BAYES' THEOREM
(DA_2022)

Andrea Giansanti

Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

DA_2022 Lecture n. 8, Rome 28th March 2022

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

3 Basic definitions

Let us consider just finite sets of events, this is, conceptually, not a big limitation. All the events we shall consider can be , formally, as subsets of a reference container set Ω, which contains every possible outcome of an experiment; e.g.

$$\Omega = \{\text{head, tail}\}$$

in the case of the toss of a coin. $x \in \Omega$ means "x is an element of Ω", or "x is an event, a subset belonging to Ω". We shall associate to each event $x \in \Omega$ a probability $p(x)$, that is a positive measure normalized to 1. In the discrete case, whre Ω is made by N events the set of the $p(x)$ is a set of N non negative numbers $p(x) \geq 0$ (each one associated to one of the $x \in \Omega$) and such that tali che $\sum_{x\in\Omega} p(x) = 1$. In the simple case of a tossed coin we just have two possible events: $x = \text{head}$ e $x = \text{tail}$ and the probability distribution is $p(\text{head}) = 1/2$, $p(\text{tail}) = 1/2$ (for a fair coin).

Here are some elementary properties that can be derived using Venn diagrams of the type shown in figure 2.

$$p(A) \geq 0 \quad , \quad p(\emptyset) = 0 \quad , \quad p(\Omega) = 1$$

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

$$A \cap B = \emptyset \;\Rightarrow\; p(A \cup B) = p(A) + p(B)$$

Given a set of N events in Ω: $\{A_1, A_2, ..., A_N\}$ they are *mutually exclusive* if the occurrence of one of them precludes the occurence of the rest of the others. In particular, if the N mutually exclusive events are a partition of Ω then $P(A_i) = 1 - P(\cup A_j), withj \neq i)$. N events in $\Omega$are *independent* if the occurrence of each one of them does not interfere with the occurrence of the others; in this case $P(\cap_i A_i) = \prod_i P(A_i)$. Two events that are not independent are said to be correlated and to express the degree of this correlation one introduces conditional probabilities. Correlated events have, quite intuitively, a non empty intersection. Let us then denote with $p(A|B)$ the probability of the occurrence of A, provided that B occurred, that is the conditional probability of A given B. We can consistently express the intersection of two correlated events A and B as:

$$p(A \cap B) = p(B)p(A|B)$$

that is, the probability of the co-occurrence of the correlated events A and B is given by the probability of A times the conditional probability of A given B. One has also:

$$p(A \cap B) = p(A)p(B|A)$$

, it is worth noting also that:

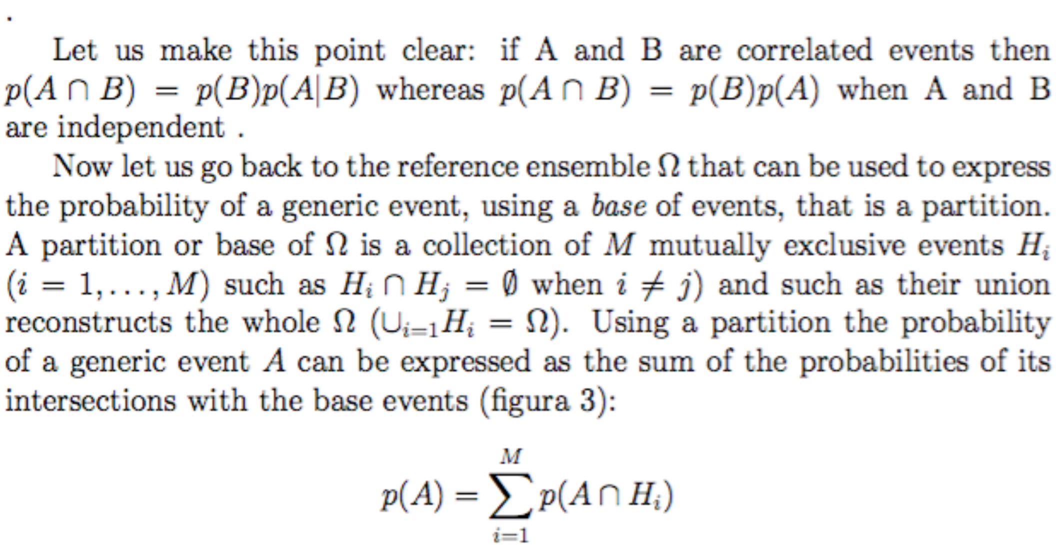
$$p(B|A) = \frac{p(A \cap B)}{p(A)} = \frac{p(B)p(A|B)}{p(A)}$$

