Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 21b
Thermodynamic database

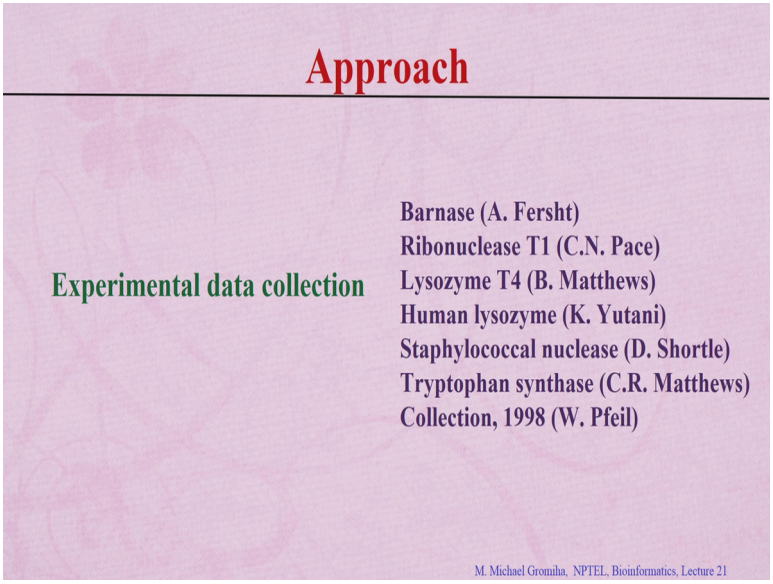
So, how to get this experimental data? Like because as we discussed about different types of experimental techniques. What are different experimental techniques?

Student: Circular dichroism.

Now, circular dichroism, different scanning calorimetric spectroscopy. So, we have different experimental techniques and for each protein they mutate their residue and use the techniques to estimate the free energy change upon mutation. This wild type residue you know like for example, 10 kilocal per mole and upon mutation if it is 12 kilocal per mole then the difference is 2 kilocal per mole. This difference is mainly due to the mutation of a particular residue from the wild type.

So, to understand this there are several experimental groups, they are working on different types of proteins.

(Refer Slide Time: 01:06)



For example Alan Fersht from the Cambridge he is using Barnase as a source right. So, using the barnase protein, they try to mutate different residues and understand the effect of mutations in the particular protein. How far the stability changes at different positions, whether the mutation is any specific residues or any secondary structures or based on (Refer Time: 01:30) solvent accessibility and so on.

So, Nick Pace gives in the new state structures. So, he is working on Ribonuclease T 1 and he analyzed the mutations and showed that hydrogen bonds are important factors for the stability of the particular proteins upon mutation. Then Bear Matthews carry carried out the investigations on T4 Lysozyme, at the 164 residues and he systematically mutated all the residues all the 164 residues and then he analyzed the stability upon mutation. He also get the structures of all these mutants and he compared the 3D structure information with the delta delta G values.

And this case he can be able to tell which mutations involved which type of interactions and specifically how much kilo calorie is required to form any specific type of interactions because of the crystallographic structures as well as the free energy values. Likewise there are different groups like Yutani reported for Lysozyme, David Shortle the staphylococcal nuclease and a C R Matthews on Tryptophan Synthase.

See earlier days they tried to use some proteins and currently we see the literature they are mainly focusing on the important proteins of functions or the disease causing mutations. They try to get the new proteins and identify the important locations and they trying to understand how the stability changes, and how we attribute this stability with the different other features. When they started the experiments in earlier years the systematically carried out the mutations to understand the factors influence the stability.

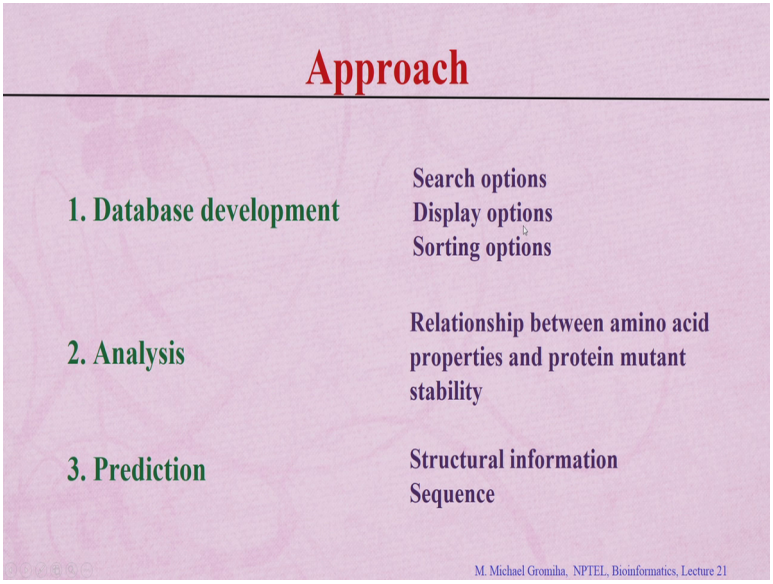
The current scenario because we are going very far ahead, they want to see in particular mutants which are important for function or important for any diseases right. So, they are trying to do it. This is the reason if you see the current literature there are many proteins you know the mutations, but not many mutations in a particular protein. If you look in this T 4 lysozyme or the barnase or ribonuclease you give 100s of data, but several proteins they have only 1 or 2 mutants it is available in literature. So, this data they publish in the literature and they are available in the literature it is very important to collect this information and use it for the application purposes.

