BAYES THEOREM II (DA_2022)

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

DA_2022 Lecture n. 9, Rome 30th March 2022

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

outline

- Discussion of Bulashevska2008
  intrinsically disordered proteins
- Discussion of Puga2015a (Altman series) by Martina Roiati
- Confusion Matrices
- ROC curves
- jacknife

Concrete example of bayesian methodology

Journal of Theoretical Biology 254 (2008) 799–803

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Using Bayesian multinomial classifier to predict whether a given protein sequence is intrinsically disordered

Alla Bulashevska [a,*], Roland Eils [a,b]

[a] Department of Theoretical Bioinformatics, German Cancer Research Center, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[b] Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology (IPMB), University of Heidelberg, Germany

ARTICLE INFO

Article history:
Received 19 November 2007
Received in revised form
19 May 2008
Accepted 19 May 2008
Available online 14 June 2008

Keywords:
Unfolded proteins
Disorder prediction
Model-based classification
Multinomial model

ABSTRACT

Intrinsically disordered proteins (IDPs) lack a well-defined three-dimensional structure under physiological conditions. Intrinsic disorder is a common phenomenon, particularly in multicellular eukaryotes, and is responsible for important protein functions including regulation and signaling. Many disease-related proteins are likely to be intrinsically disordered or to have disordered regions. In this paper, a new predictor model based on the Bayesian classification methodology is introduced to predict for a given protein or protein region if it is intrinsically disordered or ordered using only its primary sequence. The method allows to incorporate length-dependent amino acid compositional differences of disordered regions by including separate statistical representations for short, middle and long disordered regions. The predictor was trained on the constructed data set of protein regions with known structural properties. In a Jack-knife test, the predictor achieved the sensitivity of 89.2% for disordered and 81.4% for ordered regions. Our method outperformed several reported predictors when evaluated on the previously published data set of Prilusky et al. [2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21 (16), 3435–3438]. Further strength of our approach is the ease of implementation.

© 2008 Elsevier Ltd. All rights reserved.

ABSTRACT

Intrinsically disordered proteins (IDPs) lack a well-defined three-dimensional structure under physiological conditions. Intrinsic disorder is a common phenomenon, particularly in multicellular eukaryotes, and is responsible for important protein functions including regulation and signaling. Many disease-related proteins are likely to be intrinsically disordered or to have disordered regions. In this paper, a new predictor model based on the Bayesian classification methodology is introduced to predict for a given protein or protein region if it is intrinsically disordered or ordered using only its primary sequence. The method allows to incorporate length-dependent amino acid compositional differences of disordered regions by including separate statistical representations for short, middle and long disordered regions. The predictor was trained on the constructed data set of protein regions with known structural properties. In a Jack-knife test, the predictor achieved the sensitivity of 89.2% for disordered and 81.4% for ordered regions. Our method outperformed several reported predictors when evaluated on the previously published data set of Prilusky et al. [2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21 (16), 3435–3438]. Further strength of our approach is the ease of implementation.

© 2008 Elsevier Ltd. All rights reserved.

https://www.disprot.org/about
https://protein.bio.unipd.it/projects
https://proteopedia.org/wiki/index.php/Main_Page

The focus of the paper

In this paper, we introduce a new prediction method, which exploits the Bayesian classification procedure to predict disordered property for a given protein or protein region from its primary sequence. Bayesian Markov chain model-based classification has already found its application in proteomics for the prediction of protein subcellular locations (Bulashevska and Eils, 2006). This approach represents each class with a single probabilistic summary. Since the AAC of disordered regions is distinct from that of ordered, we propose to use multinomial models for the description of class-conditional densities. The intuition behind this approach is that each protein sequence belonging to a certain class can be considered as a realization of an independent random process that emits symbols from an alphabet of 20 amino acids.