and then, in general one has:

$$p(B|A) \neq p(A|B);$$

they are equal just in the case when $p(A) = p(B)$. If the occurrence of A is independent from the occurrence of B, then one has $p(A|B) = p(A)$ and the co-occurrenece of uncorrelated events A and B is just:

$$p(A \cap B) = p(A)p(B)$$



Let us make this point clear: if A and B are correlated events then $p(A \cap B) = p(B)p(A|B)$ whereas $p(A \cap B) = p(B)p(A)$ when A and B are independent .

Now let us go back to the reference ensemble Ω that can be used to express the probability of a generic event, using a *base* of events, that is a partition. A partition or base of Ω is a collection of M mutually exclusive events H_i $(i = 1, \ldots, M)$ such as $H_i \cap H_j = \emptyset$ when $i \neq j$) and such as their union reconstructs the whole Ω $(\cup_{i=1} H_i = \Omega)$. Using a partition the probability of a generic event A can be expressed as the sum of the probabilities of its intersections with the base events (figura 3):

$$p(A) = \sum_{i=1}^{M} p(A \cap H_i)$$

Warning: the degree of correlation of two events g(A,B)=P(A|B)/P(A) is a symmetric notion, whereas

causal relations require asymmetry

Correlation is required for Causation but is not sufficient for

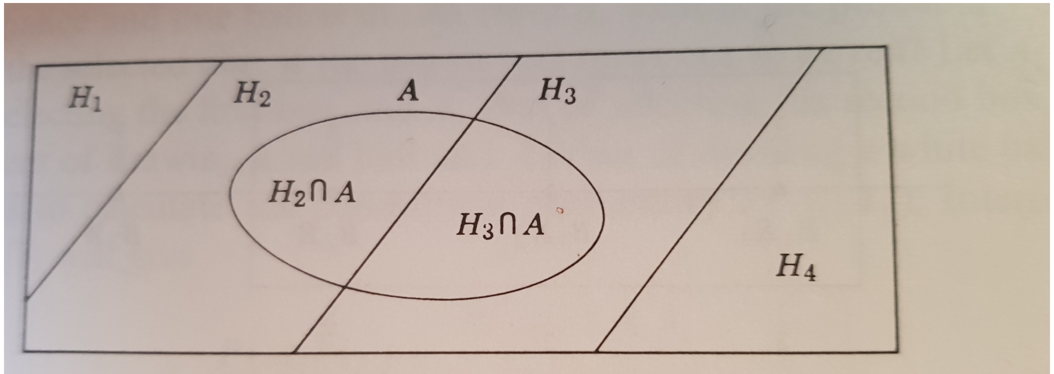


Fig. 11. A partitioning of a sample space.

5 Bayes' Theorem

Let us consider the methodological setting. Suppose you have a fact, an event to consider, E that you want to explain, to interpret, not making use of senses nor by concotting opoinions, but in a possibly transparent way, based on a quantitative analysis. Consider the "total" reference event of the calculus of probability Ω, we have introduced above. Then introduce a proper partition made by parts $\{H_i\}$ of Ω, to be used as a causative base to interpret E. In other words we want to determine the relative correlation of each one of the mutually exclusive H_i events in the partition with the event E. We shall express these correlation through conditional probabilities: of the form: $p(H_i|E)$. Let us start again from the general formula defining conditional probabilities, using events E and H_i: $p(H_i|E)p(E) = p(E \cap H_I) = p(E|H_i)p(H_i)$ and then, isolating $p(H_i|E)$. we get:

$$p(H_i|E) = \frac{p(H_i)p(E|H_i)}{p(E)}$$

,

which is equal to: $\frac{p(H_i)p(E|H_i)}{\sum_j p(E \cap H_j)}$, having used the projection of $p(E)$ over the base $\{H_i\}$, that is: $p(E) = \sum_j p(E \cap H_j)$.

$$\frac{p(H_i)p(E|H_i)}{\sum_j p(H_j)p(E|H_j)}$$

.

