

Information Theory, Coding and Cryptography
Dr. Ranjan Bose
Department of Electrical Engineering
Indian Institute of Technology, Delhi

Module - 08
Superinformation
Lecture – 08

Hello and welcome to this in lecture on Superinformation. So, we would start with the basic motivation followed by a brief background; then we will talk about what superinformation actually is some interesting results and applications and basically the conclusion or the summary.

(Refer Slide Time: 00:46)

Information Theory, Coding and Cryptography


Uncertainty and Information

- **Intuitive Feel** : Occurrence of a less probable event conveys more information
- **Mathematical Measure:**
 - Consider a discrete random variable X with n possible outcomes $x_i, i = 1, 2, \dots, n$.
 - The **self information** of the event $X = x_i$ is defined as

$$I(x_i) = \log\left(\frac{1}{P(x_i)}\right) = -\log P(x_i).$$

When the base of the logarithm is 2 the units of $I(x)$ are in bits

Physically: Entropy is a measure of randomness



So, let us quickly revisit the intuitive field that we have developed so, far about uncertainty and relating it to information.

So, we saw that the less probable event conveys more information if you know everything about an occurrence of an event we do not need to communicate it and hence it carries a limiting information. Now the mathematical measure we saw required a log and 1 over $P(x_i)$ for the self information of x_i and it is given by minus $\log P(x_i)$.

Now, entropy is a measure of randomness that is fine now in today's lecture we would see where we can really use this measure in a real world situation. Are we looking at

samples and streams of data where randomness is present? Do we analyze large chunks of information or streams of data where we need to process the information? One of the very interesting application areas is indeed genomic signal processing and analyzing information content in genomic data. The other place where we are really troubled with randomness and so, to psychiatric behavior is the stock market; the question is can we use information theory to our advantage here?

So, where I will be deriving our impetus from? Well it is a measure of randomness and we love randomness today and therefore, we will see where entropy works and it where it fails.

(Refer Slide Time: 02:27)


Information Theory, Coding and Cryptography

Entropy – The Average Self Information

- The **average self information** or **entropy** of a random variable X is defined as

$$H(X) = \sum_{i=1}^n P(x_i) I(x_i) = - \sum_{i=1}^n P(x_i) \log P(x_i).$$

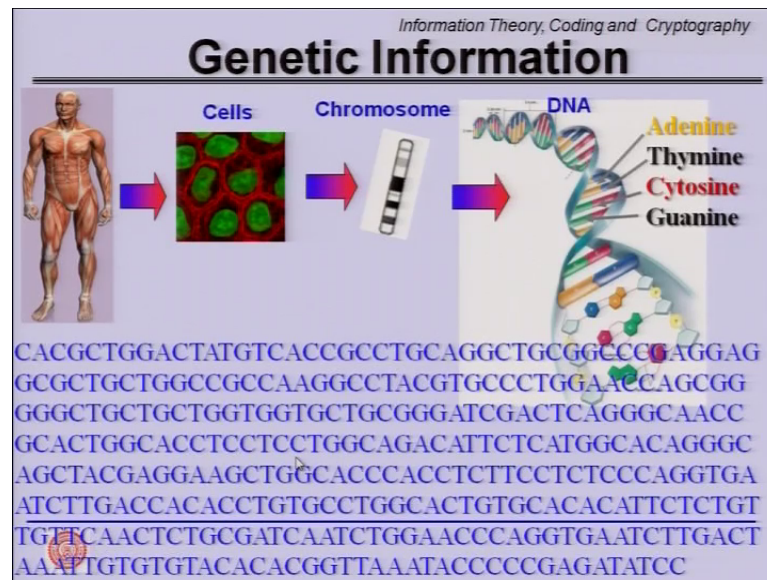
- $H(X)$ has been widely used to study the information content of **DNA sequences** earlier.
- However, as the name implies, it gives the information content **average** sense, and may not provide the true picture in many cases



Now, entropy is the average self information when we toss a coin or roll the die; we are not really interested in whether just the head comes up or just the die faces number 6, we would like to know, what is the average information content? And therefore, we average it by multiplying the self information with probability of occurrences.

Now this standard definition of Shannon's entropy has been used widely for DNA sequence analysis ok. All of us know what DNA is well we are dependent on it the human race is dependent on it. It is worthwhile exploring and is sequences, but as the name implies entropy is talking about the data in an average sense; can it hurt? Are we losing out something let us look at this average sense is what we would like to see whether it is working to our benefit or is really harming us.

(Refer Slide Time: 03:39)



So, very quick recap reacting information, human being, cells; cells have chromosomes. So, we have 23 pairs of chromosomes and inside that we have this helical double helical structure and these are this ARGC these are the symbols. So, we have only 4 symbols that make up this DNA chain and A combines with T and G comes combines with.

And A stands for adenine, T for thymine, C for cytosine and G for guanine. So, only 4 symbols are there; so, in some sense we are dealing with a quarter neri symbol stream we have only A, T, G and C; the first letters of these.

So, a genetic sequence might look like this meaningless line of data CAG C CTTG and we have enough of it huge amounts of data people have a hard time compressing and storing it; let alone analyzing it. Herein lies the genetic information question is randomness sure there is should we apply entropy analysis? What is the harm and this is exactly what we are trying to do; at least as a start we will look at randomness here and we can measure the randomness by applying the mathematical technique of entropy.

(Refer Slide Time: 05:19)

Information Theory, Coding and Cryptography

The problem with Entropy


- Consider the three sequences:

S_1 : AAAAGGGGTTTTCCCC

S_2 : AGTCAGTCAGTCAGT

S_3 : ATTGACCCTGTCGAGA

All the three sentence have EQUAL ENTROPY !!!
 $H(X)$ is **not** a good measure for analyzing sequences of symbols with repeating patterns, as is common in DNA sequences



Now, a minute of thought for the problem with entropy; so let me give you 3 toy sequences S_1 , S_2 and S_3 . The first one again has only 4 kinds of letters A, C, T and G, but the first sequence happens to be A A A then again 4 G's 4 T's and 4 C's. The other guy the other sequence number 2 has A GTC and then again AG TC and then again AGTC and so, and so forth and the third guy is hey I am looking for a pattern, but I cannot see anything.

Now, these 3 sequences could have been present somewhere here; hidden somewhere here and there millions and millions of these symbols you just cannot have enough of it. So, somewhere there could be these 3 toy sequences hidden there; now we say where is our entropy going to take us? So, we find out the entropy of the 3 sequences now what is entropy? Entropy is average self information.

