Predicting the functional consequences of amino acid polymorphisms using hierarchical Bayesian models

C. Verzilli[1,*], J. Whittaker[1], N. Stallard[2], D. Chasman[3]

[1]Dept. of Epidemiology and Public Health, Imperial College London, London, UK

[2]Medical and Pharmaceutical Statistics Research Unit, The University of Reading, Reading, UK

[3]Variagenics/Nuvelo, Cambridge, MA, USA

* Funded by the Wellcome Trust

Contents

- Protein structure and amino acid polymorphisms

Contents

- Protein structure and amino acid polymorphisms
- Data from Lac Repressor and Lysozyme mutagenesis experiments

Contents

- Protein structure and amino acid polymorphisms
- Data from Lac Repressor and Lysozyme mutagenesis experiments
- Modelling

Contents

- Protein structure and amino acid polymorphisms
- Data from Lac Repressor and Lysozyme mutagenesis experiments
- Modelling
- Results

< > - +

Contents

- Protein structure and amino acid polymorphisms
- Data from Lac Repressor and Lysozyme mutagenesis experiments
- Modelling
- Results
- Discussion

Protein structure and AA polymorphisms

Central dogma:

DNA → RNA → protein

Protein structure and AA polymorphisms

Central dogma:

DNA → RNA → protein

Triplets of bases (codons) in coding regions of the DNA

... GTG CAC CTG ACT CCT GAG GAG ...

Protein structure and AA polymorphisms

Central dogma:

DNA → RNA → protein

Triplets of bases (codons) in coding regions of the DNA

... GTG CAC CTG ACT CCT GAG GAG ...

are translated to a sequence of amino acids in a protein

Protein structure and AA polymorphisms

Central dogma:

DNA → RNA → protein

Triplets of bases (codons) in coding regions of the DNA

... GTG CAC CTG ACT CCT GAG GAG ...

are translated to a sequence of amino acids in a protein

... Val His Leu Thr Pro Glu Glu ...

< > - +

Protein structure and AA polymorphisms

Central dogma:

DNA → RNA → protein

Triplets of bases (codons) in coding regions of the DNA

... GTG CAC CTG ACT CCT GAG GAG ...

are translated to a sequence of amino acids in a protein

... Val His Leu Thr Pro Glu Glu ...

which folds spontaneously to a three-dimensional structure.

Protein structure and AA polymorphisms

- There are 20 different possible amino acids (AA) that vary in their physico-chemical properties.

Protein structure and AA polymorphisms

- There are 20 different possible amino acids (AA) that vary in their physico-chemical properties.
- DNA variants that change amino acid sequence (*nsSNP*) may change protein structure and hence function

Protein structure and AA polymorphisms

- There are 20 different possible amino acids (AA) that vary in their physico-chemical properties.
- DNA variants that change amino acid sequence (*nsSNP*) may change protein structure and hence function
- Example: sickle cell anæmia:

Protein structure and AA polymorphisms

- There are 20 different possible amino acids (AA) that vary in their physico-chemical properties.
- DNA variants that change amino acid sequence (*nsSNP*) may change protein structure and hence function
- Example: sickle cell anæmia:
  - GAG (GLu) $\longrightarrow$ GTG (Val) mutation in β-globin

Protein structure and AA polymorphisms

- There are 20 different possible amino acids (AA) that vary in their physico-chemical properties.
- DNA variants that change amino acid sequence (*nsSNP*) may change protein structure and hence function
- Example: sickle cell anæmia:
  - GAG (GLu) $\longrightarrow$ GTG (Val) mutation in β-globin
  - introduces an hydrophobic patch on the surface of the molecule

Protein structure and AA polymorphisms

- There are 20 different possible amino acids (AA) that vary in their physico-chemical properties.
- DNA variants that change amino acid sequence (*nsSNP*) may change protein structure and hence function
- Example: sickle cell anæmia:
  - GAG (GLu) $\longrightarrow$ GTG (Val) mutation in β-globin
  - introduces an hydrophobic patch on the surface of the molecule
  - major changes to its properties.

Protein structure and AA polymorphisms

- There are 20 different possible amino acids (AA) that vary in their physico-chemical properties.
- DNA variants that change amino acid sequence (*nsSNP*) may change protein structure and hence function
- Example: sickle cell anæmia:
  - GAG (GLu) $\longrightarrow$ GTG (Val) mutation in β-globin
  - introduces an hydrophobic patch on the surface of the molecule
  - major changes to its properties.
- Any nsSNP which disrupts structure is a strong candidate in disease/pharmacogenetic studies

Objective

- Identify nsSNPs likely to disrupt function in a novel protein, based on training data where the functionality of nsSNPs is known.

Objective

- Identify nsSNPs likely to disrupt function in a novel protein, based on training data where the functionality of nsSNPs is known.
- Explanatory variables:

Objective

- Identify nsSNPs likely to disrupt function in a novel protein, based on training data where the functionality of nsSNPs is known.
- Explanatory variables:
  - *structural* data: hydrophobicity, relative B factor, surface accessibility of native amino acid.

Objective

- Identify nsSNPs likely to disrupt function in a novel protein, based on training data where the functionality of nsSNPs is known.
- Explanatory variables:
  - *structural* data: hydrophobicity, relative B factor, surface accessibility of native amino acid.
  - *sequence*-based data: conservation of native amino acid in table of multiple sequence alignment.

< > - +

Objective

- Lots of recent interest in this (Gunther *et al*, 2003; Stitziel *et al*, 2003; Wang and Moult, 2001; Terp *et al*, 2002; del Sol Mesa *et al*, 2003; Ng and Henikoff, 2002; Chasman and Adams, 2001; Sunyaev *et al*, 2000,2001; Saunders and Baker, 2002)

Objective

- Lots of recent interest in this (Gunther *et al*, 2003; Stitziel *et al*, 2003; Wang and Moult, 2001; Terp *et al*, 2002; del Sol Mesa *et al*, 2003; Ng and Henikoff, 2002; Chasman and Adams, 2001; Sunyaev *et al*, 2000,2001; Saunders and Baker, 2002)
- None of these use statistical models: we will build a probabilistic model for protein function.

Data

- Data from site-directed mutagenesis experiments on proteins with known 3D structures, specifically Lac repressor (Markiewicz *et al*, 1994) and Lysozyme (Rennell *et al* 1991) proteins.

Data

- Data from site-directed mutagenesis experiments on proteins with known 3D structures, specifically Lac repressor (Markiewicz *et al*, 1994) and Lysozyme (Rennell *et al* 1991) proteins.
  - the native, wild-type, amino acids are replaced, one at a time, by non-native amino acids.

Data

- Data from site-directed mutagenesis experiments on proteins with known 3D structures, specifically Lac repressor (Markiewicz *et al*, 1994) and Lysozyme (Rennell *et al* 1991) proteins.
    - the native, wild-type, amino acids are replaced, one at a time, by non-native amino acids.
    - Effect on protein functionality recorded.