Introducing the normalization aka partition function: $Z = \sum_j p(H_j)p(E|H_j)$, we eventually get Bayes' formula in compact form:

$$p(H_i|E) = \frac{1}{Z}\, p(H_i)p(E|H_i)$$

.

Introducing the normalization aka partition function: $Z = \sum_j p(H_j)p(E|H_j)$, we eventually get Bayes' formula in compact form:

$$p(H_i|E) = \frac{1}{Z}\, p(H_i)p(E|H_i)$$

Eikosograms (RW Oldford)

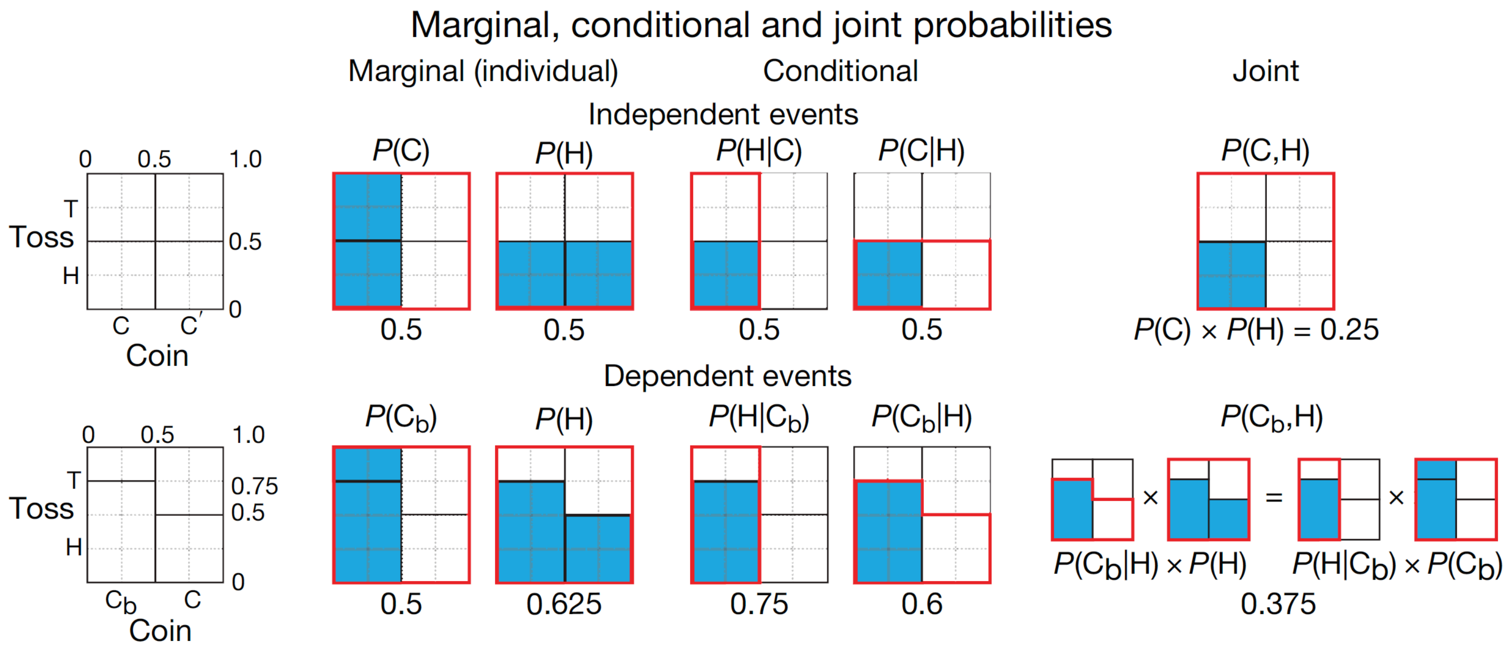


Figure 1 | Marginal, joint and conditional probabilities for independent and dependent events. Probabilities are shown by plots[3], where columns correspond to coins and stacked bars within a column to coin toss outcomes, and are given by the ratio of the blue area to the area of the red outline. The choice of one of two fair coins (C, C′) and outcome of a toss are independent events. For independent events, marginal and conditional probabilities are the same and joint probabilities are calculated using the product of probabilities. If one of the coins, C_b, is biased (yields heads (H) 75% of the time), the events are dependent, and joint probability is calculated using conditional probabilities.