Original observation

i.e. depletion of the typically buried W, C, F, I, Y, V, L and enrichment of the typically exposed K, E, P, S, Q, R. However, some significant differences were also found. Short disordered regions are more depleted in C, I, V and L, while long disordered regions are more enriched in K, E and P but are less enriched in Q and S. In addition, long disordered regions are depleted in G and N, while short disordered regions are enriched in G and D. In order to

we model short ($\leqslant$30 residues), middle (31–100 residues) and long ($>$100 residues) disordered regions separately.

2.2. Multinomial models

Multinomial models assume a *bag-of-amino acid* sequence representation, which considers the appearance of each amino acid as an independent event. The order in which amino acids occur in a given amino acid sequence is ignored; the only information retained is a vector of counts $\mathbf{n} = (n_1, \ldots, n_{20})$, where n_i is the number of occurrences of amino acid i in the sequence.

We assume that the probability of a sequence s to come from a certain class c is given by a multinomial probability function governed by its vector of parameters $\theta_c = (\theta_{c1}, \ldots, \theta_{c20}) \in [0,1]^{20}$:

$$p(s|\theta_c) = \frac{n!}{\prod_{i=1}^{20} n_i!} \prod_{i=1}^{20} \theta_{ci}^{n_i}, \tag{1}$$

where $n = \sum_i n_i$ denotes the length of the sequence. The parameter θ_{ci} denotes the cth class-conditional probability of amino acid i to occur in a sequence. The parameters of the model corresponding to class c are estimated from the training regions belonging to the class c. Thus, the parameter θ_{ci} is calculated as

$$\theta_{ci} = \frac{n_{ci}}{\sum_{i=1}^{20} n_{ci}}, \tag{2}$$

where n_{ci} is the number of occurrences of amino acid i in the sequences of class c. This way of estimating parameters of the

2.3. Bayesian multinomial classifier

Bayesian classification is a widely applied method in the machine learning and statistical community, which is based on Bayes' theorem (Bayes rule). According to Bayes' rule, the class for an unlabeled sequence s can be inferred using the posterior probability:

$$p(c|s) = \frac{p(c)p(s|c)}{p(s)} = \frac{p(c)p(s|c)}{\sum_c p(c)p(s|c)}. \quad (3)$$

Note!

We assume class prior probabilities $p(c)$ to be equally distributed. We further assume that the sequences of each class are generated from multinomial models. Thus, given the parameters $\{\theta_c\}$ of the models for each class, the term $p(s|c)$ denoting the prior probability of a sequence s to belong to the class c can be computed using the formula (1) for $p(s|\theta_c)$ from previous subsection.

Since we model short, middle and long disordered regions separately, the estimation of the class-conditional densities involves four subproblems (for short, middle, long disordered and ordered classes), in which each of the class-conditional density is estimated based on the data belonging to the corresponding class only.

Bayesian classifier is a *probabilistic* classifier, which yields for each query instance the posterior probability for each class, a numeric value that represents the degree to which an instance is a member of a class. To produce a discrete output, the following decision rule is usually applied: the class should be the one which maximizes the posterior probability.

NOTE!

To classify an input sequence as disordered or ordered, we sum the posterior probabilities for short, middle and long disordered subtypes into a single value describing the posterior probability of a sequence to be disordered and then use the standard decision rule to come up with a discrete output, i.e. predict one of the two classes (disordered/ordered) showing the biggest posterior probability.

2.4. Performance evaluation

The prediction performance of our predictor was validated with *Jack-knife test* (or *leave-one-out cross-validation*) (Mardia et al., 1979). By Jack-knife test the learning step is performed with all training instances except the one for which the class is to be predicted.

The prediction quality was evaluated using the standard measures of *sensitivity* (SN) and *specificity* (SP), where the sensitivity, or *true positive rate*, is the percentage of disordered sequences correctly predicted, and the SP, or *true negative rate*, is the percentage of ordered sequences correctly predicted. We calculate the *overall accuracy* (ACC) as the average of SN and SP, which is more suitable than the percentage of all correctly predicted sequences for data sets with imbalanced class distributions. We also show receiver operating characteristic (ROC) curve and report area under the ROC curve (AUC) calculated using the R package ROCR (Sing et al., 2005).