Data

- Data from site-directed mutagenesis experiments on proteins with known 3D structures, specifically Lac repressor (Markiewicz *et al*, 1994) and Lysozyme (Rennell *et al* 1991) proteins.
  - the native, wild-type, amino acids are replaced, one at a time, by non-native amino acids.
  - Effect on protein functionality recorded.
- For our data:

< > - +

Data

- Data from site-directed mutagenesis experiments on proteins with known 3D structures, specifically Lac repressor (Markiewicz *et al*, 1994) and Lysozyme (Rennell *et al* 1991) proteins.
  - the native, wild-type, amino acids are replaced, one at a time, by non-native amino acids.
  - Effect on protein functionality recorded.
- For our data:
  - Roughly 12 substitutions per site

Data

- Data from site-directed mutagenesis experiments on proteins with known 3D structures, specifically Lac repressor (Markiewicz *et al*, 1994) and Lysozyme (Rennell *et al* 1991) proteins.
    - the native, wild-type, amino acids are replaced, one at a time, by non-native amino acids.
    - Effect on protein functionality recorded.
- For our data:
    - Roughly 12 substitutions per site
    - Outcome is binary indicator of functionality

Data

- Data from site-directed mutagenesis experiments on proteins with known 3D structures, specifically Lac repressor (Markiewicz *et al*, 1994) and Lysozyme (Rennell *et al* 1991) proteins.
  - the native, wild-type, amino acids are replaced, one at a time, by non-native amino acids.
  - Effect on protein functionality recorded.
- For our data:
  - Roughly 12 substitutions per site
  - Outcome is binary indicator of functionality
- We will train on the Lac repressor and validate on Lysozyme (and *vice-versa*).

Lac Repressor

- *Lac repressor* controls the synthesis of various enzymes.

Lac Repressor

- *Lac repressor* controls the synthesis of various enzymes.
    - In the absence of lactose it binds to the DNA double helix upstream of the genes that code for enzymes necessary for *E. coli* to use lactose as a source of energy, preventing their synthesis.

Lac Repressor

- *Lac repressor* controls the synthesis of various enzymes.
  - In the absence of lactose it binds to the DNA double helix upstream of the genes that code for enzymes necessary for *E. coli* to use lactose as a source of energy, preventing their synthesis.
  - Amino acids at 264 sites (out of 360) mutated to give total of 3245 observations.

Lac Repressor

- *Lac repressor* controls the synthesis of various enzymes.
  - In the absence of lactose it binds to the DNA double helix upstream of the genes that code for enzymes necessary for *E. coli* to use lactose as a source of energy, preventing their synthesis.
  - Amino acids at 264 sites (out of 360) mutated to give total of 3245 observations.
- *Lysozyme* molecule from T4 phage:

Lac Repressor

- *Lac repressor* controls the synthesis of various enzymes.
  - In the absence of lactose it binds to the DNA double helix upstream of the genes that code for enzymes necessary for *E. coli* to use lactose as a source of energy, preventing their synthesis.
  - Amino acids at 264 sites (out of 360) mutated to give total of 3245 observations.
- *Lysozyme* molecule from T4 phage:
  - synthesised from the phage DNA once it has infected a bacterium and digested the bacteria cell wall, allowing replicated copies of the phage to escape.

Lac Repressor

- *Lac repressor* controls the synthesis of various enzymes.
  - In the absence of lactose it binds to the DNA double helix upstream of the genes that code for enzymes necessary for *E. coli* to use lactose as a source of energy, preventing their synthesis.
  - Amino acids at 264 sites (out of 360) mutated to give total of 3245 observations.
- *Lysozyme* molecule from T4 phage:
  - synthesised from the phage DNA once it has infected a bacterium and digested the bacteria cell wall, allowing replicated copies of the phage to escape.
  - Amino acids at 143 sites (out of 162) mutated to give a total of 1632 observations.

Predictive features used

Feature	*Description*
Accessibility	Solvent accessible area of native AA
Relative accessibility	Accessibility relative to maximum accessibility in training set
Relative phylogenetic entropy	Normalised phylogenetic entropy of native AA
Neighbourhood rel. phylogenetic entropy	Phylogenetic entropy of structural neighbourhood of native AA
Relative *B*-factor	Normalised *B*-factor of native AA
Neighbourhood relative *B*-factor	Normalised *B*-factor of structural neighbourhood of native AA
Unusual AA	Mutant AA is not in phylogenetic profile
Buried charge	Mutant is charged AA at buried site
Turn breaking	Mutant AA occurs at glycine or proline in a turn
Helix breaking	Mutant AA occurs in helical region and involves glycine or proline
Conserved	Native AA is at conserved position in phylogenetic profile

Methods

- Prediction problem; lots of ’standard’ classification tools could be tried.

Methods

- Prediction problem; lots of ’standard’ classification tools could be tried.
- Logistic regression

Methods

- Prediction problem; lots of ’standard’ classification tools could be tried.
- Logistic regression
- Classification trees

Methods

- Prediction problem; lots of ’standard’ classification tools could be tried.
- Logistic regression
- Classification trees
- Support vector machines

Multivariate Adaptive Regression Splines

Extension of logistic regression, with logit$(p) = \eta$ where

$$\eta = \beta_0 + \sum_{k=1}^{K} \beta_k B_k(\mathbf{x})$$

with *basis functions* $B_k(\mathbf{x})$ defined as

$$B_k(\mathbf{x}) = \prod_{j=1}^{J_k} \left[s_{kj}(x_{w_{kj}} - t_{kj})\right]_+ \quad k = 1, \ldots, K$$

Multivariate Adaptive Regression Splines

Extension of logistic regression, with logit$(p) = \eta$ where

$$\eta = \beta_0 + \sum_{k=1}^{K} \beta_k B_k(\mathbf{x})$$

with *basis functions* $B_k(\mathbf{x})$ defined as

$$B_k(\mathbf{x}) = \prod_{\mathbf{j=1}}^{\mathbf{J_k}} \left[s_{kj} (x_{w_{kj}} - t_{kj}) \right]_+ \quad k = 1, \ldots, K$$

where $\mathbf{J_k}$ is the degree of interaction,

< > - +

Multivariate Adaptive Regression Splines

Extension of logistic regression, with logit$(p) = \eta$ where

$$\eta = \beta_0 + \sum_{k=1}^{K} \beta_k B_k(\mathbf{x})$$

with *basis functions* $B_k(\mathbf{x})$ defined as

$$B_k(\mathbf{x}) = \prod_{j=1}^{J_k} [s_{kj}(x_{w_{kj}} - t_{kj})]_+ \quad k = 1, \ldots, K$$

where J_k is the degree of interaction, $[\cdot]_+ = \mathbf{max}[\mathbf{0}, \cdot]_+$,