From: N. Altman’s Bayes’ Theorem

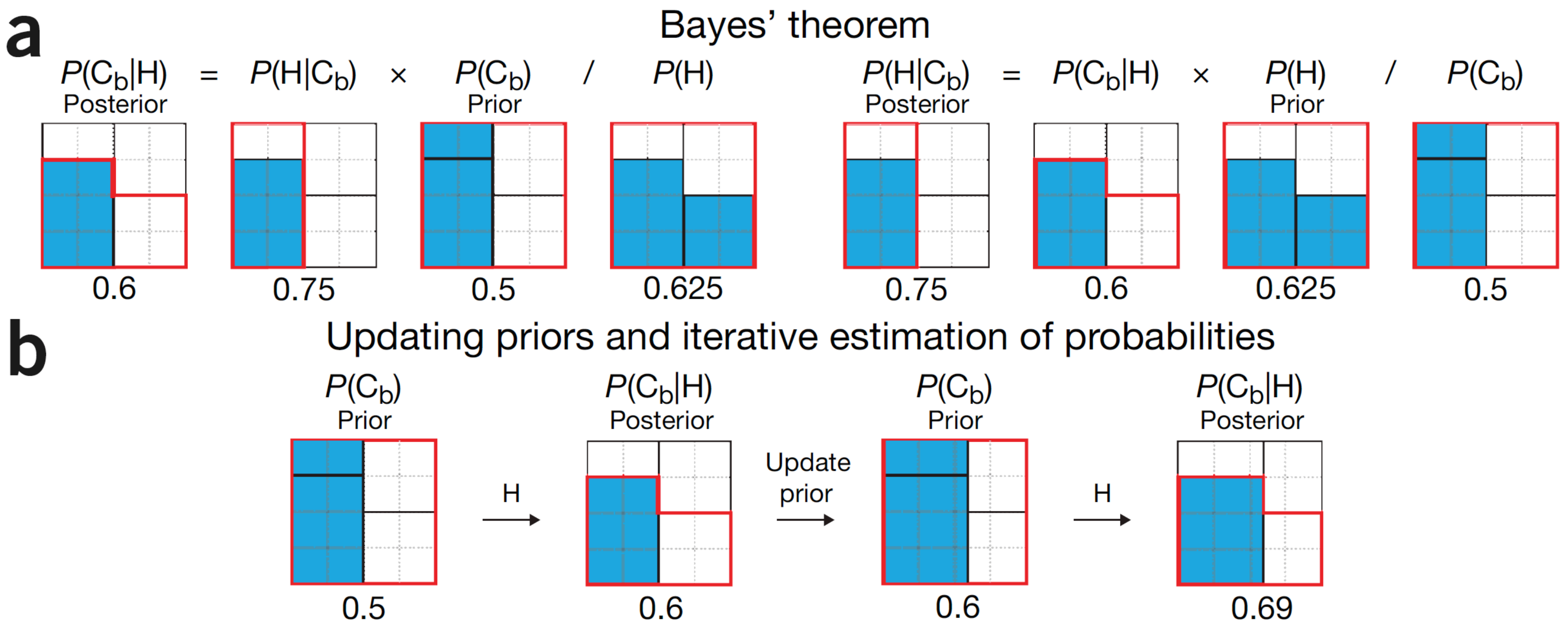


Figure 2 | Graphical interpretation of Bayes' theorem and its application to iterative estimation of probabilities. (**a**) Relationship between conditional probabilities given by Bayes' theorem relating the probability of a hypothesis that the coin is biased, $P(C_b)$, to its probability once the data have been observed, $P(C_b|H)$. (**b**) The probability of the identity of the chosen coin can be inferred from the toss outcome. Observing a head increases the chances that the coin is biased from $P(C_b) = 0.5$ to 0.6, and further to 0.69 if a second head is observed.

