COMPARISON BETWEEN HEURISTIC AND STATISTICAL ANALYSIS ON PROTEIN STRUCTURAL PROPERTIES

E. Del Prete, S. Dotolo, A. Marabotti, A. Facchiano
eugenio.delprete@isa.cnr.it

15th International Workshop on Network Tools and Applications for Biology
11th Integrative Bioinformatics International Symposium

October 15th, 2015 Bari
Protein Structure

- Biological macromolecules
  - complex structural organization
  - balance of energetic factors

- Homology among different organisms
  - different sequence, same structure
  - amino acid substitution, different structure

- Structure-function relationships
  - some of them stronger than others

- Examining protein structure
  - analyzing conformational features

2CBR, A Chain, LCN, *Bos taurus*
(obtained with PyMol)
**Getting Data**

- **Protein families with different architectural classification**
  - 1. Beta-lactamase (BLA)
  - 2. Cathepsin B (CTS)
  - 3. Ferritin (FTL)
  - 4. Glycosyltransferase (GTF)
  - 5. Hemoglobin (HGB)
  - 6. Lipocalin 2 (LCN)
  - 7. Lysozyme (LYS)
  - 8. P. Cell Nuclear Antigen (PCNA)
  - 9. P. Nucleoside Phosphorylase (PNP)
  - 10. Superoxide Dismutase (SOD)

- **153 Crystallographic structures**
  - 2.40.128.x β-β barrel
Cleaning Data

- Similar number of structures, 13-19 for each family
  - only wild-type, one for organism
  - less than 50 residues about length

- Only one chain in homo-multimeric proteins
  - chain A where available (chain E in 1M73 and chain X for 3CH2)

- Structural-geometrical properties
  - secondary structure, hydrogen bonds, accessible surface areas, torsion angles,
    packing defects, charged residues, free energy of folding, volume, salt bridges

- Percentage features and standard score form
  - better stability in evaluations
EDA: Kernel Density Distribution

- **Non-parametric estimation of p.d.f.**
  - based on a finite data sample

- **Overcoming the histogram graph**
  - a more effective way to show the distribution of a variable

- **How variables are distributed**
  - for each protein family

\[
\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right)
\]

with \( h \approx 1.06 \frac{\hat{\sigma}}{\sqrt{n}} \)

(if \( K \) is a Gaussian distribution for univariate data)
**EDA: Correlation**

- **Pearson’s correlation coefficient** - graphical correlation matrix
  \[ \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \]

- **Partial correlation coefficient** - able to avoid the collinearity
  \[ \rho_{yz|x} = \frac{\rho_{yz} - \rho_{yx} \rho_{zx}}{\sqrt{1 - \rho^2_{yx}} \sqrt{1 - \rho^2_{zx}}} \]

- **Dissimilarity measurement** - Pearson’s distance
  \[ d_{xy} = 1 - |\rho_{xy}| \]

- **T-Student test for significance** - confidence level of 0.95
  \[ t = \rho \sqrt{\frac{n - 2}{1 - \rho^2}} \]
EDA: Principal Component Analysis

- Multivariate & unsupervised statistical method
  - compressed data, new relationships
  \[
  \begin{align*}
  PC_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1c}X_c \\
  PC_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2c}X_c \\
  \vdots \\
  PC_l &= a_{l1}X_1 + a_{l2}X_2 + \cdots + a_{lc}X_c
  \end{align*}
  \]

- Summarizing initial variables into new ones
  - semi-heuristic decision on variables number

- Clusterization and outlier detection
  - interpretation allowed to investigator
  \[
  \max_{a_m} \left\{ \frac{1}{r} \sum_{i=1}^{r} \left( \sum_{j=1}^{c} a_{1j}x_{ij} \right)^2 \right\} \quad \text{with} \quad \sum_{j=1}^{c} a_{1j}^2 = 1
  \]

- Sparse PCA, a hybrid technique with regression
  - not all the variables are in the PCs
  \[
  (\Sigma - \lambda_m I)a_m = 0
  \]
Classification: Variable Importance

- **Categorizing observations**
  - by means of predictive models for classification

- **Different algorithms used:**
  - random forest (RFO)
  - recursive partitioning (RPA)
  - stochastic gradient boosting (GBM)
  - boosting model (C50)
  - flexible discriminant analysis (FDA)
  - nearest shrunken centroid (NSC)

- **Different scores for variable importance estimation**
  - percentage of variables occurrence (ranking)

---

**Resampling:**
- **a)** Training set of 70% of data
- **b)** Testing set of 30% of data
- **c)** 10-fold cross validation
- **d)** Repeating 10 times
Classification: Performance

- **Accuracy**
  - proportion of true results among the total number of cases examined
  \[ ACC = \frac{TP + TN}{P + N} \]

- **Sensitivity**
  - proportion of positives that are correctly identified as such
  \[ TPR = \frac{TP}{P} \]

- **Specificity**
  - proportion of negatives that are correctly identified as such
  \[ TNR = \frac{TN}{N} \]

- **Kappa coefficient**
  - reliability of a statistical classification, related to the possible best classification
  \[ K = \frac{Pr(o) - Pr(e)}{1 - Pr(e)} \]
R Tools

- R environment in Rstudio IDE
  - user and developer
  - Comprehensive R Archive Network (CRAN) & Bioconductor

- corrplot, Hmisc, ppcor

- sparcl, GeneNet, caret

- lattice, ggplot2, directlabels

Classification and Regression Training:

- a) data splitting
- b) pre-processing
- c) feature selection
- d) model tuning using resampling
- e) variable importance estimation
Density Panels