Multivariate Adaptive Regression Splines

Extension of logistic regression, with logit$(p) = \eta$ where

The linear predictor of a MARS model is

$$\eta = \beta_0 + \sum_{k=1}^{K} \beta_k B_k(\mathbf{x})$$

with *basis functions* $B_k(\mathbf{x})$ defined as

$$B_k(\mathbf{x}) = \prod_{j=1}^{J_k} [\mathbf{s_{kj}}(x_{w_{kj}} - t_{kj})]_+ \quad k = 1, \ldots, K$$

where J_k is the degree of interaction, $[\cdot]_+ = max[0, \cdot]_+$, $\mathbf{s_{kj}} \in \{\pm\mathbf{1}\}$,

Multivariate Adaptive Regression Splines

Extension of logistic regression, with logit$(p) = \eta$ where

$$\eta = \beta_0 + \sum_{k=1}^{K} \beta_k B_k(\mathbf{x})$$

with *basis functions* $B_k(\mathbf{x})$ defined as

$$B_k(\mathbf{x}) = \prod_{j=1}^{J_k} [s_{kj}(x_{\mathbf{w_{kj}}} - t_{kj})]_+ \quad k = 1, \ldots, K$$

where J_k is the degree of interaction, $[\cdot]_+ = max[0, \cdot]_+$, $s_{kj} \in \{\pm 1\}$, $\mathbf{w_{kj}}$ indexes the predictor included

< > - +

Multivariate Adaptive Regression Splines

Extension of logistic regression, with logit$(p) = \eta$ where

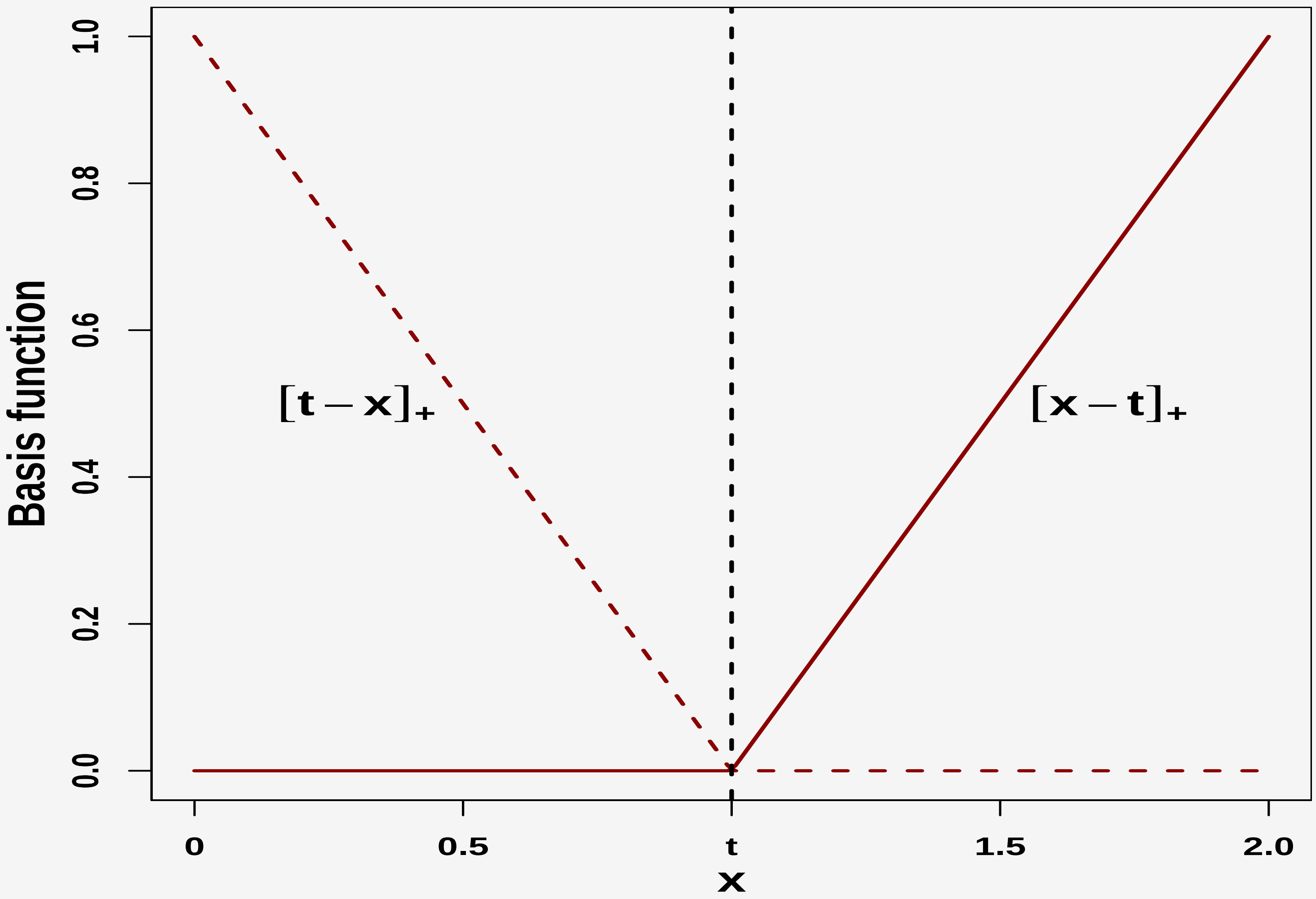
$$\eta = \beta_0 + \sum_{k=1}^{K} \beta_k B_k(\mathbf{x})$$

with *basis functions* $B_k(\mathbf{x})$ defined as

$$B_k(\mathbf{x}) = \prod_{j=1}^{J_k} [s_{kj}(x_{w_{kj}} - \mathbf{t_{kj}})]_+ \quad k = 1, \ldots, K$$

where J_k is the degree of interaction, $[\cdot]_+ = max[0, \cdot]_+$, $s_{kj} \in \{\pm 1\}$, w_{kj} indexes the predictor included and $\mathbf{t_{kj}}$ are knot points.

Multivariate Adaptive Regression Splines



Example of MARS basis functions

Methods

- However, note we have multiple observations on each site: therefore clustered data.

Methods

- However, note we have multiple observations on each site: therefore clustered data.
- Need a method that allows for this.

Methods

- However, note we have multiple observations on each site: therefore clustered data.
- Need a method that allows for this.
- We extend the Bayesian Multivariate Adaptive Regression Spline model (Holmes and Denison, 2003) to clustered training data.