So, in 1998 Pfeil is from Germany. So, he collected the data and he published a book on thermodynamic data by Springer. See even they publish your book these not accessible to several investigators and it is required to get the data and show these in the form of computer readable form.

So, in this case what can you do it is very important to develop a database. Earlier days because so not good computers are available as internet facility was not there. So, they collected the data and they published or they kept the data with themselves for analyzing data currently with the availability of fast computers and the high storage capacity it is easily possible you can see they transfer the data. So, it is you can do it develop a database. In one of the earlier classes we discussed about the development of database. What are the various features of database and how to analyze the data from this database right.

So, now in these thermodynamic data for the proteins are the mutants we collect all the data, put in computer readable form and it is very important to provide several options to search the data because each user has a different options to search the data because they are interested on different types of data I will explain some of the details in the later part.

(Refer Slide Time: 05:13)



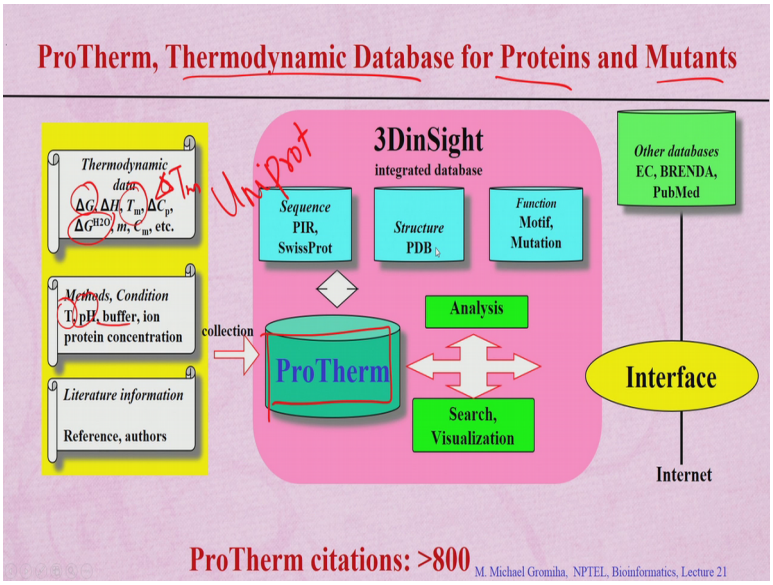
So, it is important to give the search option so that they can search for any data they require and we need to display the results that we need specific format where the user requires.

Then it is also important to give the data to be downloadable in that case they can download the data and they can analyze the data. So, it is important one to develop a database for the mutants as well as for the proteins. Once the database is ready the data are available then you can use it for the analysis, either you can analyze the residues which are stabilizing or which are destabilizing and what are the factors which are important to understand the stability, are there any relationship between amino acid properties and the stability upon mutation, any specific properties which can influence the stability of the mutations.

Once you relate the properties and the stability then it is possible to develop models. In this case you can predict the stability upon mutation for unknown data. So, these three different aspects, the first one is developing a database that should be available and it should have the search options and the result you have to give the option to download. Once the database is ready, then the second aspect, you can analyze the data to understand, what are factors which are important for the stability? Once you analyze the data we will get the hypothesis these are the various aspect of the bioinformatics I discussed in the first class. And then we can develop models on prediction.

These aspect you can use for different aspects for example, stability, binding affinity of the protein interactions, protein DNA interactions, for any applications you can use this type of approach. So, how to make a database, what are the important factors we need to consider for developing a database right.

(Refer Slide Time: 07:15)



So, first you need to put in the important information what you are interested in. Here I put a thermodynamic data for the proteins and the mutants. So, you put the ProTherm here this is the database name. So, mainly we give the data. So, what the experimental data we get? Delta G.

Student: (Refer Time: 07:28).

Obtained from thermal denaturation or you can obtain from the denaturant denaturation because you put a H_2O because if you denature the protein we add the denaturant. So, we get the values any particular concentration. If you extrapolate these values to 0 concentration you will get the value at 0 concentration. This is to have the data uniform.

Otherwise you have the value for different concentrations in this case it is not reliable not good to compare. So, you extrapolate the value to 0, so we get the value of the 0 concentration. Then you can get the thermal you can see melting temperature and the changing melting temperature we will get delta T m we can get. These are experimental data this experimental data depends upon the conditions. What the conditions we use, as well as the method what are different methods we use, use the CD or DSC or fluorescence we use and different conditions for example, temperature pH different buffers and the ions and the concentration. So, the major aspect of the database is the data along with the conditions this is fine.

Now, to you facilitate the users for giving more information to the users you can add the literature where exactly we get the data. Now, if the any of the users they are interested in particular protein or you want to know more about the data then they can go to the reference and they can use that reference try to understand more about the particular protein on the stability.

So, in this case one part is done. Then if you take any particular protein, so you can see it has either particular sequence it has some particular structures if it is known by the x-ray crystallography NMR spectroscopy. So, can you give the sequence information, where we get the sequence information?