Indicators to evaluate methods

$$\text{Sensitivity (or recall)} : S_n = \frac{TP}{TP + FN} = \frac{TP}{N_d} \tag{1}$$

is the number of correctly identified disordered proteins normalized to the total number of disordered proteins in the sample

$$\text{Specificity} : S_p = \frac{TN}{TN + FP} = \frac{TN}{N_o} \tag{2}$$

is the ratio between the number correctly identified ordered proteins and the total number of ordered proteins in the sample;

$$\text{Rate of false positives} : f_p = \frac{FP}{TN + FP} = 1 - S_p \tag{3}$$

is the ratio between the number of ordered proteins predicted as disordered and the total number of ordered proteins in the sample;

$$\text{Accuracy} : ACC = \frac{S_n + S_p}{2} \tag{4}$$

that is the average between sensitivity and specificity. It measures the overall performance of the predictor. Then,

$$\text{Precision (or selectivity)} : \text{Pr} = \frac{TP}{TP + FP} = \frac{TP}{n_d} \tag{5}$$

Jackknife

One of the earliest techniques to obtain reliable statistical estimators is the jackknife technique. It requires less computational power than more recent techniques.

Suppose we have a sample $x=(x_1,x_2,...,x_n)$ and an estimator $\hat{\theta}=s(x)$. The jackknife focuses on the samples that *leave out one observation at a time*:

$$x_{(i)}=(x_1,x_2,...,x_{i-1},x_{i+1},...,x_n)$$

for $i=1,2,...,n$, called *jackknife samples*. The *i*th jackknife sample consists of the data set with the *i*th observation removed. Let $\hat{\theta}_{(i)}=s(x_{(i)})$ be the *i*th jackknife replication of $\hat{\theta}$. The jackknife estimate of standard error defined by

$$\widehat{SE}_{jack}=[\frac{n-1}{n}\sum(\hat{\theta}_{(i)}-\hat{\theta}_{(.)})^2]^{1/2} \qquad (3)$$

where $\hat{\theta}_{(.)}=\sum_{i=1}^{n}\hat{\theta}_{(i)}/n$.

The jackknife only works well for linear statistics (e.g., mean). It fails to give accurate estimation for non-smooth (e.g., median) and nonlinear (e.g., correlation coefficient) cases. Thus improvements to this technique were developed.