Now, self information of what? Well we have 4 kinds of events; we can say it is a 4 sided die that I toss or if you can imagine a 4 sided coin on one side A is written the other side G another side T another side C and each time I toss this 4 sided coin or this die I get one of these symbols out. So, I have got a source it is a genetic source, but it generates my symbols this seemingly random S_1 , S_2 , S_3 .

Now, I want to find out the entropy of these 3 sources and then I can talk about my compression, I can talk about source coding theorem and all the beautiful tools that I have learned, but let me first find out the entropy. So, when I find out the average self

information right for A G T and C I found out that in sequence S 1; the equal number of is equal number of GS TS and CS. So, probability is one by 4 for each of them and same with S 2 and S 3. So, when I get on the back of the envelope calculation for my entropy it is nothing, but summation of $P_i \log \frac{1}{P_i}$ over P_i summation i is equal to 1 through 4, but we get the same answer for the 3 sequences.

Entropy says S 1 has the same randomness as S 2 which is the same randomness as S 3 entropy is a measure of randomness and I have given you the 3 measures. So, there is a flaw because to a very untrained i also these 3 sequences apparently have different levels of randomness. The entropy is fooled by the simple sequences we have trouble at hand; the sad part of the story is r DNA sequences have repeated patterns. If it is a repeated patterns that code for the your protein; so, protein coding. So, finally, these would later on eventually lead to the coding of the proton if there are proteins.

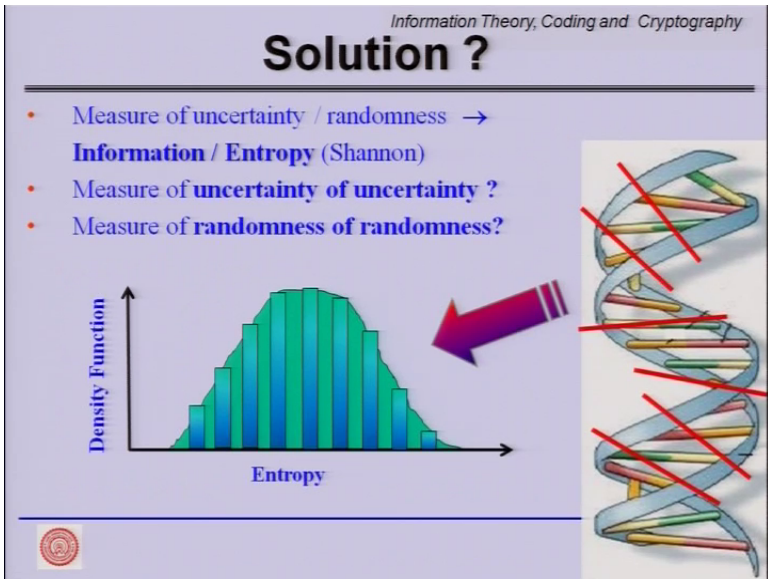
So, the sequence of this ARGC will dictate what kind of protein formation will happen and eventually all the other properties of the human being. So, repeated pattern is important and it routinely occurs in DNA sequences. We have started off with a practical example to look at the fully import of how entropy can or cannot be used. So, lot of literature has been using the standard vanilla flavored entropy to analyze sequences DNA sequences and; obviously, they have not been able to get too much out of it.

(Refer Slide Time: 09:54)

Information Theory, Coding and Cryptography

Solution ?

- Measure of uncertainty / randomness → **Information / Entropy** (Shannon)
- Measure of **uncertainty of uncertainty** ?
- Measure of **randomness of randomness** ?



Now, what is this that we are trying to get out of these sequences? Well let us look at this in a greater detail. So, we have on this right hand side this double helical structure and the runs in the ladder are the connections and on the 2 side of the rugs are either A and T or G and C that is how nature has designed us, but the issue is suppose I divide this double helical structure into parts. So, I have a first part and then another section and another section. So, I break it up into sections and each section is subjected to the calculation of entropy; how do I do it? Well it is possible that these 3 were the first 3 sections of the DNA chain S_1 S_2 and S_3 ok, but they could be different.

So, if you look at the different sections they may or may not have the same entropy. So, when we do this calculation, but you will generate one particular number and then what you do is you make a simple histogram. So, on the x axis we have a entropy and we find the variations in entropy; it is not a surprising fact because we have this long chain and for different sections the entropy varies and clearly you have a distribution.

Now, for the sake of simplicity in illustration only I have put together this nice bell shaped distribution; in nature it is slightly different, but let us understand where we are going. So, the first interesting observation and first failure where we see that entropy would have failed unless we break it up into sections is that there is a distribution.

I could have taken the whole sequence long sequence 100, 1000 ARG C symbols they are called the nucleotides and you find entropy and you get one number you get 2.73 bits one big number for this gene. What does it tell me? It is just a number, but if you look at different sections there is a distribution.

So, if entropy is a measure of uncertainty or if entropy is a measure of randomness then there is a randomness of randomness ok; this is a very interesting observation. So, far we have only been talking about randomness what if the randomness is not constant? That is the level of randomness in this section is not the same as this, is not the same as this; it itself varies. What if we can have a measure for randomness of randomness?

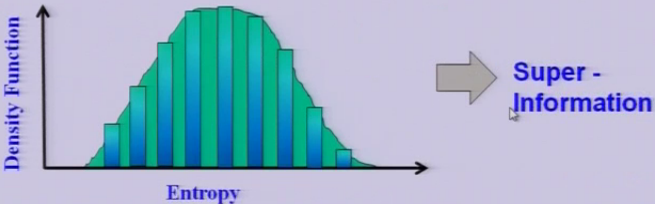
It is like rate of change of distance speed rate of change of speed acceleration rate of change of acceleration jerk like. So, you can keep going; so, if you could find self information then you found average self information and now which we defined as entropy; I would like to find out entropy of entropy that is the basic motivation.

(Refer Slide Time: 13:47)


Information Theory, Coding and Cryptography

Solution ?

- Measure of uncertainty / randomness → **Information / Entropy** (Shannon)
- Measure of **uncertainty of uncertainty** ?
- Measure of **randomness of randomness** ?