Student: (Refer Time: 09:20) uniprot.

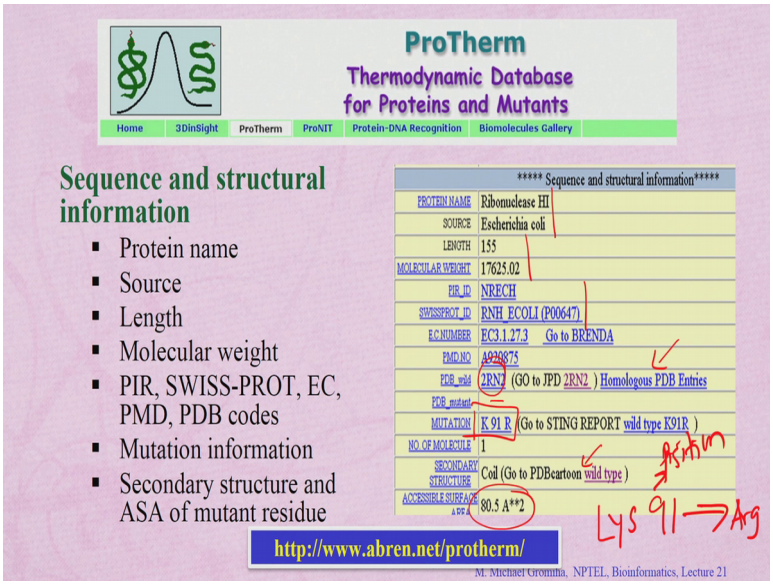
Uniprot, earlier days we have two different databases called PIR. What is PIR?

Student: Protein (Refer Time: 09:28).

Protein information resource and Swiss-Prot and the (Refer Time: 09:30) in uniprot. So, currently we know the uniprot id. Then for the structure PDB this is a unique resource protein data bank. So, you can we give the PDB id. Then, here the some motifs and the mutations for example, if it is mutated it will be the wild type residue and why are we mutate and to which residue we mutated. So, we give all the information. So, when you give all the information this is complete then we can start this, use this database by different search options and you obtain the results and then you can use for your analysis and for developing any algorithms for predicting the stability.

Now, now we will explain about the utilities of database and the future classes I will explain about how we can use database for understanding the factors which influence the stability as well as development of models.

(Refer Slide Time: 10:33)



First part we give sequence and structure information. So, for example, if you have the data right, but we need to see which protein in this case we give the protein name like a ribonuclease HI. So, obtained from e coli because source is also important sometimes we take T4 lysozyme, on the lysozyme you cannot say from the chicken or human then the sequences are different right. So, in this case it is very important to provide the source information. So, we give the name and the source.

But when you give this information you can also get the length and the molecular weight and give the codes where we can get more information for the particular protein. So, we give the PIR, Swiss-Prot and currently we give the uniprot id and this is the enzyme then we can give the enzyme number and we give the link to Brenda. What is Brenda?

Student: Enzyme database (Refer Time: 11:14).

Its comprehensive enzyme database. Say if you give a particular protein is your enzyme then you can give the link. So, you can get more information if anybody interested to know about the particular enzyme. Then PMD number is the protein mutant database they give the literature information regarding the particular mutant and then we give the PDB information. This is the wild type and if the mutant the structure is known mainly for the case of lysozyme in that case most of the mutants the structure is known, then we give the data of PDB mutant, if it is not known then it is blank. Then we give the homologous PDB entries. What is homologous PDB entries?

Student: Proteins (Refer Time: 12:02).

If you can see the protein is similar to the wild type structure for example, 2RN2 you can see the proteins which are similar to 2RN2. So, there if you will you will get as the homologous PDB entries. Then we give the mutation information for example, here K91R. What is the meaning of this K91R?

Student: Lysine.

A lysine.

Student: at position 91 (Refer Time: 12:24).

Position number, 91 this is the position.

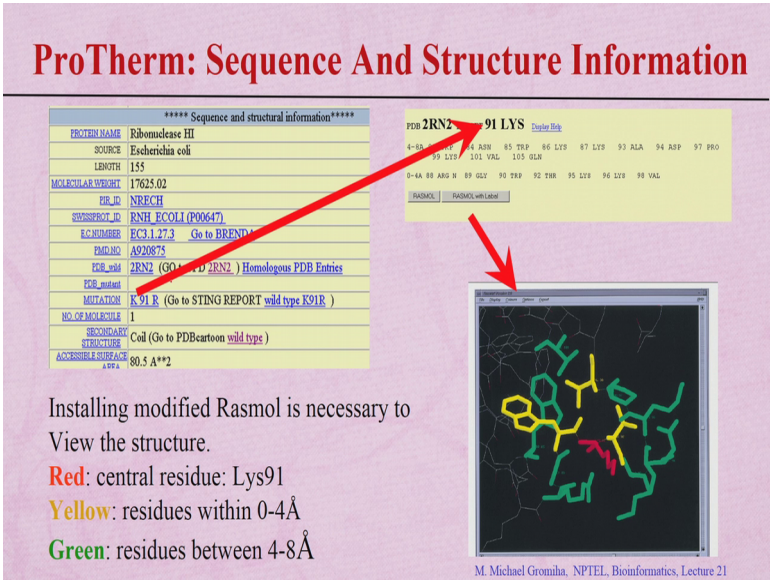
Student: (Refer Time: 19:27) to arginine.

