Simulate clinical significance for illustration
  clin_significance <- sample(c("Pathogenic", "Benign", "VUS"), length(ann), replace = TRUE)
  return(clin_significance)
}
get_uniprot_data <- function(ann) {
  # Simulate UniProt data for illustration
  uniprot_data <- data.frame(
    protein_id = sample(letters, length(ann), replace = TRUE),
    structure_info = sample(c("Helix", "Sheet", "Loop"), length(ann), replace = TRUE)
  )
  return(uniprot_data)
}
```

```
# Clinical significance from ClinVar
clin_significance <- get_clinvar_significance(ann)

# Protein structure from UniProt
uniprot_data <- get_uniprot_data(ann)
```

4. **Bayesian Inference**:

```
library(brms)
library(rstan)

annotated_data <- data.frame(
  variant = sample(1:100, 100, replace = TRUE),
  gene = sample(letters, 100, replace = TRUE),
  disease = sample(c(0, 1), 100, replace = TRUE)
)

# Define the prior distributions
prior <- c(
  set_prior("normal(0, 1)", class = "b"),
  set_prior("cauchy(0, 2)", class = "sd")
)

# Fit a Bayesian model
fit <- brm(
  formula = disease ~ variant + (1|gene),
  data = annotated_data,
  family = bernoulli(),
  prior = prior,
  chains = 4, iter = 2000
)

# Extract posterior probabilities
posterior <- posterior_samples(fit)
```

5. **Posterior Probability Calculation**:

```
calculate_posterior_probabilities <- function(posterior) {
  post_prob <- apply(posterior, 2, mean)
  return(post_prob)
}

# Calculate the posterior probability for each variant-disease pair
post_prob <- calculate_posterior_probabilities(posterior)

# Output the results
write.csv(post_prob, "posterior_probabilities.csv")
```

Summary

This methodology outlines a comprehensive Bayesian approach to estimate the probability of genetic variants causing specific diseases. By integrating population frequencies, biological data, and clinical annotations into a Bayesian framework, we can pre-calculate disease probabilities for known and novel variants.

We integrate of various data sources to provide robust estimates of disease probabilities associated with genetic variants.