Super - Information



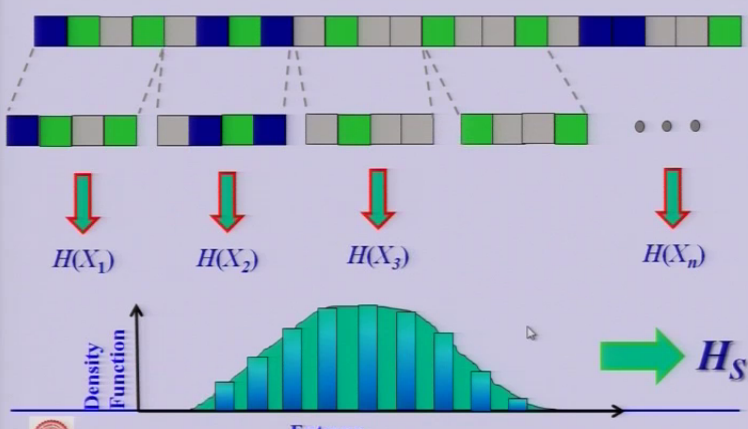
And this randomness of randomness is what we define as super information.

Most entropy is classically information as a measure of information what do we call this well we call this super information. And what is it telling us? It is the uncertainty of uncertainty or randomness of randomness, but that is just a concept we better find out a mathematical measure for this.

(Refer Slide Time: 14:18)


Information Theory, Coding and Cryptography

Finding Super Information



$H(X_1)$ $H(X_2)$ $H(X_3)$ $H(X_n)$

Entropy H_s



So, how do we find this super information first we have a long chain of symbols and since very taking an example of biological genomic sequences, we are talking about ATG

CCAT some random things and these colors possibly indicate the different kinds of symbols.

First, we divide them into blocks now how big are these blocks? That is the thousand rupee question, but right now we would look at some arbitrary block size and I find out $H(X_1)$ I get one number here and then I get $H(X_2)$ $H(X_3)$ and so, and so forth up to $H(X_n)$. And then I find out the histogram and if I normalize it to make the area under the curve 1, assume it is a kind of a continuous distribution then I can have an a kind of density function and this is called H S; H subscript S stands for super information.

So, I do get a one final one number because I am going to find out if this is a probability continuous random variable pdf; if I have a pdf then I can always find out the entropy we have already studied that.

(Refer Slide Time: 15:56)

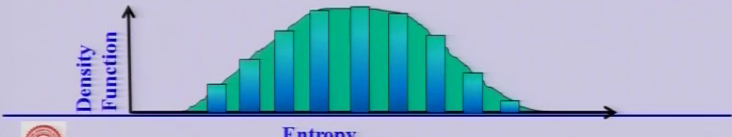
Information Theory, Coding and Cryptography

Super Information

- **Divide** the entire sequence of symbols into N blocks, of length B each.
- Construct the **histogram** of $H(X_i)$
- **Normalize** the histogram to form the probability measure

$$p_j(X_i, M) = \frac{H_j(X_i, M)}{\sum_{k=1}^M H_k(X_i, M)}, \quad j = 1, 2, \dots, M.$$

- **Super-information** is defined as

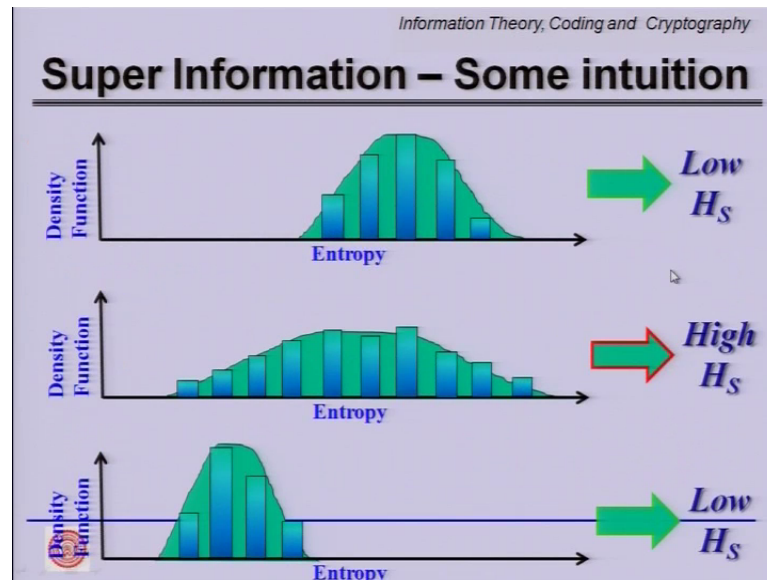
$$H_s(X_i, B, M) = -\sum_{j=1}^M p_j(X_i, M) \log p_j(X_i, M)$$


So, what is the step what are the steps we are going to take divide the entire sequence of symbols into n blocks of length B right.

Construct the histogram of this $H(X_i)$ where i is the i th block normalize the histogram to form the probability measure. So, if you see I have in the numerator; so, $H_j(X_i, M)$ right and then you can have this M could be the number of blocks that you have here and you can divide it so, that you have the sum of probability is equal to 1.

And finally, these probabilities can be used to find another. So, entropy of entropy once you have a probability you can always plug into the standard formula for finding out the entropy and you get the entropy of entropy right. So, it is a pretty straightforward way, but what will it tell us what is this randomness of randomness.

(Refer Slide Time: 17:12)



So, let us get some intuition into it; again look at the first curve on the x axis I have got this entropy, on the y axis we have this density function. And suppose for whatever long sequence that I have I have broken up into blocks of size B. And these guys these subsections have a distribution, but it is not widely distributed, it is not spread out. So, the H_S the super information value will be low; it tells me that the entire sequence has high entropy, but there is no variation of this entropy.

So, a big sequence all sections having very high values of entropy all sections are highly random; the whole sequence is highly random, but the super information is low. Again if you look at the third curve I have got this long sequence, I divided it into sections and I find out a distribution of the entropies within that and again I find all of them have low entropy. Then again my super information is low, but if you have this long sequence and some of them are very high entropy values; some of them are very low some of them lie in the middle, but there is a wide range of entropy only then it is high super information.

So, just if I give you a big sequence which is very very random all the way. So, this random, but the randomness of randomness is low; similarly low randomness and hardly

much of our distribution spread, low super information only when there is a wide distribution range does the super information value go up. So, in nature what does it tell us? Does nature believe in just randomness, does nature believe in randomness of randomness? It is worthwhile exploring.