And this is mutated to arginine. Here they did the analysis because lysine and the arginine both are positively charged they want to see the length what will happen. So, this were they changed lysine to arginine. Then we give a secondary structure here this in coil, is coil if you want to see the PDB in the cartoon you can click here. So, you can see the cartoon structure and locate the mutant you can see this in coiled structure. What are the other regular secondary structures?

Student: Helix and stand.

So, here this is in coil region right. So, in accessible surface area, which we discussed earlier, you can see how far a residue is accessible to a solvent. So, this 80.5 angstrom square this buried or exposed? Its partial exposure because it is more than 80 (Refer Time: 13:16), 50. So, you can see a partial exposures or in exposed region. So, if you have a particular mutation we give all the information regarding the protein and the mutants.

(Refer Slide Time: 13:24)

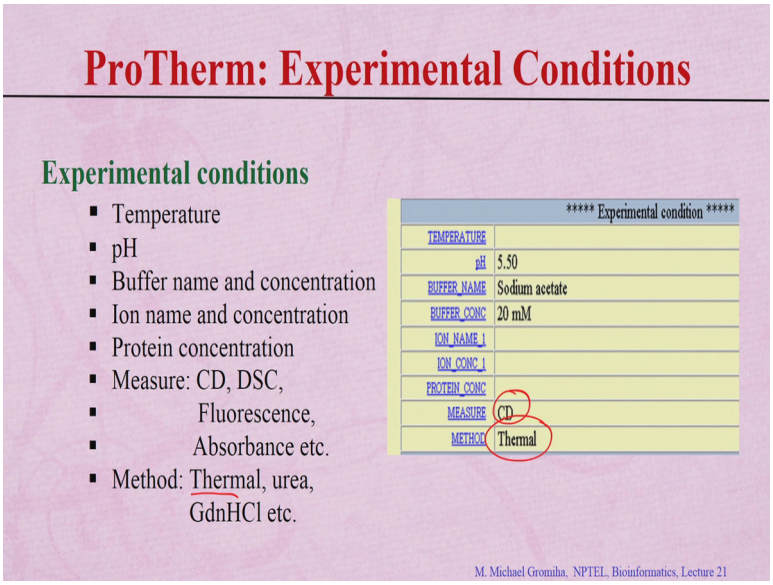


So, then we give the link this is the K is 91 to R. So, we display the mutation you can see the lysine 91 here is the lysine 91, K 91 and see the residues which are occurring near to this particular mutant.

So, which is the yellow one, this one here this is 0 to 4 angstrom, that is 0 to 4 angstrom and the green one, you can see this is 4 to 8 angstrom. Then here I display the data here 0 to 4 is 88 arginine, 89 glycine and the other residues if you see this change here. And if you see this 4 to 8 this is the green one. So, here also you can see the residues which are occurring within 4 to 8 angstrom what is the green residue tryptophan you can see a tryptophan 90 here this is a two rings here.

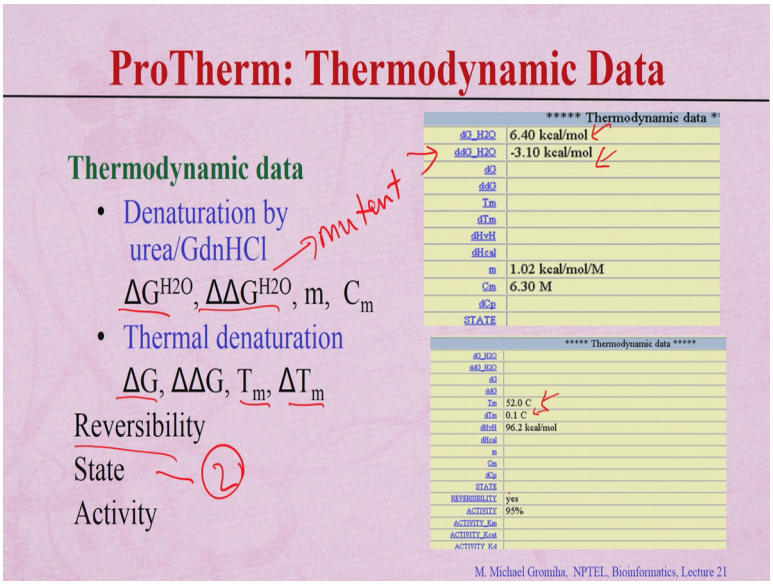
Here also you can see the tryptophan here you can see a, to be here in the 4 to 8 angstrom. So, we will gave a data. You give the picture presentation as well as give the data you can see which are the residues which are surrounded by the any particular mutant residue. So, we gave experimental protein information from the uniprot and the PDB as well as a mutant. Now, we give the conditions what are the experimental conditions they use to get the data.

(Refer Slide Time: 14:47)



The important one is temperature and the ph and the buffer name and concentration on the protein concentration and different measures whether they use CD or DSC or fluorescence, here they use CD and which method they use thermal or the denaturant here they use the thermal. So, what are the information which you can obtain in the literature, whether they report the literature when they do the experiments then all then experimental conditions.