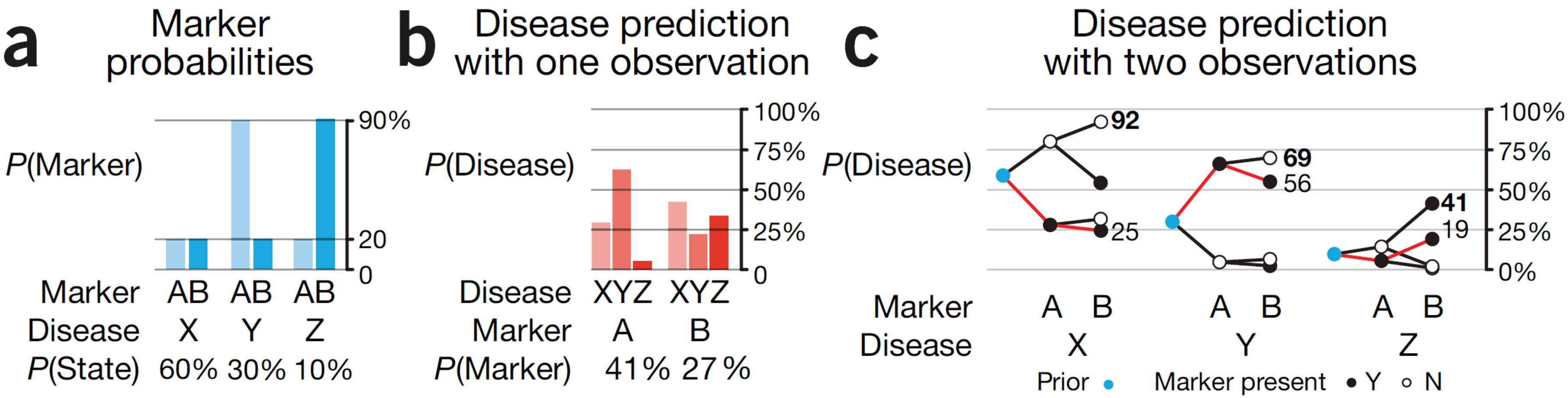
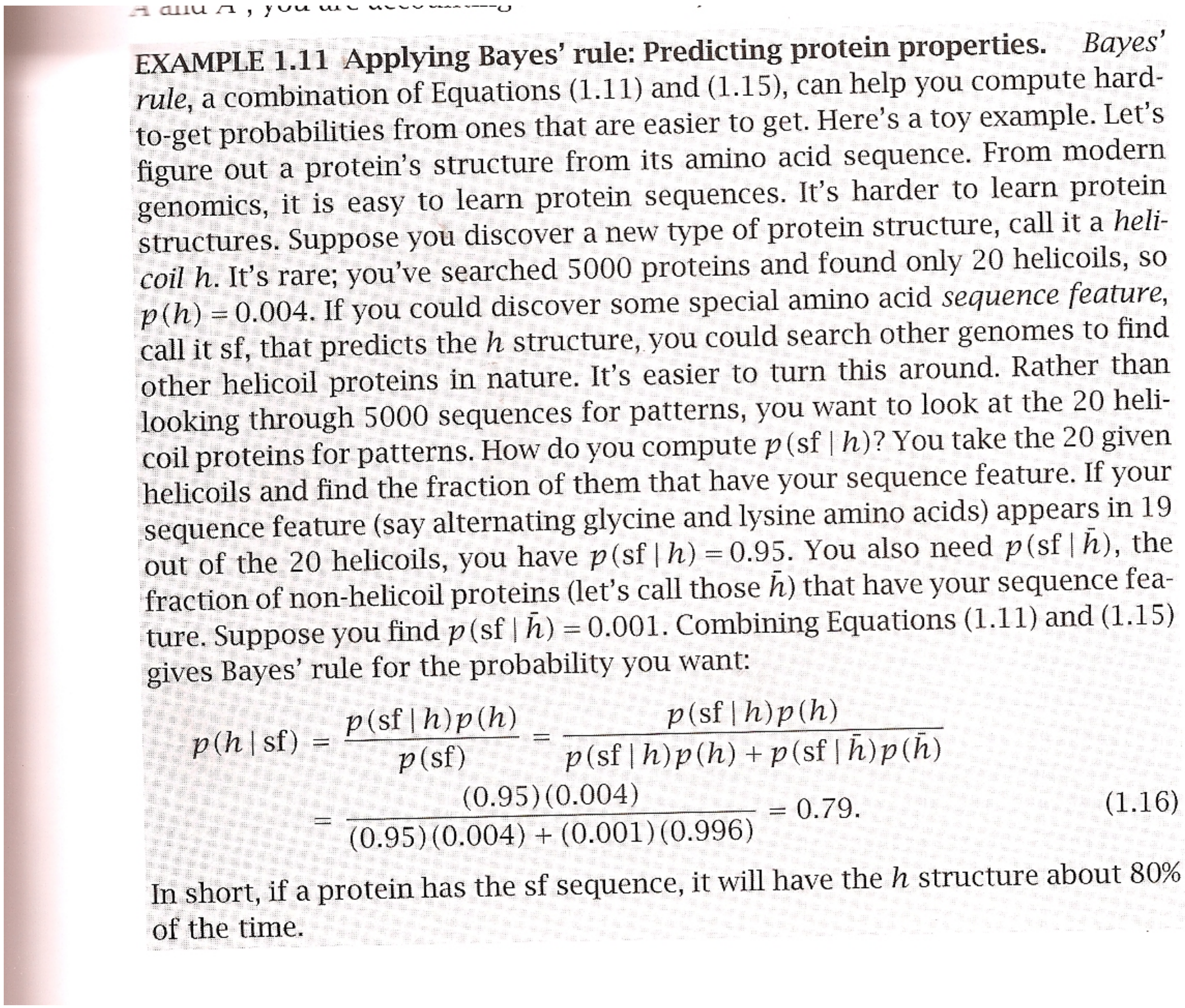