(Refer Slide Time: 19:51)

Information Theory, Coding and Cryptography


Back to the toy sequences

S_1 : AAAAGGGGTTTTCCCC

S_2 : AGTCAGTCAGTCAGT

S_3 : ATTGACCCTGTCGAGA

All the three sentence have EQUAL ENTROPY but not the same Super Information



Symbols/Block (B)	Number of blocks (N)	$H_s(S_1)$	$H_s(S_2)$	$H_s(S_3)$
2	8	0	0	0.543
3	5	0.971	0	0.971
4	4	0	0	1.500
5	3	0.918	0	1.585

So, but before we proceed let us look at those toy sequences once again $S_1 S_2 S_3$ we had standard equal levels of entropy, but this time we also find out the super information, but super information requires us to break it up into blocks.

So, suppose we break up into block sizes of 2 and there were 16 symbols in all 4 4 za 16. So, 8 number of blocks is 8 and if you find out we find that look at the sequence 1 A A block 1 A A block 2 and so, and so forth. If you look at it you will find that the super information is actually 0 for sequence 1 and 0 for sequence 2; if the block size B is 2. But the third guy gives a nonzero value; so, certainly at the resolution. So, B can be taken as a block resolution and a resolution of 2; I can start seeing some variations in history and therefore, I get a nonzero value here.

If you change the block size block length to 3 you start getting some other values and so, and so forth. So, what this table tells us is that if you have different sequences which were apparently having equal entropy values; the super information values need not be the same. But it depends on at what resolution you are looking at. So, this is quite important and it will become our design parameter.

Resolution means what is my block size? So, if you go back if you see this is a long sequence I break it up into 4 4 4 4. So, that B is 4 I will not get the same answer if my B was 8 because its distribution will change because this random is in the sequence.

So, it is in a way at what fine level are we actually exploring this long sequence? Fine is there a notion of the amount of resolution we can get in versus the information; just like you have a time bandwidth product constraint do you have a resolution versus information that it conveys; is there a product like that? Those other questions we will ask ourselves ok, you cannot have very fine resolution and also very good estimates of the information.

Because finally, for every block you are finding out the entropy what is entropy? Entropy is average of these 4 symbols, but if there too few symbols the average does not have a very good meaning; the mean so weak law of large numbers.

The mean value gets closer to the real mean if the number of sample points are large, but the block sizes become large, you become coarse in your analyzing the sequence the analysis of the sequence become coarse. So, there is a tradeoff apparently there should be some optimal block size for the kind of data we are dealing with ok; that is what we mean by resolution. I cannot have infinite resolution both in time and space or time and frequency. Similarly, we cannot have infinite resolution both in the block size and the amount of information it gives us.

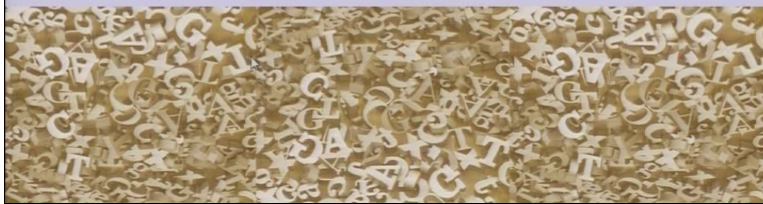
(Refer Slide Time: 23:59)

Information Theory, Coding and Cryptography

Current Status?

DNA is a code, but little is known about :

- Its exact **information** content
- The nature of **redundancy**
- **Statistical** properties
- How to **efficiently segment** Introns and Exons
- **Error control codes** within



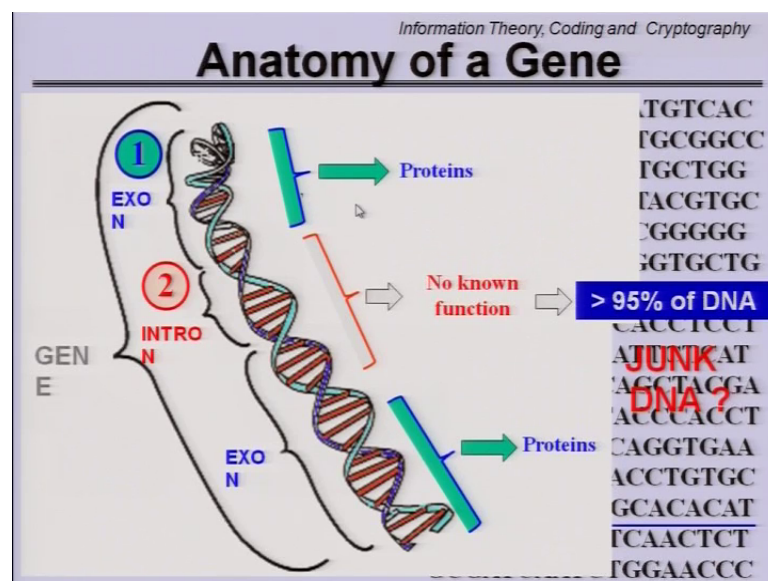
So, we come back to this DNA and this is really interesting because there are lots of symbols out there and just to see what is information content, what nature is embedded in us that is being conveyed ; well we have this DNA as a code. And we would be interested to finding out what is the information content?

What is the nature of redundancy there, what are the statistical properties are there correlations right. And we would like to look at 2 interesting aspects of this DNA sequence one is called the introns; the non coding region and exons which are the coding regions. We will subsequently tell you what these are ah in a matter of couple of slides and the other question that we asked ourselves is do these sequences have an inbuilt error control pattern.

Because when we go from one generation to other it is like a communication system genetic information is embedded the question is how does nature do error correction? As if there are errors in this code there will be all kinds of funny people in various generations coming up right four noses, one eye right we do not know, but nature has a inherent way to correct for errors.

But first let us find out what is the redundancy? And the bigger question is once we have an understanding of water introns and exons; what is the method to separate them or segment them based on information theory; so, very very practical application.

(Refer Slide Time: 25:55)



So, let us quickly do a one on the anatomy of a gene. So, this is our long double helix structure; now this is divided into segments by nature itself and these are of 2 types which are inter linked. So, exons and introns; so, if you see an exon is followed by an intron is followed by an exon is followed by an intron and so and so, forth.

But what is more exciting is that this exons code for proteins; only the genetic information present in exons the ATG CCAT whatever only they tell which proteins to generate and then build up the body. The introns; however, barring a few marginal applications have no known function and that is the current state of the art yes some theories exist that they do this.