(Refer Slide Time: 15:11)

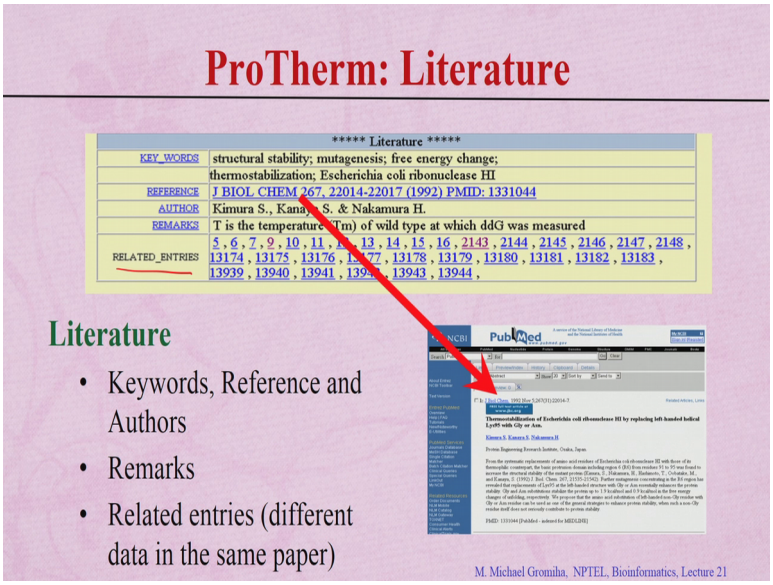


Now, you get the data whether you get the data by denatural denaturation a product to delta G H_2O for the wild type, this is for the mutant. When we extrapolate the data we can get the slope and the midpoint of the concentration we will get the m and C. Likewise when you get thermal denaturation you can get the T_m or the delta T_m actually delta G you can if you know the heat capacity and the any particular temperature you can get the delta G.

In the reversibility what is the meaning of reversibility? Now, if you folded and unfolded state and the other conditions you can come back from the unfolded state to folded state that is called as reversible and then the states that how many states for example, an unfolded to folded state, they are having intermediate states. So, this will give you the state if it is 2 then saying a folded state and the unfolded state, then you give the activity of that particular protein.

So, if you see the data, here delta G H2O is 6.4 and delta delta G is minus 3.1 kilocalorie per mole; that means, it destabilize the protein up to a T kilo cal per mole. So, it is delta delta G H2O. Then you hear the term thermal thermodynamic data we can give that T_m is 52 and dT_m is a 0.1 and you give the enthalpy values and the because no m and C m because thermal denaturation, the denaturant we give all these values.

(Refer Slide Time: 16:46)



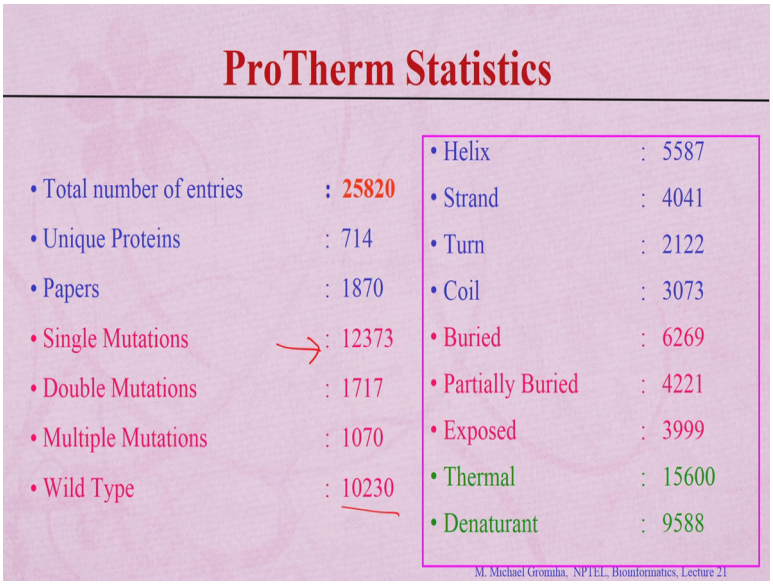
So, reversibility it is yes because it is completely this protein is reversible protein. Then we go with the literature lets we have the protein information and we have the mutant

information and we have the experimental data and the experimental conditions. Now, we link with the literature. So, here we use the keywords are used in this paper and the original reference where we get the data for example, here this is JBC population this one and we give the link to the pubmed, this is literature database from that you can get the pretext for most of the journals. So, you can obtain the papers and if you want to more information it is possible to get the information.

Then here we add one more aspect that is related entries. So, this means so for any particular mutant or particular delta G values you can we gave the data what are the entries which contains similar protein and similar type of mutants. So, these are the entries deal with the (Refer Time: 17:41) and the different types of sources for example, this e coli.

So, you can see all these related entries and you can compare what are the different conditions, what are the different papers and how they got the different types of data and what influences the stability based on experimental information.

