**COS 424: Interacting with Data**

Lecturer: Olga Troyanskaya
Scribe: Martin Suchara

Lecture 20#

# 1 Introduction to Molecular Biology

Cells are fundamental building blocks of living organisms. Cells contain a nucleus, mitochondria and chloroplasts, endoplasmatic reticulum, ribosomes, vacuoles, etc. The nucleus is important organelle because it houses chromosomes which include the DNA. The DNA is in essence a blueprint of the organism as it encodes information needed to synthesize proteins. Molecular biologists would like to understand how human biology works with the hope to treat diseases like cancer. One can look at simpler organisms such as yeasts to understand how human biology works. Admittedly, unicellular yeasts are very different from humans who have approximately $10^{14}$ cells. However, the DNA is similar across all living organisms. For example, humans share 99% of DNA with chimps. Naturally, we would like to know what information contained in that 1% of DNA is so critical to determine all the distinguishing features of humans, and we will try to answer this question.

## 1.1 DNA

DNA stands for deoxyribonucleic acid. DNA is an extremely long molecule that forms a double-helix. The double-helix backbone of the molecule consists of sugars and phosphates, and there is one base attached to each sugar. There are four types of bases: cytosine (C), guanine (G), adenine (A) and thymine (T). The DNA consists of two strands, and each base attached to one strand forms a bond with a corresponding base on the other strand. A only links with T and C links with G. A triplet of bases encodes an amino acid. Protein is a sequence of amino acids, and the functional subunit of DNA that encodes a protein is called a gene. DNA is depicted in Fig. 1.

## 1.2 Gene Expression - from DNA to protein

Gene expression is a two-step process in which DNA is converted into a protein it encodes. The first step is DNA transcription. In this step, the information from the archival copy of DNA is imprinted into short-lived mRNA. The structure of RNA is a little different, it contains ribose instead of deoxyrybose, and the four bases that bind to it are cytosine (C), guanine (G), adenine (A) and uracil (U). During transcription, DNA unfolds, and mRNA is created by pairing mRNA bases with the bases of RNA. In this process C in DNA translates to G, G to C, A to U, and T to A. After mRNA is translated, it is transported to the ribosome. The second step, protein translation occurs at the ribosome. During translation, the sequence of codons (triplets of bases) of mRNA is, with the help of tRNA, translated into a sequence of aminoacids.

Gene expression seems to be a straightforward process, but it is the control of gene expression that causes most phenotypic differences in organisms. Since many diseases result from complex changes on the molecular level, we need to observe and model these processes

