Lecture Title and Date

25m10e - Networks - Network Prediction 3/5/2025

Objectives of the Lecture

By the end of this lecture, students should:

1. Understand the basic principles of predicting connectivity between nodes in a network (interactions)
2. Be able to compare and contrast different methods to assess network interaction predictions
3. Understand the naive Bayes' Rule and its advantages over other prediction strategies
4. Be able to construct a basic Bayesian Network using Bayesian formalism

Key Concepts and Definitions

- Network assessment terminology:
    - Union - if any is TRUE, output is TRUE
    - Intersection - if any is FALSE, output is FALSE
    - Majority - output is the most common input, FALSE if tied
    - Weighted voting/supervised classification - scores are given weights and "voting" occurs in weighted fashion (outcome $R = \vec{w} \bullet \vec{f} + w_0$
- Bayes Rule: $P(Y|X) = \frac{P(X|Y)*P(Y)}{P(X)}$
    - Prior distribution $P(Y)$: probability distribution of parameter Y prior to updating (uniform distribution over range if absolutely no information known about Y)
    - Likelihood $P(X|Y)$: probability distribution of observing X if Y is true
    - Posterior distribution $P(Y|X)$: probability distribution of parameter Y given observations of X
- Naive Bayes: an assumption for a multiple parameter Bayesian model in which all parameters are written as independent
- Receiver Operating Characteristic (ROC) Curve: A plot of the true-positive rate or sensitivity against the false-positive rate or (1-specificity)
- Directed Acyclic Graph (DAG): a network that defines relationships of influence between nodes
- Bayesian Network: a DAG constituting a joint distribution which is the product of conditional probabilities for each node

Main Content/Topics

In order to construct and validate a network, we need to develop methodologies that integrate information we collect about the network. For example, if we want to understand how the subunits of a protein are interconnected, how do we integrate noisy experimental evidence to accurately characterize subunit interactions (Figure 1)? We might think of using an example protein, like that of the experimentally solved structure of RNA Polymerase II (RNAPII), to validate our proposed network and apply what we learn in this example case to other proteins. The information gleaned from RNAPII validation will enable us to identify a method which effectively incorporates noisy measurements into an accurate prediction of subunit interaction.

Let's say we use multiple experimental approaches to assess the connectivity of RNAPII subunits, giving us a list of plausible connections (Figure 2). The approach most likely to identify true interactions (without regard to false positives) would be one using a **union** approach: if any method says the subunits interact, then we'll say they interact. However, this is likely to overestimate the true connectivity of the protein. The method most likely to avoid false positives would be the opposite to the union approach, or an **intersection** in which any negative result would disqualify any positive result, but this has the opposite problem to the union approach by sacrificing true positives. One could instead treat each experiment as providing a "vote" for connectivity, creating a case where the predicted result is found by **majority** rule. This approach will help eliminate the effects of outlier experiments, but is vulnerable to highly correlated experiments.

None of the above approaches are particularly well suited to real-life situations in which we are often more confident about some experiments over others. A solution to this problem is found in applying weights to certain experiments, in an approach known as supervised classification (Figure 3). By adding weights to our experimental votes, we now include some kind of measure of confidence in our measurements that help us better identify the predictive power of each observation. While weights enable us to make more informed predictions, they also present another problem, namely how do we identify the weights we should use for each experiment? Enter **Bayes Rule** (Figure 4). Bayes Rule, while a complicated looking formula, is actually very simple when broken into its constituent parts.

The **prior distribution** $P(Y)$ is our starting point for the distribution of our weights. Do we expect each experiment to be equally informative (uniform distribution of weights) or is there reason to trust one kind of experiment over another? We multiply the prior by the **likelihood** $P(X|Y)$ divided by $P(X)$ to identify the distribution of outcomes X (in this case regions of RNAPII that are interacting) we would expect given our weight Y. Taken together, this information tells us how we should update our understanding of the distribution of weights given information about the outcomes we are likely to observe with those assumptions, creating a **posterior distribution** $P(Y|X)$. Bayes Rules therefore provides a rigorous mathematical approach to assess the informativeness of certain experiments, such as those done on RNAPII connectivity.