Methods

- However, note we have multiple observations on each site: therefore clustered data.
- Need a method that allows for this.
- We extend the Bayesian Multivariate Adaptive Regression Spline model (Holmes and Denison, 2003) to clustered training data.
    - Accounts for clustering in the data

Methods

- However, note we have multiple observations on each site: therefore clustered data.
- Need a method that allows for this.
- We extend the Bayesian Multivariate Adaptive Regression Spline model (Holmes and Denison, 2003) to clustered training data.
  - Accounts for clustering in the data
  - Deals with potentially nonlinear effects of predictors

Methods

- However, note we have multiple observations on each site: therefore clustered data.
- Need a method that allows for this.
- We extend the Bayesian Multivariate Adaptive Regression Spline model (Holmes and Denison, 2003) to clustered training data.
  - Accounts for clustering in the data
  - Deals with potentially nonlinear effects of predictors
  - Uses Bayesian model averaging to make predictions

Methods

- However, note we have multiple observations on each site: therefore clustered data.
- Need a method that allows for this.
- We extend the Bayesian Multivariate Adaptive Regression Spline model (Holmes and Denison, 2003) to clustered training data.
  - Accounts for clustering in the data
  - Deals with potentially nonlinear effects of predictors
  - Uses Bayesian model averaging to make predictions
  - Fitted via MCMC

Hierarchical BMARS

- For amino acid site i and mutation m assume

$$p(y_{im} = 1|\boldsymbol{\beta}, \mathbf{x}_{im}, b_i) = \Phi\left(\beta_0 + \sum_{k=1}^{K} \beta_k B_k(\mathbf{x}_{im}) + b_i\right) = \Phi(\eta_{im}+b_i)$$

with $b_i \sim N(0, \sigma^2)$.

Hierarchical BMARS

- For amino acid site i and mutation m assume

  $$p(y_{im} = 1|\boldsymbol{\beta}, \mathbf{x}_{im}, b_i) = \Phi\left(\beta_0 + \sum_{k=1}^{K} \beta_k B_k(\mathbf{x}_{im}) + b_i\right) = \Phi(\eta_{im} + b_i)$$

  with $b_i \sim N(0, \sigma^2)$.
- Probit link for technical reasons—allows us to work out full conditionals for regression parameters and hence we can use gibbs sampling to update these

Hierarchical BMARS

- For amino acid site i and mutation m assume

$$p(y_{im}=1|\boldsymbol{\beta},\mathbf{x}_{im},b_i)=\Phi\left(\beta_0+\sum_{k=1}^{K}\beta_k B_k(\mathbf{x}_{im})+b_i\right)=\Phi(\eta_{im}+b_i)$$

  with $b_i \sim N(0,\sigma^2)$.
- Probit link for technical reasons—allows us to work out full conditionals for regression parameters and hence we can use gibbs sampling to update these
- Reversible jump MCMC to add, delete or modify a basis function at each iteration.

Application to mutagenesis data

- Fitted values given by

$$\hat{y}_{new} = I\left[\frac{1}{N}\sum_{t=1}^{N}\Phi(\mathbf{B}^{(t)}(\mathbf{x}_{new})\boldsymbol{\beta}^{(t)}) > \alpha\right].$$

Application to mutagenesis data

- Fitted values given by

$$\hat{y}_{new} = I\left[\frac{1}{N}\sum_{t=1}^{N}\Phi(\mathbf{B}^{(t)}(\mathbf{x}_{new})\boldsymbol{\beta}^{(t)}) > \alpha\right].$$

- Predictive performance assessed by calculating area under ROC curves, where ROC curves plot sensitivity versus specificity as α varies in $[0, 1]$.

Application to mutagenesis data

- Fitted values given by

$$\hat{y}_{new} = I\left[\frac{1}{N}\sum_{t=1}^{N}\Phi(\mathbf{B}^{(t)}(\mathbf{x}_{new})\boldsymbol{\beta}^{(t)}) > \alpha\right].$$

- Predictive performance assessed by calculating area under ROC curves, where ROC curves plot sensitivity versus specificity as α varies in $[0, 1]$.
  - sensitivity: proportion of mutations affecting function correctly classified

Application to mutagenesis data

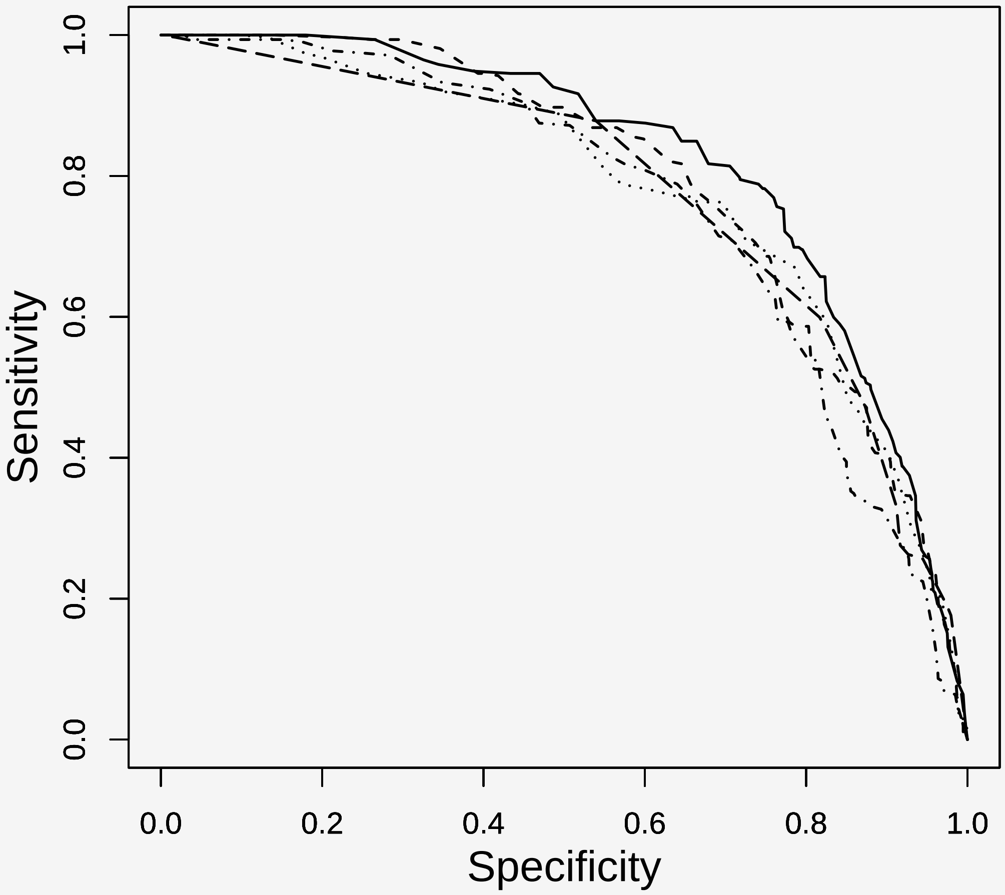
- Fitted values given by

$$\hat{y}_{new} = I\left[\frac{1}{N}\sum_{t=1}^{N}\Phi(\mathbf{B}^{(t)}(\mathbf{x}_{new})\boldsymbol{\beta}^{(t)}) > \alpha\right].$$

- Predictive performance assessed by calculating area under ROC curves, where ROC curves plot sensitivity versus specificity as α varies in $[0, 1]$.
  - sensitivity: proportion of mutations affecting function correctly classified
  - specificity: proportion of mutations *not* affecting function correctly classified