So, therefore, they are called non coding region and before this small functionalities were discovered they were called as junk DNA. Exons at the other hand code for protein and nature has this alternate sequence exon intron exon intron and so, and so, forth. What is more surprising is 95 percent of the DNA is intron; 95 percent of our DNA has no known function. I mean I would hate to believe it because we are spending a lot of energy carrying those information and transferring it to next generation. So, much of energy is being used and nature is known to be highly efficient.

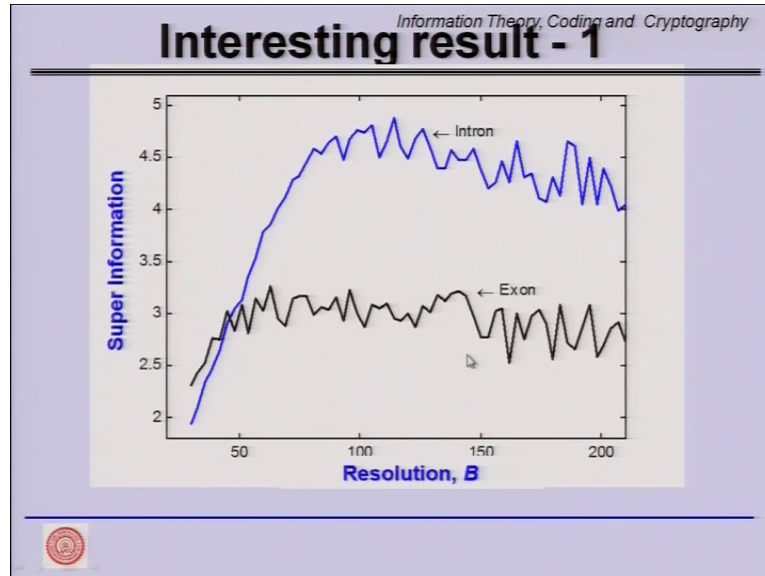
So, why is there 95 percent of this ATGC; this long chain of streams having no known really no known functions well there are some basic functions. But the most important functions is done by only the 5 percent of the sequence which is the exons. And therefore, it is important for us to quickly have a very good way to differentiate between exons and introns.

So, the question we can ask ourselves is there a mathematical way from information theory which can tell me after their information carrying; if you want to stick with the name of junk then there is hardly any information contained in the introns. But they are not conveying anything to the next generation whereas; it is the exons which according to biology is conveying the information.

So, hopefully let us pray that there is the information theoretic difference between in transit axles; because together they look exactly like this ARG C AG GT TAC; it keeps going it does not stop you do not even know when one stops one when it starts it is just a sequence; it just keeps going. So, question is can we separate the junk data from the

actual exons which code for protein. So, the question we are going to ask ourselves is can we analyze the human DNA using super information?

(Refer Slide Time: 29:39)

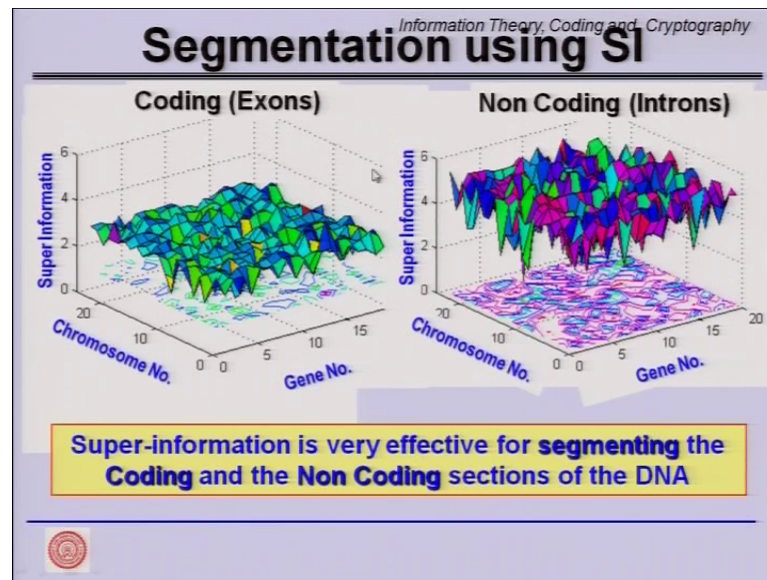


So, these are some of the results which we are going to discuss here on the x axis we have this resolution parameter B; if you remember B was the block size all right.

So, we have done experiments for blocks as a 50, 100, 150, 200 it keeps going on the y axis is the plot of super information please remember for a long chain you only get one number. So, if you carry out for 50, 51, 52 you do all this analysis; you get of wavering value for exons for different resolutions. And similarly for introns you get another set of super information values what is really exciting is there is a big shift. So, somehow it seems that the randomness of randomness in introns is much higher one and a half times that of in exons.

Exons apparently are random, but the randomness of randomness is pretty low as compared to introns. So, information theoretically it comes out as a very very surprising result that they should just separate out like this and this is true for almost all DNA sequences found in the human body.

(Refer Slide Time: 31:22)



If you do a more exhaustive analysis; we have got these 3 d plots just look at the left hand side we talked about the exons the coding sections right. Here you have the gene numbers here we have the chromosome number. So, it basically tells you that these tests have been run for a wide variety of genes over all the 23 chromosomes and you can see that there are variations, but they are pretty much below 3 bits this is the super information level.


If you do the same thing for the non coding introns that the junk part; you have close to 4 and a half to 5 is the bits of super information available there. So, apparently super information is a very effective mechanism for segmenting coding and non coding regions, it is very important for this to find out.

(Refer Slide Time: 32:25)

Information Theory, Coding and Cryptography

An interesting conjecture

- One possible purpose of introns is that they serve as **'bricks'** or building blocks for exons, and are useful for adaptive evolution.
- However, these 'bricks' are of different sizes, probably suitable for different purposes.
- These bricks are **chiseled** into the desired shape and size (exon) during the process of adaptive evolution.
- Figure for Exons is a depiction of the **finished building** (protein) made out of chiseled bricks.
- Figure for Introns can be compared to a **pile of bricks** of different sizes, ready to be used for further enhancements of the building.



So, it makes us conjecture about things what is the purpose of this non coding regions. And one of the ways to look at it is a purpose of intron which is pretty much people are still working on it is that this service bricks or building blocks for exons. So, evolution is there very evolving, but how do we evolve? Well proteins undergo evolution, but who are coding for proteins exons are, but how will new exons come out? I cannot kill my old exons they may get changed once in a while, but maybe the new exons are being generated from the introns. So, the introns form the bricks for building they are the building one. So, this breaks are chiselled.