The final note on Bayes Rules relevant to our course is on that of parameter independence. The simplest assumption one can make about repeated experiments of a

phenomenon is that those experiments are independent, i.e. one experiment has nothing to do with any other experiment. This assumption applied to Bayes Rule gives rise to the **naive Bayes**, a form of Bayes Rule in which all parameters/weights are independent and can be assessed individually without consideration of other experiments (Figure 5). This assumption allows one to decompose a multiple-parameter Bayesian formula into an addition of multiple single-parameter Bayesian formulas, greatly simplifying one's ability to solve a problem by hand. Students should be aware that they may need to add in "dummy counts" for missing data if such a case arises (Slide 19). Students can work through the naive Bayes example of RNAPII connectivity by following along from slides 21-26. They should notice at the end that the **Receiver Operating Characteristic (ROC) Curve** (Figure 6), a plot of the sensitivity over the specificity for our experimental assessments, shows that the Bayesian approach outperforms all the other metrics regardless of our selected confidence threshold.

However, the naive Bayes assumption is quite rare in practice, especially in our RNAPII example in which we include repetitions of the same kind of experiment which are likely to be highly correlated (slides 28 & 29). While correlation in experiments is not discussed in detail in this lecture, students should be aware that in real life situations one may need to construct relationships between all experiments/parameters to conduct an accurate assessment of the task at hand.

Finally, networks which follow Bayesian principles are pleasantly known as **Bayesian Networks**, which neatly arrange the organization of parameters and outcomes using a **Directed Acyclic Graph (DAG)** formalism (Figure 7). The DAG formalism is a straightforward network in which nodes are connected by edges indicating the direction and weight of interaction. Bayes Rules are applied to mathematically describe the relationships between nodes under the rules of conditional probability. Bayesian networks can help to drastically simplify more complicated networks via filtering for the most informative elements in a network, as exemplified by gene co-expression networks (Figure 8). Filtering based on Bayesian probabilities can help remove uninteresting nodes and preserve only the most important edges, helping one to draw meaningful predictions from otherwise cumbersome or "busy" networks.