Application to mutagenesis data

Train on Lac repressor/ Test on Lysozyme

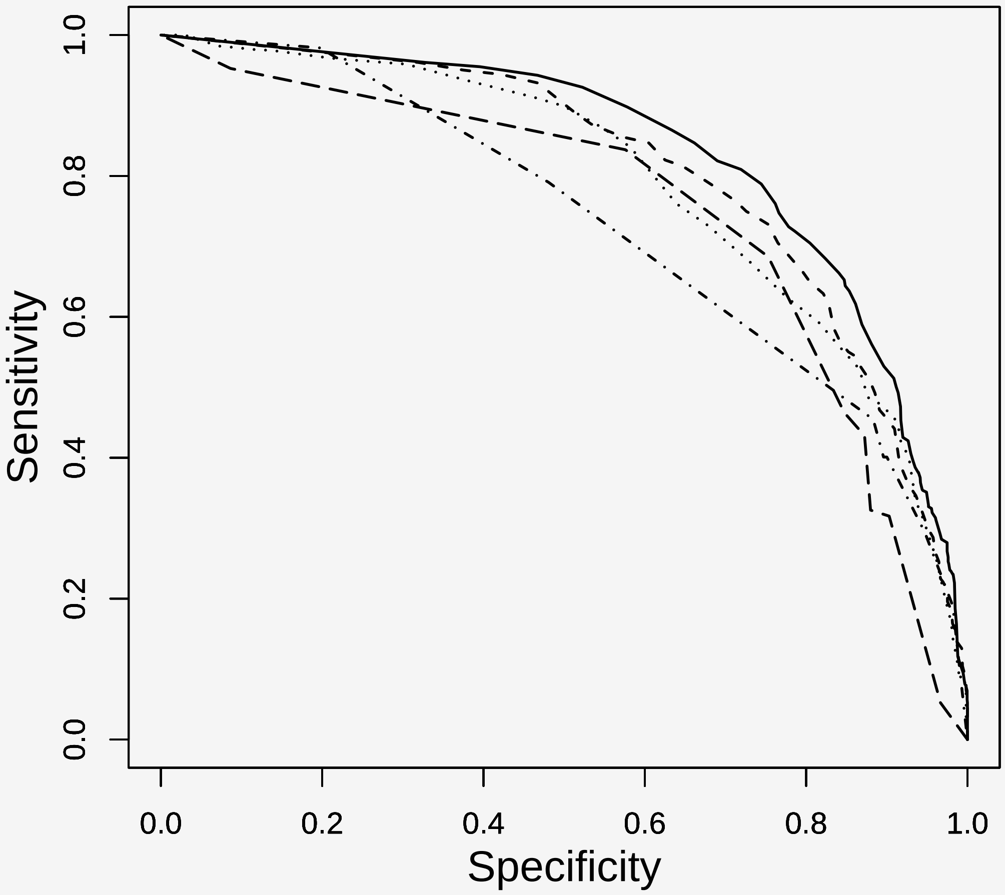


— H-BMARS (0.82) - - BMARS (0.79) · · · MARS (0.78)

- · - SVM (0.76) – – Tree (0.77)

Application to mutagenesis data

Train on Lysozyme/ Test on Lac repressor



— H-BMARS (0.83)   - - BMARS (0.80)   · · · MARS (0.79)

- · - SVM (0.73)   – – Tree (0.75)

Application to mutagenesis data: SVM caveat

- SVM used the radial kernel.

Application to mutagenesis data: SVM caveat

- SVM used the radial kernel.
- Need to tune this to control the smoothness of the decision boundary: 5 fold CV used.

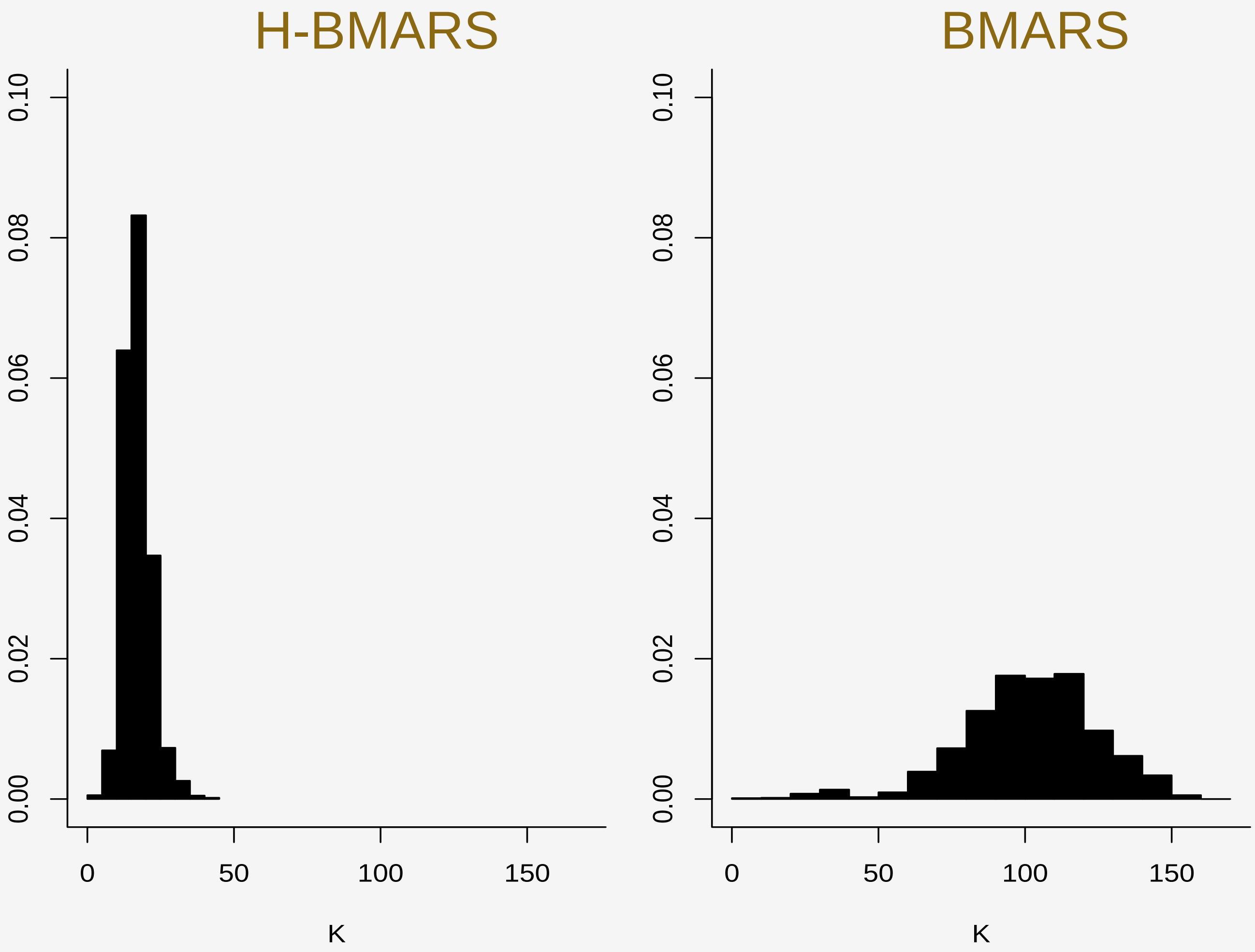
Application to mutagenesis data: SVM caveat