https://www.datasciencecentral.com/profiles/blogs/resampling-methods-comparison

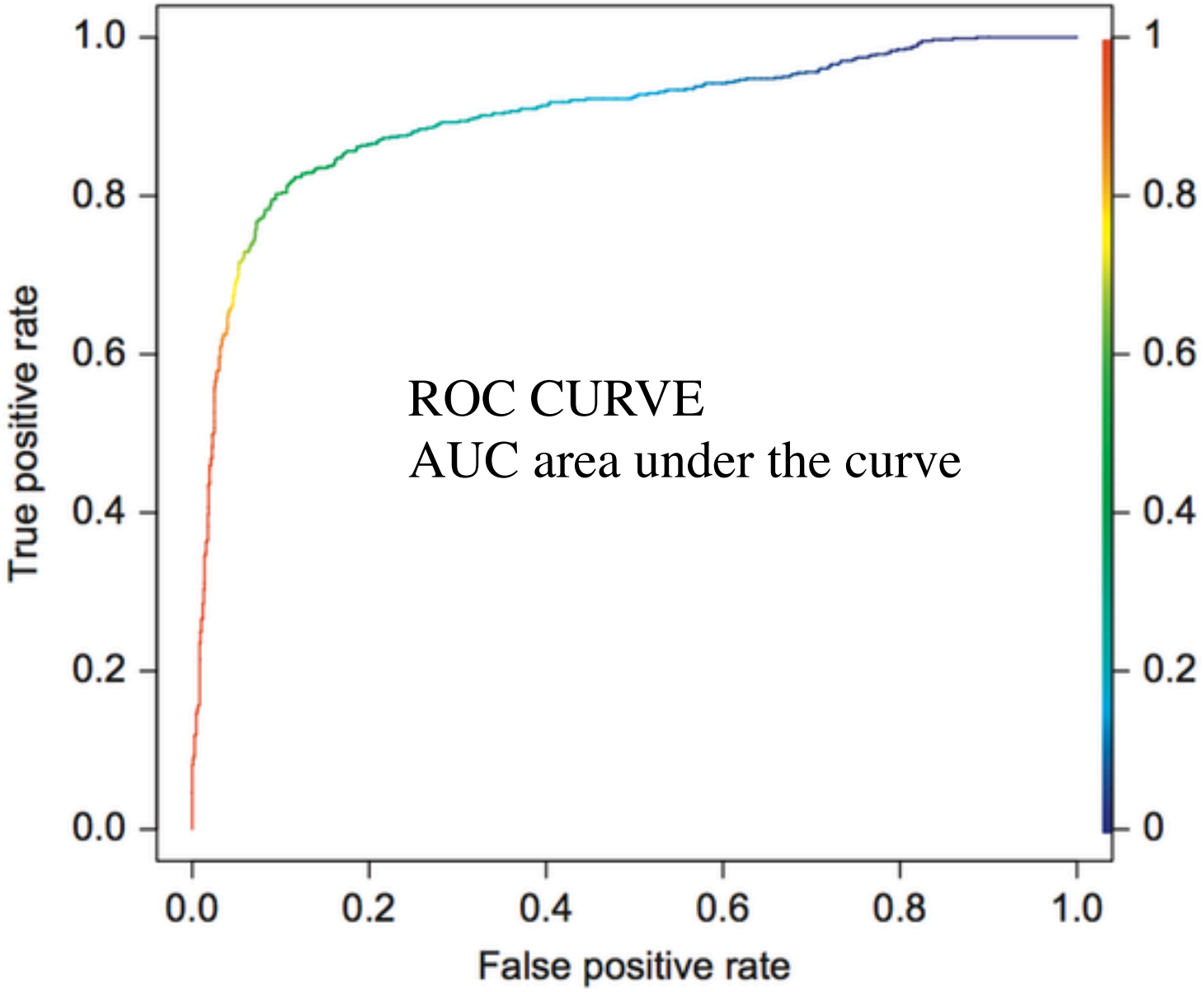

Fig. 1. ROC curve.

Table 5
Comparison of the prediction accuracies of our predictor and four other algorithms obtained for the data set of Prilusky et al. (2005)

	SN	SP	ACC
FoldIndex	76.9 30/39	88.1 133/151	82.5
DISOPRED	56.4 22/39	98.7 149/151	77.6
PONDR	71.8 28/39	92.7 140/151	82.3
GlobPlot	23.1 9/39	98.0 148/151	60.6
Our predictor	89.7 35/39	89.4 135/151	89.6

CONFUSION MATRIX

Table 2
Confusion matrix of the results of our predictor

	Predicted group		
	Disordered	Ordered	Sum
Disordered	**961**	116	1077
Ordered	132	**577**	709
Sum	1093	693	1786

Bold numbers along the major diagonal represent the numbers of correctly predicted sequences for each class, the numbers off this diagonal represent the errors.

BASIC TOOL TO EVALUATE Machine Learning CLASSIFICATIONS

Given a classifier and an instance, there are four possible outcomes. If the instance is positive and it is classified as positive, it is counted as a *true positive*; if it is classified as negative, it is counted as a *false negative*. If the instance is negative and it is classified as negative, it is counted as a *true negative*; if it is classified as positive, it is counted as a *false positive*. Given a classifier and a set of instances (the test set), a two-by-two *confusion matrix* (also called a contingency table) can be constructed representing the dispositions of the set of instances. This matrix forms the basis for many common metrics.