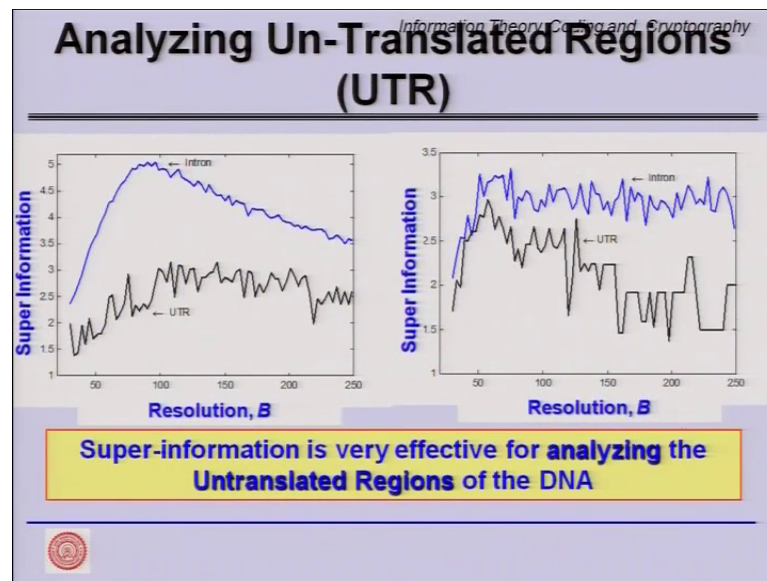
So, these are like your raw material now the moment you choose a let you bring into some structure the moment, you bring in some structure; you are basically reducing the randomness of randomness. Therefore, the moment an intron in this theory becomes an exon you have lowered the super information. So, if you look at a finished building well there is a randomness there; so there is a structure but then rooms are different, walls are different, colors, shade everything is different. So, exons are the finished building and introns are just a pile of bricks lying there.

So, eventually the pile of bricks may convert themselves to the building. So, it is an very interesting.

Student: Cements and cinchonas may convert it (Refer Time: 34:13).

So, this is the conjecture it is not proven that possibly for evolution how will new proteins come into picture well proteins are made out of or coded for by exons and where will they come from. So, introns will get converted into exons and that is just, but information theories neatly telling us, why this is all making sense why this super information must go down; it is a very interesting analogy with a very strong intuitive.

(Refer Slide Time: 34:51)



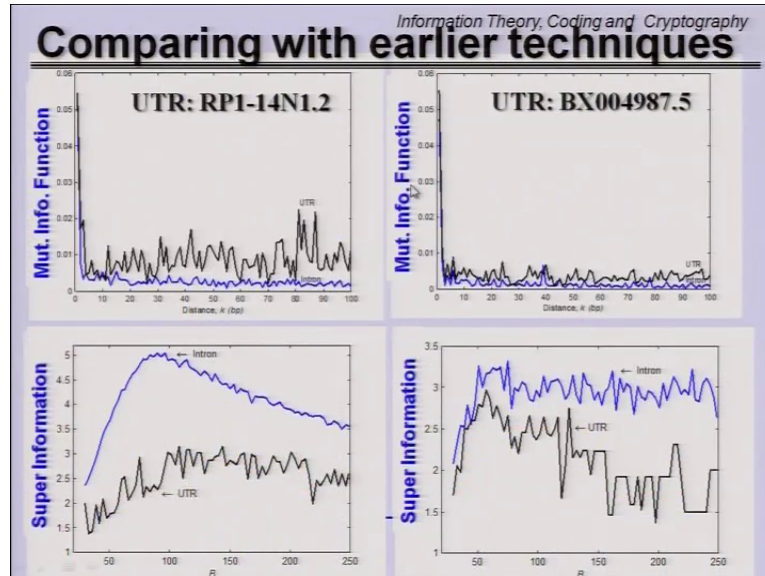
So, another very interesting application that we have is untranslated reasons. So, people are doing this genome sequencing right; human genome sequencing project. So, they know exactly know where the introns are, where the exons are.

But suppose we have untranslated region we do not know where they are and suddenly we have to draw the boundaries; nature will not give you the boundaries. So, if you subject this test to the untranslated regions you will find that some of the untranslated regions are low with respect to a standard intron values and you can immediately say that most likely these are exons whereas, if they are close to introns then that section is likely to be introns.

So, if I draw a horizontal line which is the boundary between the super information values of exons and introns that line itself will tell me anything above that is an intron, below that is an exon. A very neat mathematical way to find out and people go to the lengths of finding the melting points of exons and introns which are different and then

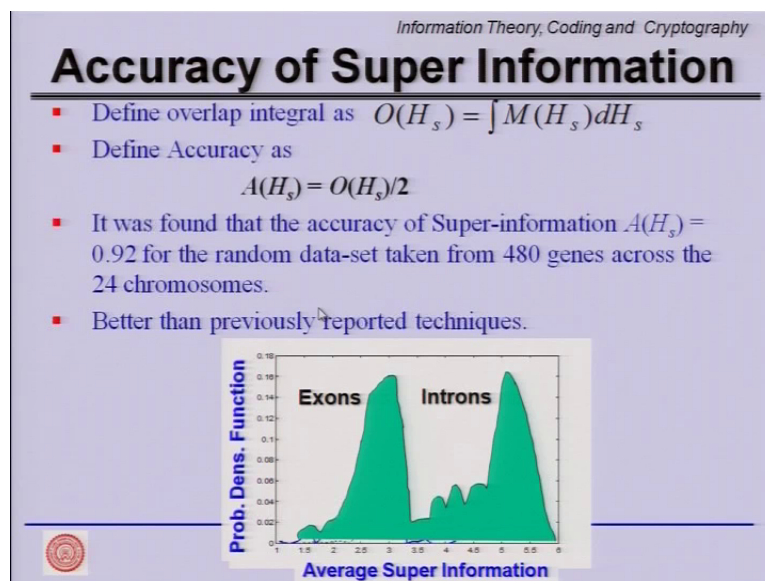
this segment is a very tedious and long drawn process, but see how effective information theory can be.

(Refer Slide Time: 36:13)



So, there are some other examples and you can also look at mutual information function to find out how one portion of the gene is related to the other portion of the gene or genetic sequence and what is the information they convey.