(Refer Slide Time: 17:57)



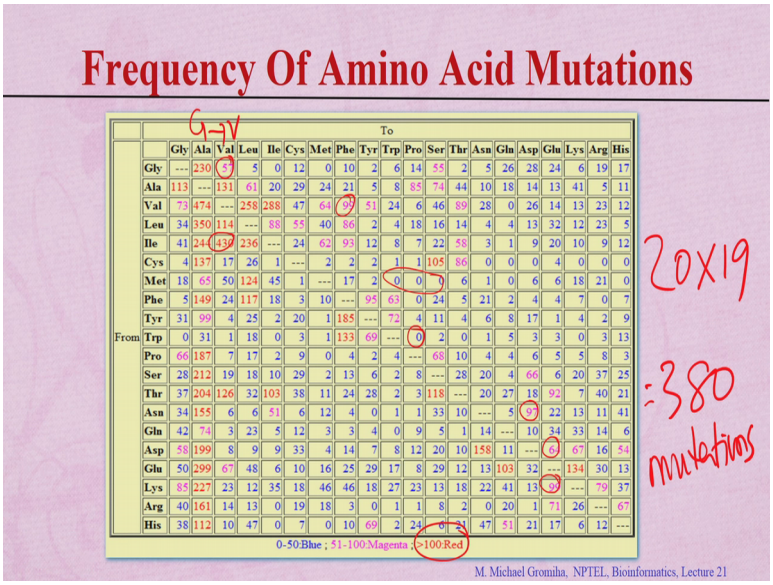
So, you give all the information and we look into the statistics here totally about more than 25,000 a data. So, this is good enough to carry out any analysis and if you see this one they obtained from a 700 proteins, more than 700 proteins and from about 2000 papers. The data are obtained from different proteins about 1800 proteins.

And you look at type of mutations, this is dominated with the single mutation this is about 12,000 single mutants and some double mutations and multiple mutations for the wild type also we have a lot of data about 10,000 wild type data. So, currently if it is PDB that about 130000 structures and wild type 10,000 data, but they are redundant, sometimes you can see the same protein at different concentrations or different experimental conditions right.

In this case if you take the hundred and ones you will get a less number of data, but that you can do for analyzing these a proteins to understand the stability with respect to this contribution of these various interactions. So, we look at the secondary structures. So, you can see the most of them one helix followed by strand and also you have a sufficient number of mutations in the turn and coil. This you can use this information to understand the factors which are influencing the stability in helix and strand whether they same or different you can see, you can carry out this analysis.

In the accessibility about 6000 buried and 4000 in the partially buried and the exposed regions right, so that also you can see the importance of the location of the residues for mutations. Then about 15000 thermal and the 9600 obtained from denatural and denaturant experiments.

(Refer Slide Time: 19:37)



So, now you see the mutations for example, if alanine is mutated to valine, alanine is mutated to glycine, we are showing the frequency of mutations. Here totally how many mutation is possible?

Student: 20 * 19 (Refer Time: 19:51)

20 into?

Student: 19.

19 this equal to 380 mutations. Now here this is from this is 2. So, glycine is mutated to alanine glycine to valine if this means glycine to valine. So, among the 380 possible mutations if you look into this table some mutations they study widely studied and some mutants very rarely studied some cases they are not have studied. Like for example, under the mutants which are widely studied, you are mainly some of the residues to alanine sequence is shown in red is more than 100. See valine to alanine, isoleucine to alanine and methionine to alanine because it is mainly because of these alanines can be muted in a studies they can get the data for the mutate to alanine this is the reason why we get more number of alanine mutations.

Here likewise other than that you can see several other mutations. Mainly similar type of residues, for example, isolation to valine 430 times, like valine to phenylalanine you can see the several types of mutations similar type of mutations. And also you can get the here you can see the asparagine to aspartic acids and the glutamine to glutamic acid lysine to glutamic acid . Why they study lysine to glutamic acids?

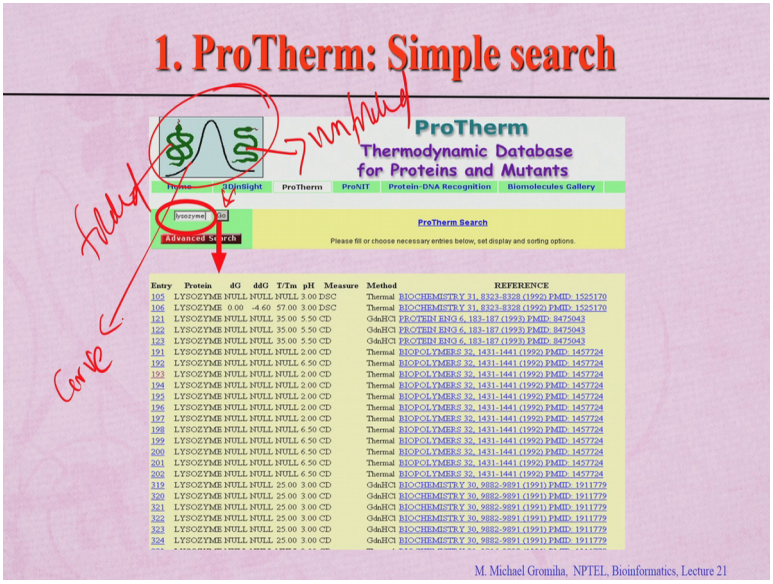
Student: (Refer Time: 21:14).