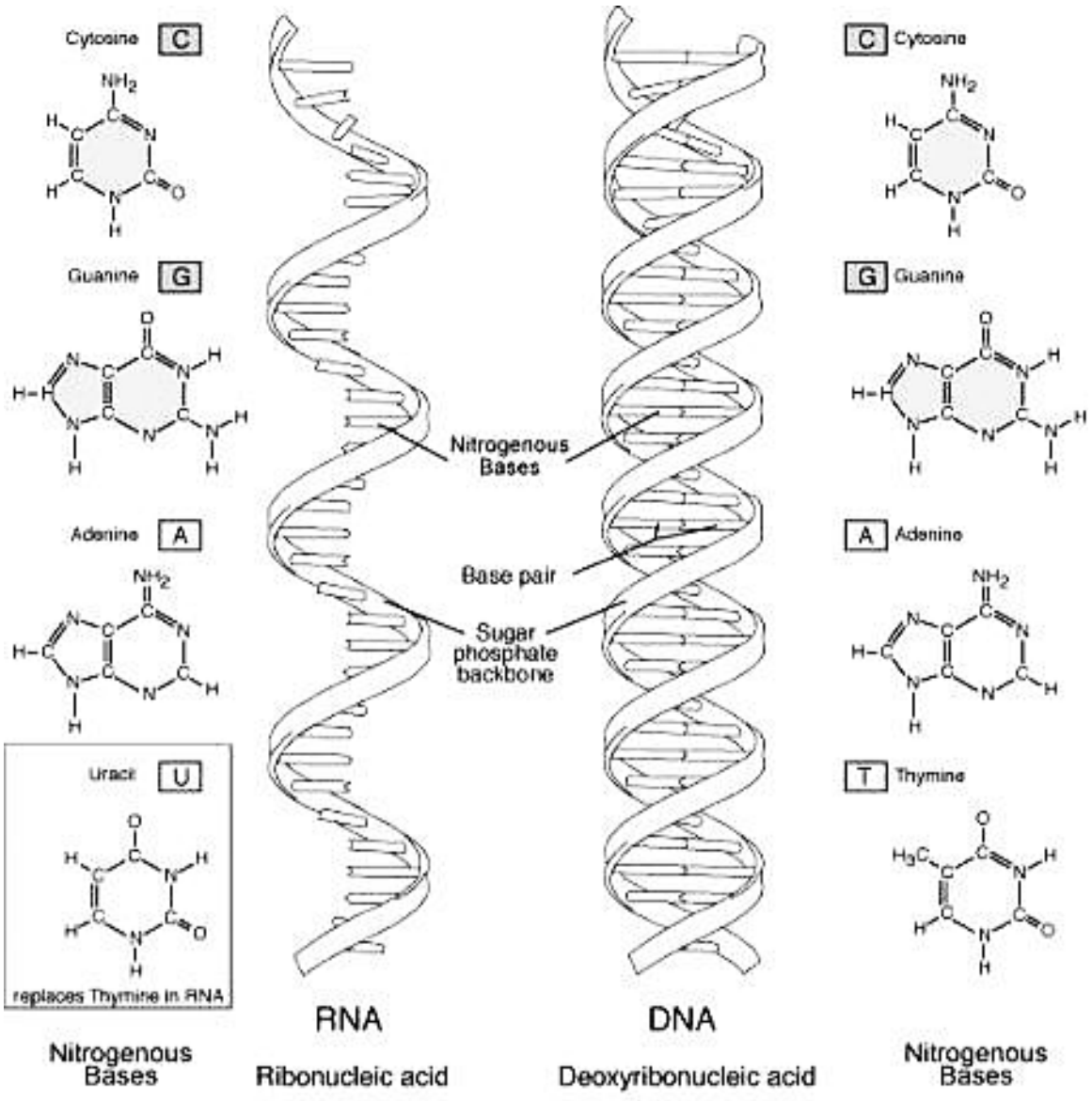


Figure 1: DNA is the blueprint for living organisms.

on the system level. Gene regulatory circuits are an example of machinery that allows us to depict gene expression graphically. In this model, arrow from A and B to D indicates that if genes A and B are made into protein, so is gene D.

## 2 Data Analysis in Genomics

Recent advances in Biology, such as sequencing of the genome, have lead to creation of enormous datasets. However, our knowledge of the inter-relationships that lead to function lags behind. Data analysis in genomics is really challenging for several reasons. First, the data are intrinsically noisy because they are results of measurements and observations.

Second, the data are heterogeneous and vary process by process. Finally, our coverage and accuracy of measurements varies process by process. To overcome these challenges and extract information about biological processes from the data, integrated analysis that uses probabilistic methods to deal with the noise are used.

## 2.1 Predicting Function of Unknown Proteins

The genome is just a sequence of letters. How do we determine what the corresponding proteins do? We already have a lot of information about some proteins, and we know about some of the interactions. This information has been captured in the Gene Ontology. The ontology is a structured vocabulary that describes proteins in terms of the associated biological processes, cellular components, molecular functions, etc.

Machine learning techniques are used to take advantage of the new advanced datasets. One possible method uses individual classifiers for each class. We note that the resulting predictions may be inconsistent. These predictions are subsequently combined and inconsistencies are resolved. This method is illustrated in the accompanying slides. 10 linear SVMs were used as the weak classifiers, and the median of the results was used. Finally, hierarchical consistency was enforced. Fig. 2. shows that enforcing hierarchical consistency improved the accuracy of the predictions for a majority of the nodes. The figure depicts the AUC of the original classifiers on the x-axis, and the AUC after the consistency enforcement on the y-axis. Since AUC is the area under the precision recall curve, the higher the value the better, and as the figure indicates, AUC increased. The improvement in AUC is not uniform. It was observed that the increase is largest for leaves of the network.

Validation is, of course, important. Validation is done as usual by holding some examples out. If a particular gene is predicted to be involved in DNA replication, an experiment can be performed. A copy of yeast that is missing the gene is created, and mitosis is observed. If both the mother and daughter cell get a copy of DNA, the gene could not have been involved in DNA replication and vice versa.

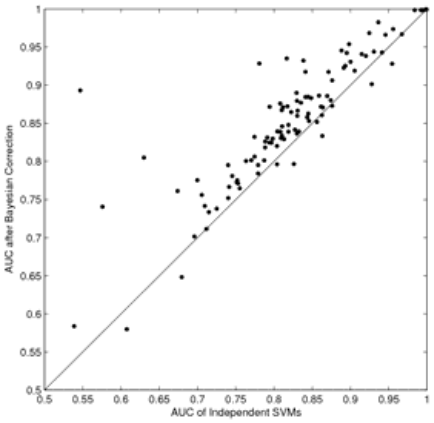


Figure 2: After consistency was enforced the AUC increased.

# 3 Predicting Biological Networks

Once we know the function of a protein, we want to learn more about how the protein interacts, i.e., we want to predict which circuits/interaction networks the protein participates in. BioPIXIE is a system that was developed to achieve this. The big ideas behind bioPIXIE are to combine information from all available sources, use information in biological context, and allow biologists to input predictions into the system. More detailed information about bioPIXIE is available at: http://pixie.princeton.edu.

## 3.1 Algorithm

BioPIXIE uses probabilistic graphical models to combine data. The user specifies some genes of interests, and the system indicates which other genes are most likely to be in the same functional neighborhood. The key part of the algorithm performs a local search in the Protein Protein Interaction network, starting at the center of the query. First, a characteristic profile of the query set is created. Then, the remaining set of proteins is searched for the closest matches. Since some of the datasets bioPIXIE uses are much more accurate than others, it was reasonable to use boosting and bagging. Nave base in bioPIXIE works well in practice, better than some more complicated techniques.

## 3.2 Example - Rad23 and DNA Repair

Rad23 is a protein that interacts with Rad4 in nucleotide repair. Recent research also suggested that Rad23 helps to repair DNA by inhibiting degradation of specific substrates. The new suggested role of Rad23 was tested in bioPIXIE by entering it along with other proteins: Pup1, Pre6, Rpn12. Since bioPIXIE shows a high probability link with Rpt6, this reconfirms involvement of Rad23 in DNA repair.

## 3.3 Testing and Robustness

The system was tested in the usual manner by pretending that particular known graphs are the queries and observing how successful the algorithm is. Robustness is another important issue. One needs to know if the underlying algorithms are robust to parameter choices, what happens when the queries are imperfect, etc.