(Refer Slide Time: 36:33)



And if you look at how accurate it is ah; you can always define an overlap integral. So, you have a distribution of exons and a distribution of introns on the x axis is the average

super information, on the y axis is the probability density and clearly there is a distinction.

But wherever there is an overlap; it means that you will make mistakes you enter into the region of exons same with exons entered into the region of introns; they are not mutually exclusive also you can find out what is the accuracy level. So, on a test done on 480 genes across 24; 23 cross was this is 23 plus 1 right. So, 24 chromosomes you have the accuracy of 0.92. So, that is the levels it is pretty accurate in terms of segmenting introns and exons. So, it is better than some of the previously reported techniques.

(Refer Slide Time: 37:37)

Information Theory, Coding and Cryptography

Superinformation and Stock Market

Let $q_o(t)$, $q_c(t)$, $q_h(t)$, and $q_l(t)$ represent the daily opening, closing, high, and low value of the stock quote for a given company, respectively. We used a "binary daily return" for a particular day, d , as follows:

$$R(d,r) = +1 \quad \text{if} \quad q_h(d) > (1 + r/100)q_o(d),$$
$$R(d,r) = -1 \quad \text{if} \quad q_l(d) < (1 - r/100)q_c(d),$$
$$R(d,r) = 0 \quad \text{otherwise,}$$

Now let us quickly look at one other interesting area where super information can actually be used and that is financial markets, stock markets. People make a lot of money on it by predicting half a percent rise or fall a can super information come to our rescue what can it tell us there is enough randomness out there in the financial markets. So, it is a good candidate for subjecting it to the analysis of super information.

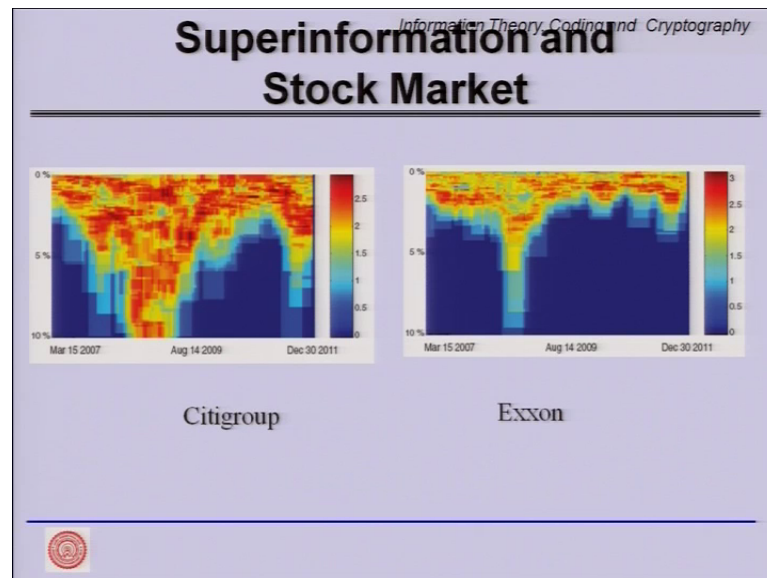
So, again coming back to the practical aspects we always have the opening, closing, high and low of the values of a stock market; you know that right. And what you need to do is buy low and sell high if you want to make money, but you do not know when it is going to happen. So, let us analyze it and again people have used all kinds of techniques, all kinds of information theoretic techniques, signal processing higher order statistics to look at the super stock a market data.

But let us look at a very simple binary daily return. So, we define a term called binary daily return for a particular day as follows what does it mean? Well it means we have a parameter r , but what it basically tells you is that whether you gained or you lost ok. So, that is what we would like to know and it is not just a random within a band of r . So, I am just having a binary gain plus loss minus just understand

Student: That is why qhd?

Yes. So, qhd q h stands for the high alright and r is a parameter we will use r is a band is a sensitivity band. So, if my high is within a band of a main value; we can also define it we can have several definitions we can define it with respect to the mean value the low or the high. Then if I come out of winner by the end of the day I declare it as a plus, but winner is not just greater than it is above a band of tolerance alright.

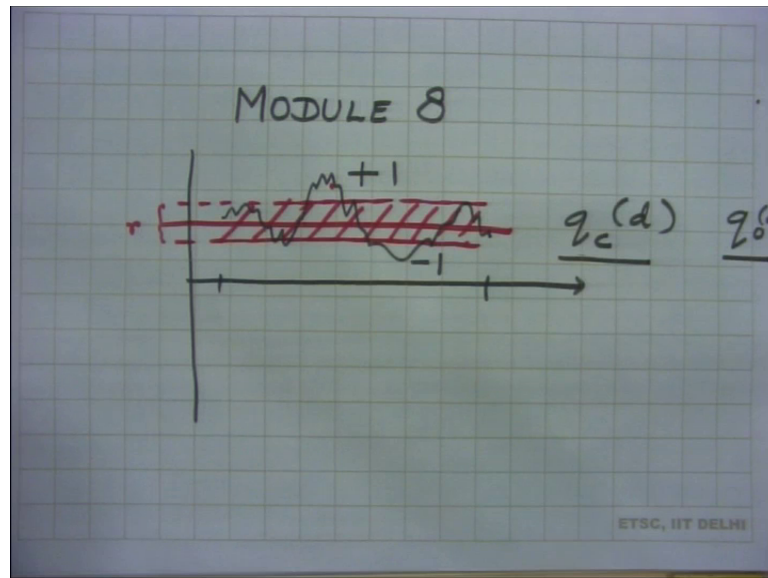
(Refer Slide Time: 40:12)



So, if let us analyze 4 stocks. So, one is Citigroup it us in the financial business the other is Exxon which is in the oil business. And let us understand the axis of this well the color represents the super information value given by this bar.

The x axis is that time ; so, we are analyzing data from 2007 to 2011; 5 years of data. So, all of this data is available on the internet on the y axis is your r the percentage it tells you within this defined band with respect to this band did I come out a winner or a loser?

(Refer Slide Time: 41:22)



So, if I look at it it is my one day and this could be the variation throughout the date and what we do is; suppose we have a mean value with respect to this mean value we can draw a region which is defined with respect to r .

Now if you end up higher up outside this band of r ; then you declare yourself a plus 1, if you come out below you declare yourself as a minus 1. So, for every day because at the end of the day at the closing right quote closing at the end of the day will be related to the quote closing opening quote at the beginning of the day. And I would like to link the closing to the opening with respect to this bar r ; so r is kind of the sensitivity right.