Figure 3 | Disease predictions based on presence of markers. (**a**) Independent conditional probabilities of observing each marker (A, B) given a disease (X, Y, Z) (e.g., *P*(A|Y) = 0.9). (**b**) Posterior probability of each disease given a single observation that confirms the presence of one of the markers (e.g., *P*(Y|A) = 0.66). (**c**) Evolution of disease probability predictions with multiple assays. For a given disease, each path traces (left to right) the value of the posterior that incorporates all the assay results up to that point, beginning at the prior probability for the disease (blue dot). The assay result is encoded by an empty (marker absent) or a solid (marker present) dot. The red path corresponds to presence of A and B. The highest possible posterior is shown in bold.

The relevance of Bayes' theorem: see DILL & BROMBERG: EXAMPLE1.11 …BIOINFORMATIC CONTEXT



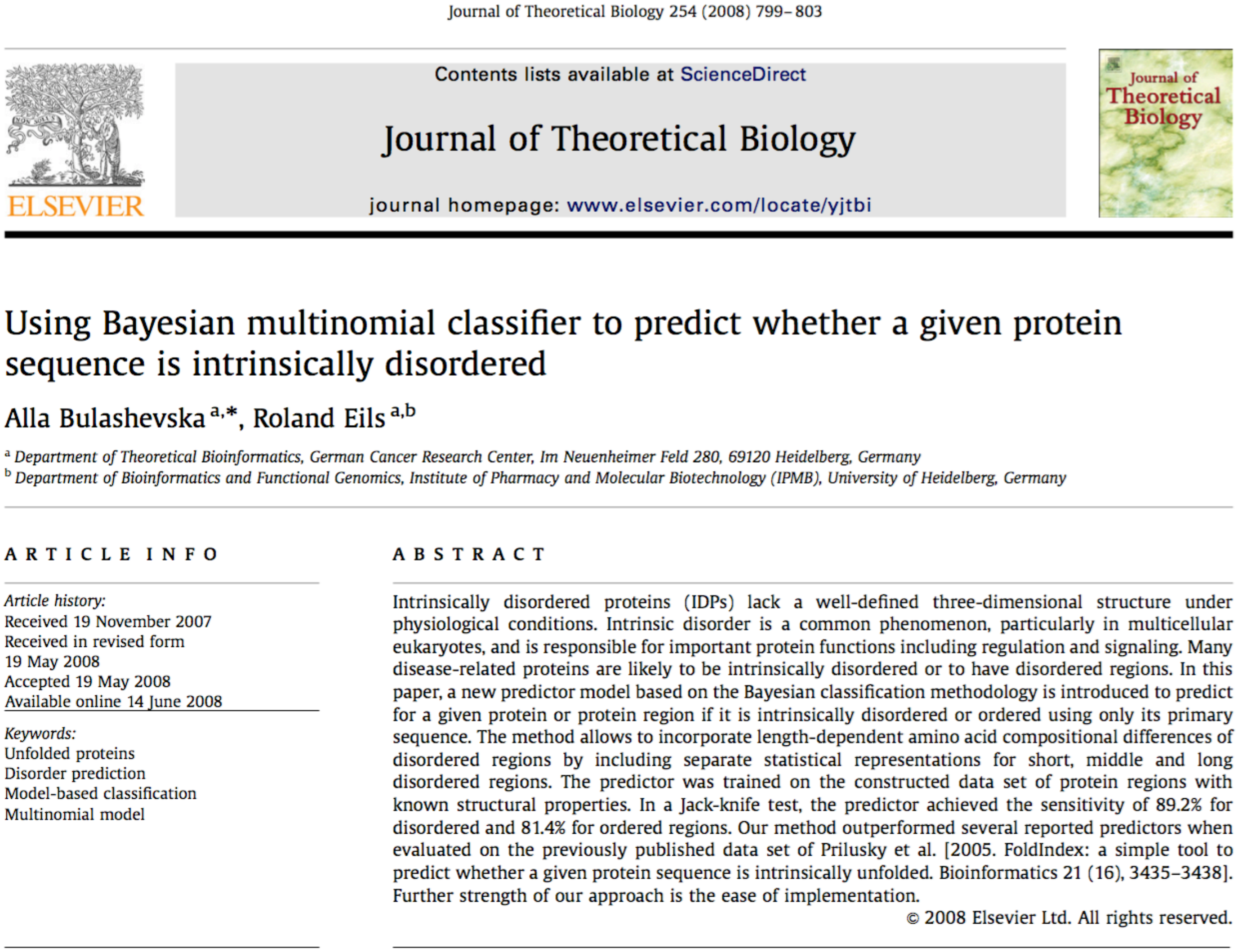
A and A', you are accounting

EXAMPLE 1.11 Applying Bayes' rule: Predicting protein properties. *Bayes' rule*, a combination of Equations (1.11) and (1.15), can help you compute hard-to-get probabilities from ones that are easier to get. Here's a toy example. Let's figure out a protein's structure from its amino acid sequence. From modern genomics, it is easy to learn protein sequences. It's harder to learn protein structures. Suppose you discover a new type of protein structure, call it a *helicoil h*. It's rare; you've searched 5000 proteins and found only 20 helicoils, so $p(h) = 0.004$. If you could discover some special amino acid *sequence feature*, call it sf, that