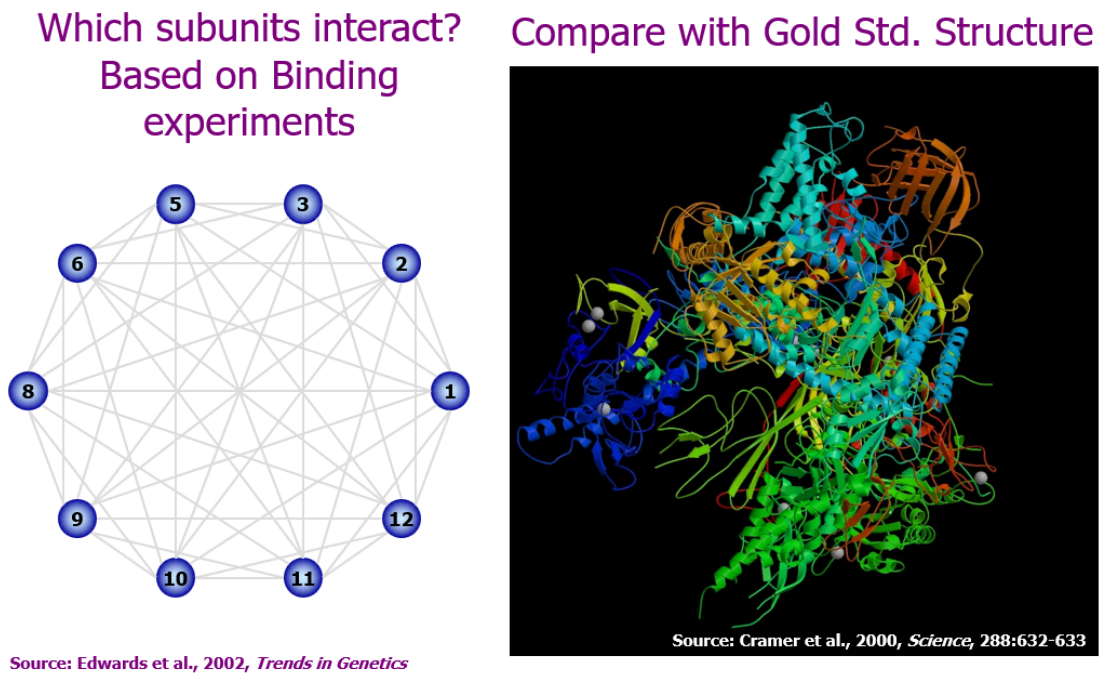

Figure 1: Slide 3

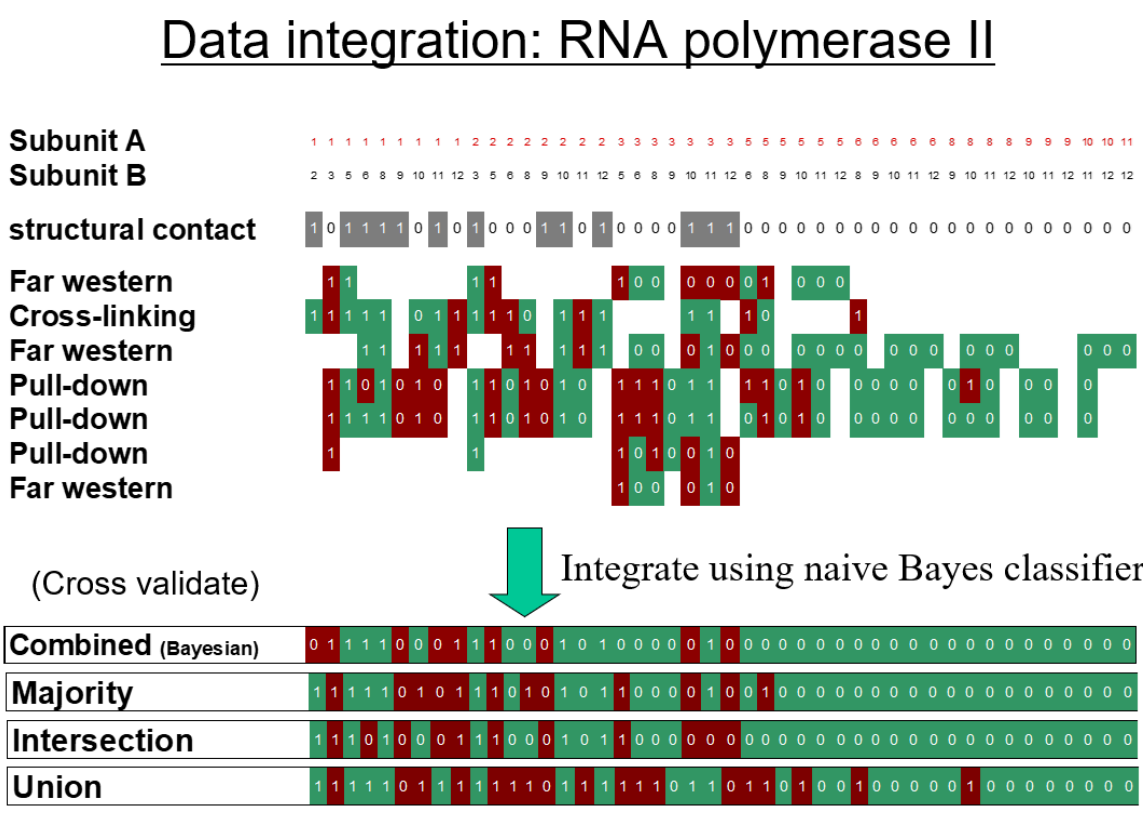

Figure 2: Slide 12

Simple Vote: $R = f_1 + f_2 + f_3 + ... + f_n$ With f = 1 or -1

If $\begin{cases} R>0; & I \text{ Interact} \\ R<0; & \sim I \text{ No interaction} \end{cases}$