- SVM used the radial kernel.
- Need to tune this to control the smoothness of the decision boundary: 5 fold CV used.
- CV misclassification surface is rather flat: taking minimum gives poor performance on test data

Application to mutagenesis data: SVM caveat

- SVM used the radial kernel.
- Need to tune this to control the smoothness of the decision boundary: 5 fold CV used.
- CV misclassification surface is rather flat: taking minimum gives poor performance on test data
- Current analysis uses the smoothest decision boundary leading to acceptable CV misclassification rate

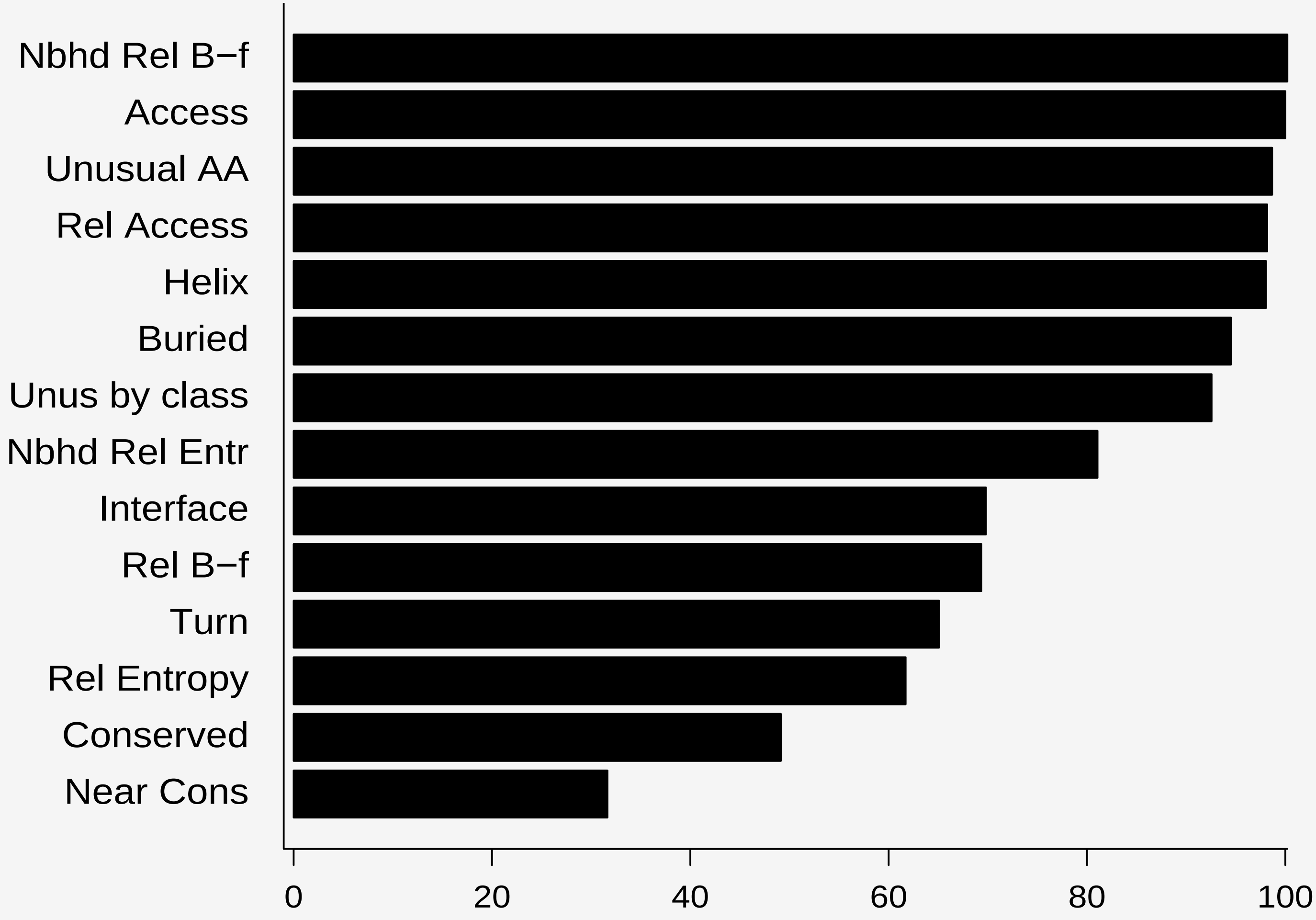