Now, to see the negatively charged positive charged what will happen if you mutate whether this is stabilized or it is destabilize right. So, here also in some cases this is completely 0 because they want to mutate tryptophan to proline, likewise you can see the other other mutations methionine into tryptophan or methionine into these residues. So, initially we started to analyze different mutants then some of the cases it is really we know that this is stabilize or destabilized and interior core they do not want to mutate (Refer Time: 21:46) to aspartic acid. So, in this case this will they will get destabilize there is a no fun to do it again right. So, they want to know the some cases if it is very

curious. This is the reason they have some mutations to some specific mutants and some cases it is very rarely occurring.

Now, how do you get the data? For example, if you have a lot of data and if you are interested to get any specific data right.

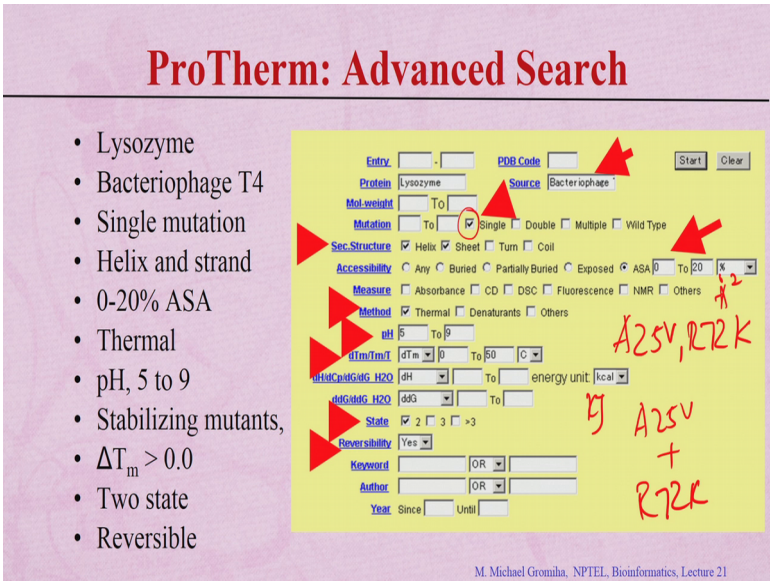
(Refer Slide Time: 22:15)



We have two different types of search options one is the basic search or one is advanced search. The basic search it will search from all entries in the data, especially the key words and other specific entries.

So, if we give the keyword or the protein name or any a source name or anything just if for example, you put lysozyme and click on go then this will show you that the all the data which contains the term lysozyme right. So, all the proteins lysozyme and we get lot of data. This case you will get almost all the information right.

(Refer Slide Time: 22:46)



Then the advanced search we go given different options, you can search with the protein name, you can search with source, you can search with the mutation for example, if you are interested only in single mutation then you can click a single signal mutation.

Then if you want to see the additivity effect of double mutation or multiple mutations for example, you have double mutation like a 25 v and R72K whether the stability obtained with this mutant it is a sum of A25V plus R72K we do not know. So, if you want to do this analysis you get the double mutant and get the single mutants and add of these two and see how far you can relate which is related you can relate or give good correlation or less correlation. If yes why it is yes if you know it is no right. So, if you want to do analysis you can do double mutation the multiple mutations more than 3 and then see how the delta G value changes.

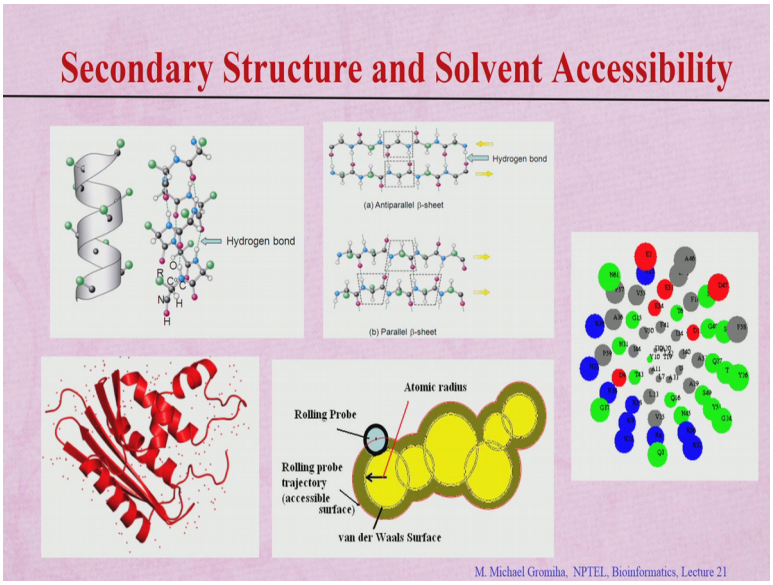
Also in several other aspects for example, thermophilic proteins if you mutate a few residues that will enhances stability more than 10 to 15 kilocalories. If you see the multiple mutations and see few mutations if you increase the stability by 10 kilocal per mole you can try to understand why this is a very high is it possible to introduce this type of mutations to some other proteins that will also increase the stability. In that case you can get the proteins with the high temperatures we can send high temperatures that have a several biotechnology applications.