Modify with feature weight:

$$R = w_1 f_1 + w_2 f_2 + w_3 f_3 + ... + w_n f_n = \vec{w} \cdot \vec{f}$$

If has prior knowledge w_0

$$R = \vec{w} \cdot \vec{f} + w_0$$

Figure 3: Slide 15

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$P(Y = y_i|X = x_j) = \frac{P(X = x_j|Y = y_i)P(Y = y_i)}{P(X = x_j)}$$

Figure 4: Slide 16

Update form of Bayes Rule

Updating our prior odds (w_0) by successively adding likelihood ratio terms (w_1,w_2,w_3) for each feature to get a final posterior odds

$$\log\left(\frac{P(I \mid f_1, f_2, f_3, ...)}{P(\sim I \mid f_1, f_2, f_3, ...)}\right) = \log\frac{TPR_1}{FPR_1} + \log\frac{TPR_2}{FPR_2} + \log\frac{TPR_3}{FPR_3} + .. + \log\frac{P}{N}$$

$$\uparrow w_1 \qquad \uparrow w_2 \qquad \uparrow w_3 \qquad \uparrow w_0$$

Figure 5: Slide 18

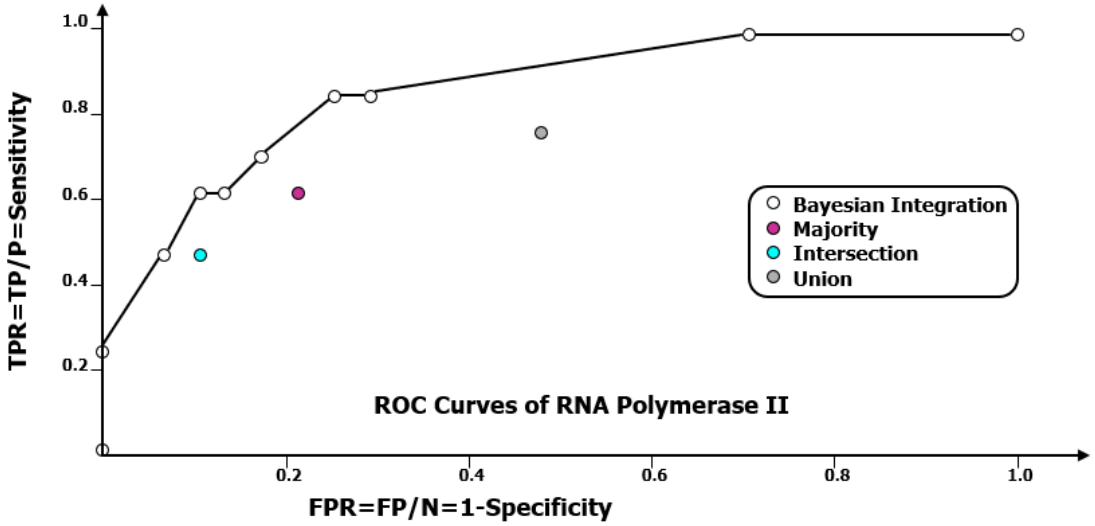

Figure 6: Slide 26

Naive Bayes

$$P(A,B,C) = P(C|A)P(B|A)P(A)$$

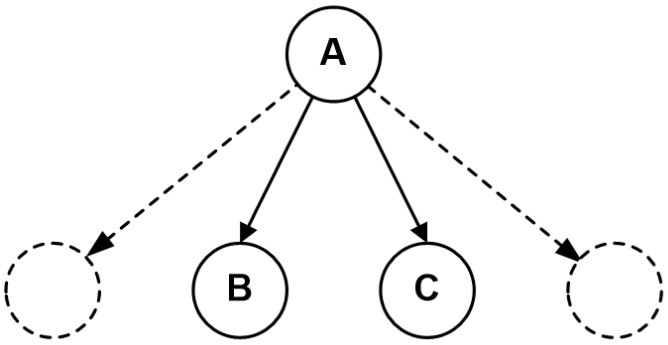

Figure 7: Slide 33

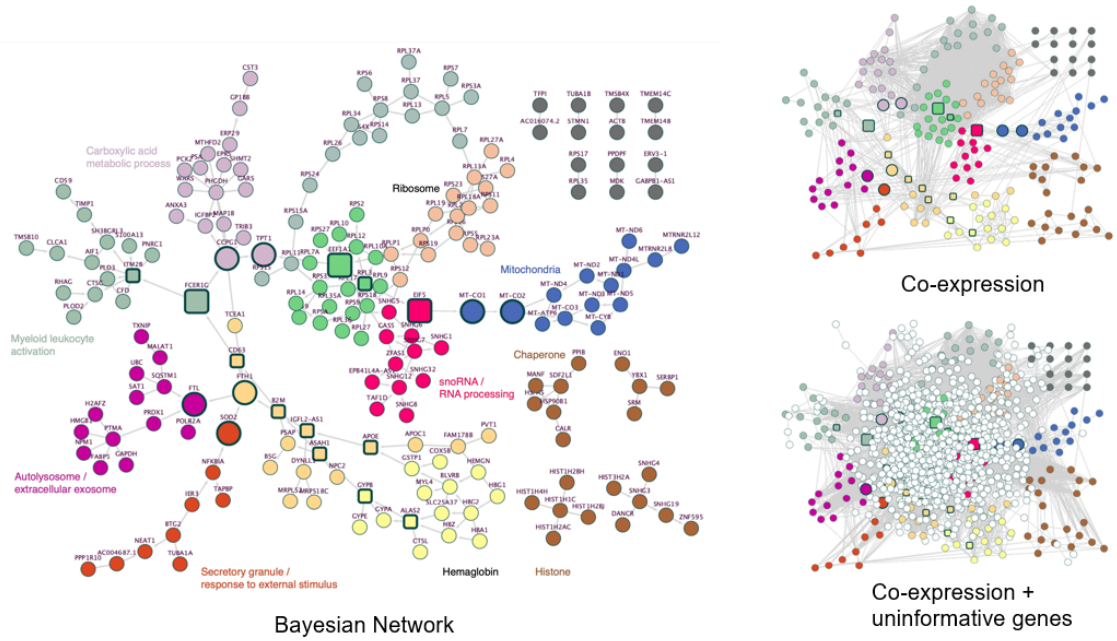

Figure 8: Slide 34

List all suggested reading here:

James, Gareth, Witten, Daniela, Hastie, Trevor, Tibshirani, Robert
An Introduction to Statistical Learning: with Applications in R
[ISLR (2nd edition)]
https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1071614177/ + https://www.statlearning.com
(Chapter 4.4.4 and 4.7.5 gives background on Naive Bayes.)

Edwards et al. (2002). Trends in Genetics, 18(10), 529–536.
Bridging structural biology and genomics: assessing protein interaction data with known complexes.
https://doi.org/10.1016/s0168-9525(02)02763-4
(Relates to the worked example.)

Notes: ISLR 4.4.4 is a good description of how naive Bayes classification compares to other approaches in classification and helps to hammer home the points made in the lecture. However, for a fuller understanding of classification, students should read the full chapter as this will prepare them better for applications in the real world. Students would also benefit from reading the Edwards et al. 2002 paper, as it covers RNAPII mapping directly relates to practical considerations of networks in bioinformatics.

References ISL/ESL (if any)

ISL: 4.4.4 and 4.7.5; ESL 6.6.3

Other Suggest references for many of the key concepts

Students should watch this 3blue1brown video on explaining Bayes Theorem to familiarize themselves with the topic and its applications:
https://www.youtube.com/watch?v=HZGCoVF3YvM