Application to mutagenesis data



Posterior distribution of the number of basis functions

< > - +

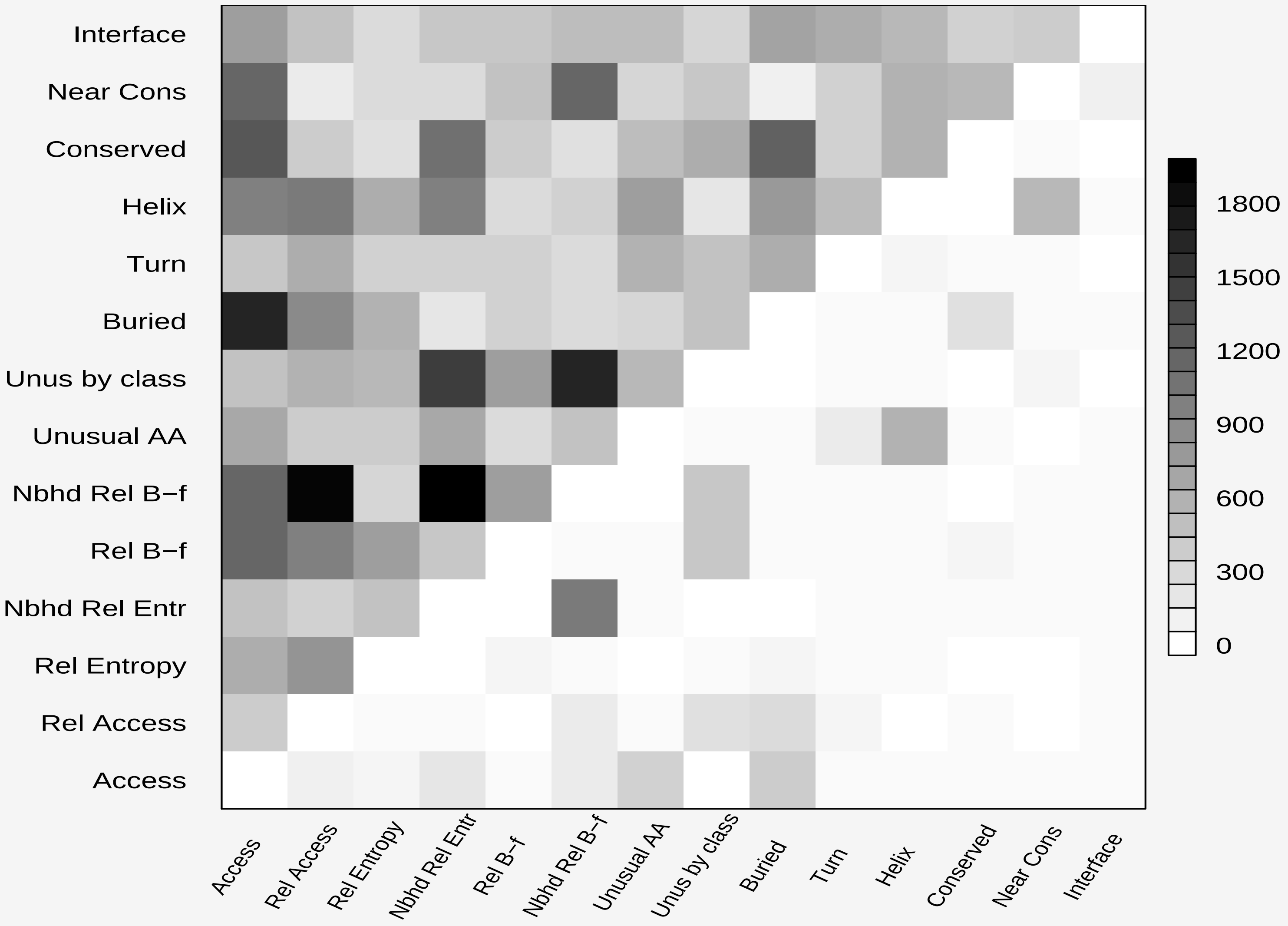
Application to mutagenesis data



Relative importance of predictors in the generated sample

< > - +

Application to mutagenesis data



Interaction terms in the generated sample

< > - +

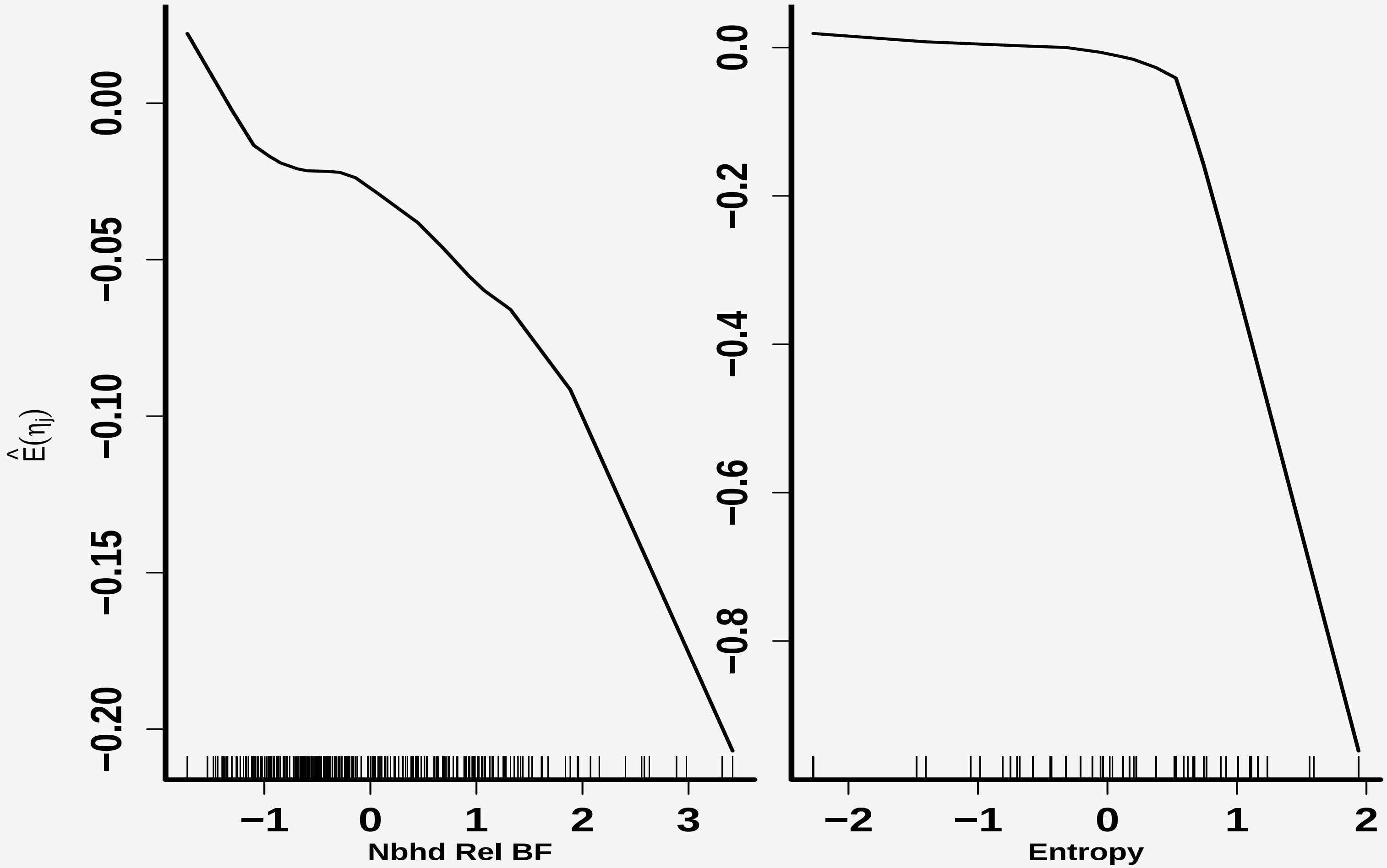
Application to mutagenesis data

The posterior main effect of generic predictor p may be quantified as

$$\hat{E}[\Phi_p(x)] = \frac{1}{L} \sum_{l=1}^{L} \sum_{\substack{k:J_k=1\\ w_{1k}=p}} \beta_k^{(l)} B_k^{(l)}(x)$$