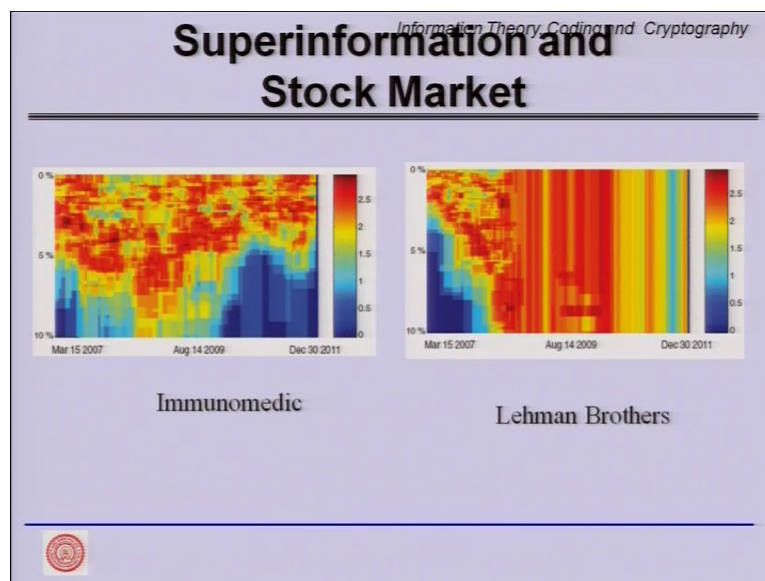
So, we would like to relate these two. So, on the y axis if you look at the diagram on the presentation on the y axis is this r percentage, on the x axis is that time fine. So, if r is large if the r value is large then we are looking at a much more tolerance. So, only when my stock is really volatile when a large value of r still lead to a higher super information. So, red means super information value is high blue means it is cool. Red means that the randomness of randomness is high, blue means randomness of randomness is low.

So, let us fix one r ; so, let us say 5 percent. So, I can compare apples with apples ; so, given a tolerance band of 5 percent and I want to know whether within my my closing quote is within 5 percent of the opening quote plus 1 if it is higher minus 1 if it is lower fair enough?.

So, if you see different days x axis if I move along the x axis in 2007 things are fine blue it is cool and then suddenly around 2008 things start heating up and you find that there are lots of yellows and reds. So, your super information value goes up and if you remember 2008 was the time when the financial crisis hit. And then if you go beyond that again you will go in the cooler region again it cools off and then when you are 2010 or 11 is the recession world recession hit again you can again see this values jumping, but what is more interesting is oil companies are less affected.

So, if you look at the 5 percent value here a long time; you can see it as blue low super information values it jiggles up little bit at during the financial crisis of 2008 it actually pinpoints and then again 2009 onwards it cools down and it does not make really much of a problem. So, the oil companies are much more immune than this from information theoretic point of view if you look at 2 other stocks.

(Refer Slide Time: 45:42)



One is a penny stock penny stocks are really volatile people get into it for short term gains; they put in some money make some money and get out. And everybody is doing the same thing; so this highly volatile nobody holds it. And if you look at this has been volatile all the way only when you increase the band about 10 percent.

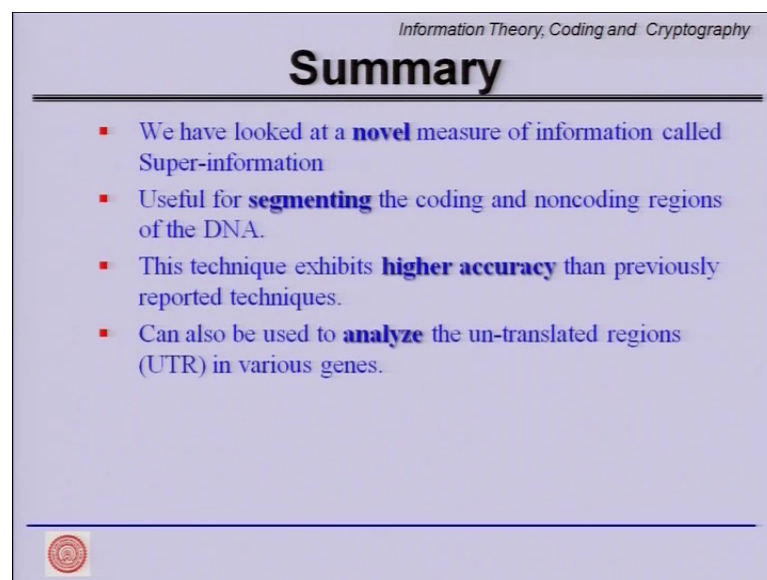
So, if you increase the tolerance do you see that it cools down little cools down it fluctuates beyond the 10 percent? So, the opening and the closing values are more than 10 percent apart that is the r equal to 10 percent band. So, only during the financial crisis

it jiggled way too much and all the other places it if you hold it; it is pretty volatile ok. So, technically speaking what Citigroup went through during the financial crisis that level of volatility was present throughout for immunomedic a penny stock.

If you look at Lehman brothers well if you see from 2007 onwards; there is an increasing trend that is happening and at 2008 it becomes red. In fact, they went bankrupt this is the time when they close shop and after that these are just vertical bars; it really has no meaning right I mean it is just worthless, the stock is worthless.

So, before 2008 before they went bankrupt you can see a lot of activity and how it is good. So, if somebody had the tool available they can almost see and predict the way it was going ok.


(Refer Slide Time: 47:37)



Information Theory, Coding and Cryptography

Summary

- We have looked at a **novel** measure of information called Super-information
- Useful for **segmenting** the coding and noncoding regions of the DNA.
- This technique exhibits **higher accuracy** than previously reported techniques.
- Can also be used to **analyze** the un-translated regions (UTR) in various genes.



So, we come to the summary of today's lecture; we have looked at a novel measure of information called super information; which is essentially the entropy of entropy where we figured out that normal systems in real world are not so, much amenable to analysis by simple entropy terms; we found the super information really adds value. So, we looked at the segmenting of coding and non coding regions for the DNA and we found that this exhibits higher accuracy than previously reported techniques. We can also analyze untranslated region and eventually we looked at the stock market data analysis using super information and it also gives a lot of insight into how markets are behaving; with that we come to an end of this lecture.