- Features distributions regularity
  - unimodal and centered

- Some valuable results
  - ROC, ROA in FTL
  - T in PNP and GTF
  - RB95 in all the families

- A good overview
  - on protein families
  - on protein structural features
Institute of Food Sciences, National Research Council – Avellino (Italy)

E. Del Prete, S. Dotolo, A. Marabotti, A. Facchiano

October 15th, 2015 Bari

Introduction
Methods
Results
Conclusions

Dissimilarity Dendrogram

- A: Alpha Helix
- B: Beta Sheet
- C: Coil
- T: Turn
- RHB: Residue Hydrogen Bond
- NPA: Non Polar Accessible Surface Area
- PA: Polar Accessible Surface Area
- CA: Charge Accessible Surface Area
- MRA_N: Mean Accessible Surface Area Per Residue
- VOL_N: Total Volume
- RPC: PhiPsi Angles Core/Allowed/Generous/Outside
- ROX: Omega Angle Core/Allowed/Generous/Outside
- PD: Packing Defect
- FEF_N: Free Energy Folding
- RB95: Buried 95% BC: Buried Charge
- RSB: Residue Salt Bridge
Features Network

- **Two kind of relationships**
  - continuous line: partial correlation
  - dotted line: partial anticorrelation

- **Pruning excessive features**
  - peripheral ones

  a) Phi-Psi Angles
  b) Omega Angle
  c) ASA information

*Use of GGMs*
PCA & sPCA

- Clustering for protein families
  - GTF, FTL, CTS, …

- SOD is in a wide open position
  - multi-structural architecture

- (Possible) outliers detection
  - Pseudomonas putida SOD

- sPCA is coherent with the dendrogram
  - 60% of explained variance
  - sPCs are clusterized
## Features Classification

<table>
<thead>
<tr>
<th></th>
<th>RFO</th>
<th>RFO_S</th>
<th>RPA</th>
<th>RPA_S</th>
<th>GBM</th>
<th>GBM_S</th>
<th>C50</th>
<th>C50_S</th>
<th>FDA</th>
<th>FDA_S</th>
<th>NSC</th>
<th>NSC_S</th>
<th>PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1.00</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0.67</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>1.00</td>
</tr>
<tr>
<td>RHB</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td></td>
<td></td>
<td>0.83</td>
</tr>
<tr>
<td>NPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>PA</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>CA</td>
<td>11</td>
<td>12</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>MRA_N</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td>0.83</td>
</tr>
<tr>
<td>VOL_N</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>RPC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>RPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>RPA_S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>RPA_S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>RPO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>RPO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>ROC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>ROA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>ROA_S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>ROG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>ROG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>ROG_S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>ROO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>ROO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>PD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>FEF_N</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1.00</td>
</tr>
<tr>
<td>RB95</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>BC</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.58</td>
</tr>
<tr>
<td>RSB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.08</td>
</tr>
<tr>
<td>ACCURACY</td>
<td>0.95</td>
<td>0.95</td>
<td>0.59</td>
<td>0.60</td>
<td>0.94</td>
<td>0.95</td>
<td>0.89</td>
<td>0.88</td>
<td>0.95</td>
<td>0.94</td>
<td>0.89</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>KAPPA</td>
<td>0.95</td>
<td>0.94</td>
<td>0.54</td>
<td>0.56</td>
<td>0.93</td>
<td>0.94</td>
<td>0.88</td>
<td>0.86</td>
<td>0.94</td>
<td>0.93</td>
<td>0.88</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>SENSITIVITY</td>
<td>0.97</td>
<td>0.97</td>
<td>0.50</td>
<td>0.50</td>
<td>0.93</td>
<td>0.93</td>
<td>0.86</td>
<td>0.86</td>
<td>0.93</td>
<td>0.93</td>
<td>0.91</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>SPECIFICITY</td>
<td>0.99</td>
<td>0.99</td>
<td>0.95</td>
<td>0.95</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>PERFORMANCE</td>
<td>0.97</td>
<td>0.96</td>
<td>0.65</td>
<td>0.65</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.91</td>
<td>0.90</td>
<td>0.95</td>
<td>0.95</td>
<td>0.92</td>
<td>0.86</td>
</tr>
<tr>
<td>P-VALUE</td>
<td>1.00E-29</td>
<td>1.00E-29</td>
<td>8.3E-10</td>
<td>1.83E-10</td>
<td>1.46E-27</td>
<td>1.46E-27</td>
<td>5.05E-22</td>
<td>5.05E-22</td>
<td>1.46E-27</td>
<td>1.46E-27</td>
<td>1.37E-26</td>
<td>2.44E-17</td>
<td></td>
</tr>
</tbody>
</table>
Features «Selection»

- Comparing all the used techniques

- Variables subset as typical for the protein families dataset
  - A, B, C, T, RHB, PA, CA, MRA_N, VOL_N, RB95

- Free energy of folding (FEF_N) strictly related to volume
  - because of the prediction formula…

- Structural defects seem to influence the present study
  - not so strong in all the methods
Conclusions and Future Works

- Graphical multivariate procedures are good tools
  - characterization
  - fingerprints

- Predictive models for classification to perform feature selection
  - knowledge < information < data

- How to improve the work?
  - multivariate regression models
  - protein families number

a) Transglutaminase
b) Superoxide Dismutase
c) Glycosyl Hydrolase
Acknowledgments and Credits

- Flagship Project «InterOmics»

- Bioinformatics and Computational Biology Laboratory
  
  E. Del Prete, S. Dotolo, A. Facchiano

  Department of Chemistry and Biology, University of Salerno (Fisciano, Italy)

  A. Marabotti