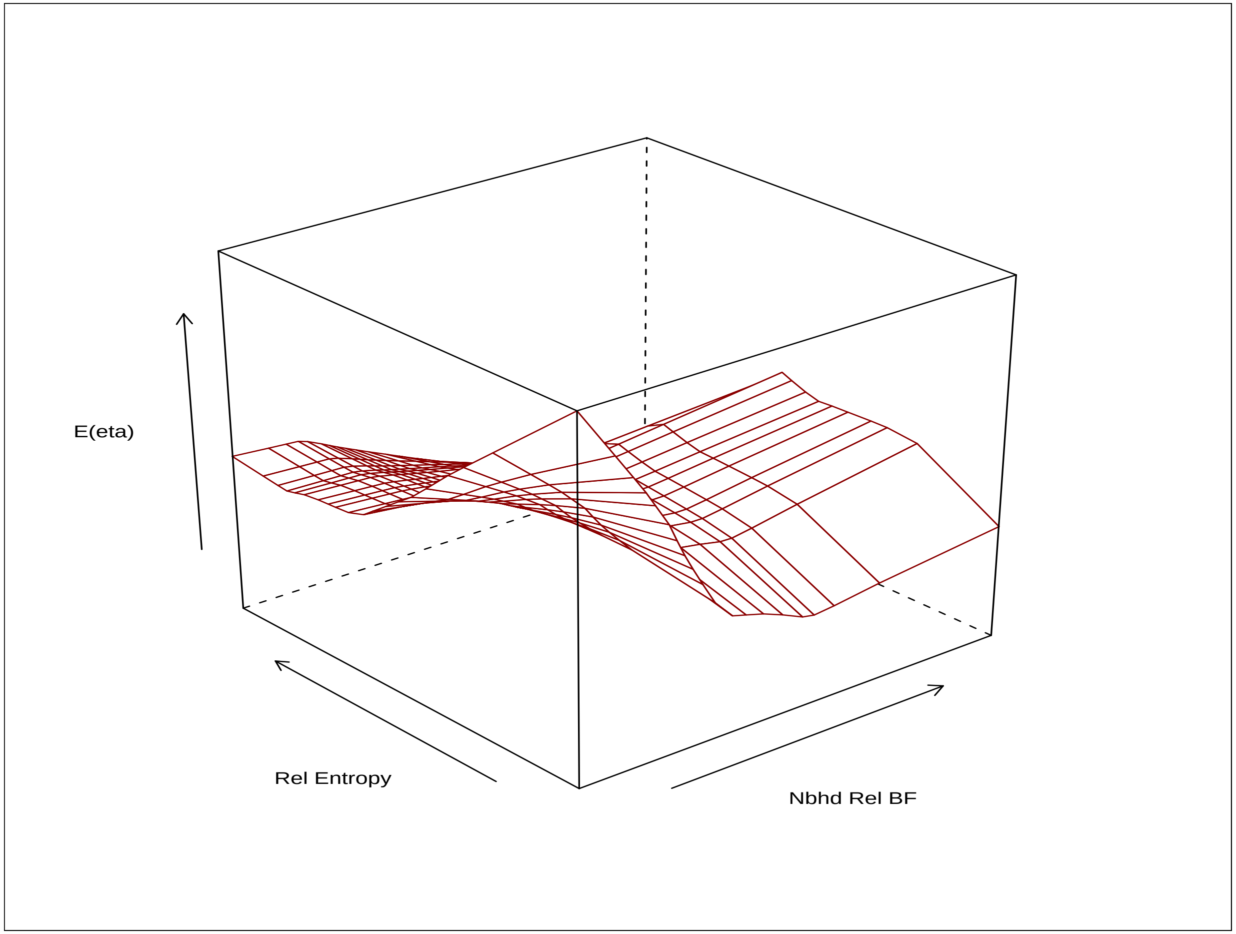
from a posterior sample of size L.

Application to mutagenesis data



Posterior mean main effects (H-BMARS)

Application to mutagenesis data



Posterior mean interaction between Entropy and Nbhd Rel BF

Summary

- Hierarchical BMARS achieves better out-of-sample specificity and sensitivity than competing methods here.

Summary

- Hierarchical BMARS achieves better out-of-sample specificity and sensitivity than competing methods here.
- Heterogeneous misclassification rates of less than 20% compared to $27 - 35\%$ reported by other authors using the same data.

< > - +

Summary

- Hierarchical BMARS achieves better out-of-sample specificity and sensitivity than competing methods here.
- Heterogeneous misclassification rates of less than 20% compared to $27 - 35\%$ reported by other authors using the same data.
- Acknowledging clustered nature of data leads to enhanced interpretability and improved mixing.

Summary

- Hierarchical BMARS achieves better out-of-sample specificity and sensitivity than competing methods here.
- Heterogeneous misclassification rates of less than 20% compared to $27 - 35\%$ reported by other authors using the same data.
- Acknowledging clustered nature of data leads to enhanced interpretability and improved mixing.
- Allowing higher degree of interaction gave very similar results

Summary

- Hierarchical BMARS achieves better out-of-sample specificity and sensitivity than competing methods here.
- Heterogeneous misclassification rates of less than 20% compared to $27 - 35\%$ reported by other authors using the same data.
- Acknowledging clustered nature of data leads to enhanced interpretability and improved mixing.
- Allowing higher degree of interaction gave very similar results
- Solvent accessibility and molecular rigidity (B-factor) are good predictors of functionality.

Summary

- Hierarchical BMARS achieves better out-of-sample specificity and sensitivity than competing methods here.
- Heterogeneous misclassification rates of less than 20% compared to $27 - 35\%$ reported by other authors using the same data.
- Acknowledging clustered nature of data leads to enhanced interpretability and improved mixing.
- Allowing higher degree of interaction gave very similar results
- Solvent accessibility and molecular rigidity (B-factor) are good predictors of functionality.
- Code available as R package.

Discussion

- Improved performance more likely to come from improved model inputs than more sophisticated modelling

Discussion

- Improved performance more likely to come from improved model inputs than more sophisticated modelling
    - In particular, need more mutant AA-specific covariates (as opposed to cluster-specific)

Discussion

- Improved performance more likely to come from improved model inputs than more sophisticated modelling
  - In particular, need more mutant AA-specific covariates (as opposed to cluster-specific)
  - Lysozyme very different to Lac repressor: consider prediction in specific protein families?

Discussion

- Improved performance more likely to come from improved model inputs than more sophisticated modelling
  - In particular, need more mutant AA-specific covariates (as opposed to cluster-specific)
  - Lysozyme very different to Lac repressor: consider prediction in specific protein families?
- How useful is this?

Discussion

- Improved performance more likely to come from improved model inputs than more sophisticated modelling
  - In particular, need more mutant AA-specific covariates (as opposed to cluster-specific)
  - Lysozyme very different to Lac repressor: consider prediction in specific protein families?
- How useful is this?
  - Predictions will be weak in general

Discussion

- Improved performance more likely to come from improved model inputs than more sophisticated modelling
  - In particular, need more mutant AA-specific covariates (as opposed to cluster-specific)
  - Lysozyme very different to Lac repressor: consider prediction in specific protein families?
- How useful is this?
  - Predictions will be weak in general
  - But functional biology is hard and expensive

Discussion

- Improved performance more likely to come from improved model inputs than more sophisticated modelling
  - In particular, need more mutant AA-specific covariates (as opposed to cluster-specific)
  - Lysozyme very different to Lac repressor: consider prediction in specific protein families?
- How useful is this?
  - Predictions will be weak in general
  - But functional biology is hard and expensive
  - Will nsSNPs disrupting function be interesting for genetic epidemiology/pharmacogenetics?