T. Fawcett / Pattern Recognition Letters 27 (2006) 861–874

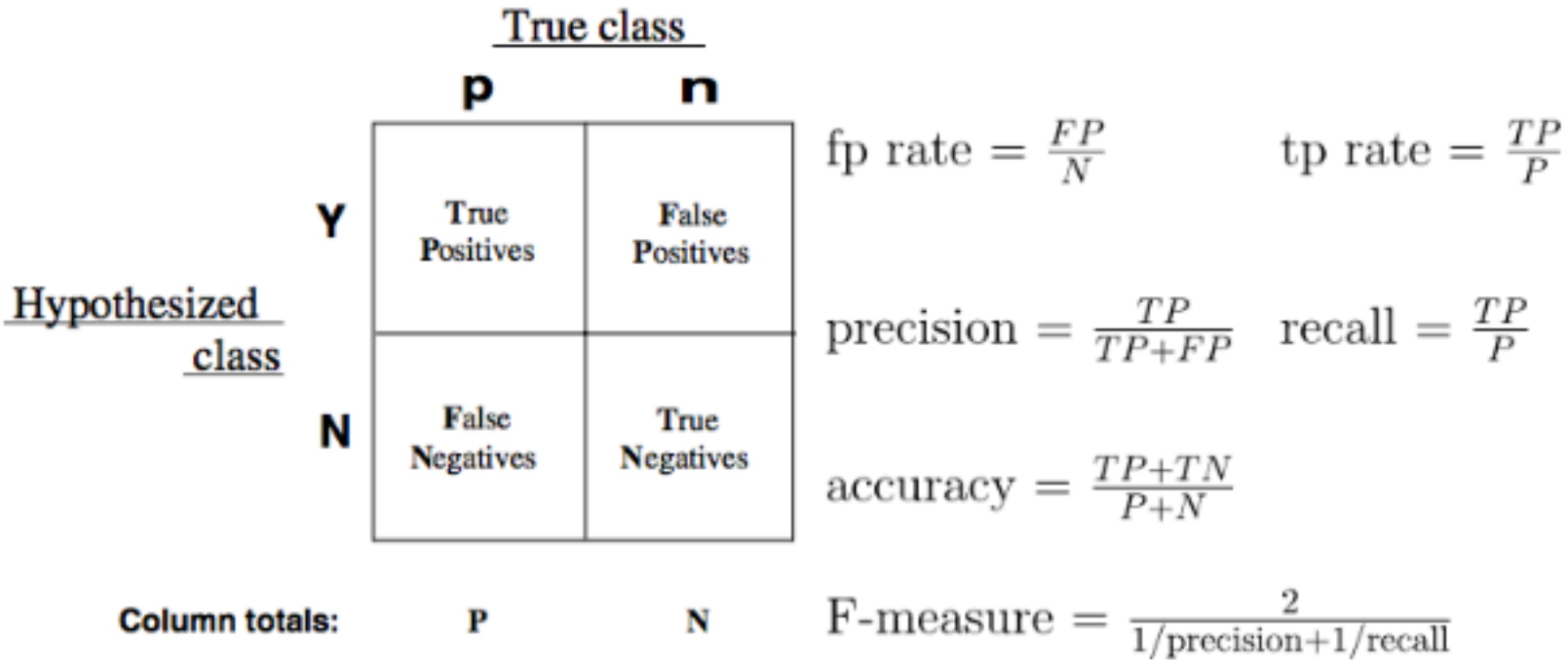

Fig. 1. Confusion matrix and common performance metrics calculated from it.

Homework (difficult)
study par. 3.9 in Rosner's textbook on ROC curves
study the difference between prevalence and incidence
(3.10) answer to the review questions 3EE
Read the paper by fawcett on ROC curves and make a 1 page resume titled ROC_yourname: on the theme **What is a ROC curve?** (this is a written test to be evaluated for the final exam)

Send it to andrea.giansanti@roma1.infn.it
subject: DA_2022 ROC