predicts the h structure, you could search other genomes to find other helicoil proteins in nature. It's easier to turn this around. Rather than looking through 5000 sequences for patterns, you want to look at the 20 helicoil proteins for patterns. How do you compute $p(\mathrm{sf} \mid h)$? You take the 20 given helicoils and find the fraction of them that have your sequence feature. If your sequence feature (say alternating glycine and lysine amino acids) appears in 19 out of the 20 helicoils, you have $p(\mathrm{sf} \mid h) = 0.95$. You also need $p(\mathrm{sf} \mid \bar{h})$, the fraction of non-helicoil proteins (let's call those $\bar{h}$) that have your sequence feature. Suppose you find $p(\mathrm{sf} \mid \bar{h}) = 0.001$. Combining Equations (1.11) and (1.15) gives Bayes' rule for the probability you want:

$$p(h \mid \mathrm{sf}) = \frac{p(\mathrm{sf} \mid h)p(h)}{p(\mathrm{sf})} = \frac{p(\mathrm{sf} \mid h)p(h)}{p(\mathrm{sf} \mid h)p(h) + p(\mathrm{sf} \mid \bar{h})p(\bar{h})}$$
$$= \frac{(0.95)(0.004)}{(0.95)(0.004) + (0.001)(0.996)} = 0.79. \qquad (1.16)$$

In short, if a protein has the sf sequence, it will have the h structure about 80% of the time.

Realistic 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

ELSEVIER

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.

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,

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

Study Materials

- Slides
- Puga2015a
- D’Agostini9512295 3.1-3.5
- Eikosograms (RW Oldford)

https://cran.r-project.org/web/packages/eikosograms/vignettes/Introduction.html

Rosner 3.7