This you will get the data for a single mutation double mutation and multiple mutation. Then you can try to get the data based on different secondary structures for example, helix or strand or turn or coil. Then accessibility, you get two options we have one is we can get from the buried or partially buried or exposed the data we have already or we can give any range in percentage or you can also do it for the Angstrom square. Angstrom score is the real value obtained from the program and percentage is normalized with the extended accessibility you can do that. Then measure and the method pH you can define because neutral pH. So, I will put 5 to 9 then the delta T_m or T_m for example, if you want to identify only the stabilizing mutant.

In this case you put delta T_m or delta delta G H2O or delta delta G which is you should destabilized then should be minus values, if you do like this you can get the stabilizing structures. Or sometimes if you want to identify the mutants which are extreme stability not around 0, in this case we can see there is minus 100 to minus 2 and here 2 plus 2 to plus 100 then we get the stabilizing and destabilizing extreme values and see what is the difference or if you want to see these minor changes what are the mutants which do not make much changes around 0, minus 0.5 to plus 0.5 if you get the mutations and then see how this is attributed with no change in the stability. All these things if you do these numbers and we can get all the values, energy values you can give either kilocal or kilo joule. So, both are possible in this database.

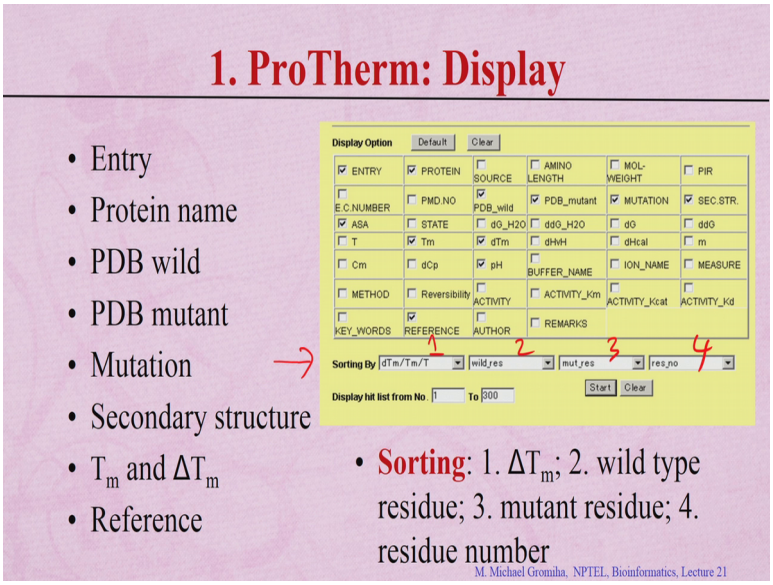
The state you can say two state or three states or more than three states reversibility also you can define. One also you can search with the keywords and authors and then if you want a current data you can put the year you change the year you will give the data which are reported in the literature.

(Refer Slide Time: 26:20)



So, these are the various secondary structures and accessibility that we discussed earlier. So, on this aspect you can search in this particular database.

(Refer Slide Time: 26:26)

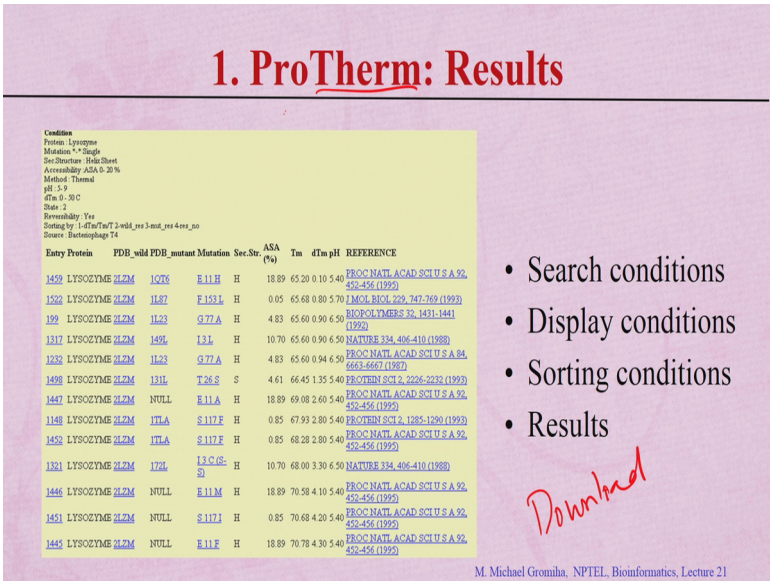


Then you can display options I put all the terms which we use in the database, but if you display all the results it will be a big mess. So, we give options to the users they can choose. For example, if they use thermal denaturation, there is no point to click on the m, C m and delta G H2O because we do not have any data, but denatural denaturation there will be no data on T_m. So, accordingly what they require they can click on the terms and

they will get the data. It is also possible to sort the data based on the wild type residue or the mutant residue or you can see the residue number or the depending of delta G values we have a 4 different sorting option even it can use all sorting option.

It will give the priority this is a priority 1, priority 2 and priority 3 and priority 4 for example, if you have if you start with wild type residue, mutant residue alanine to valine there are many alanine to valine. So, to display from the beginning, then if your third if you use the mutant number then show the first, number less least number at the first. If you want to have this mutation with the highest delta G values then if you start with that one then all the alanine to valine we can see at the first place with the highest delta G or delta T_m. So, that you can also do it by ascending and descending any order you can use it and then we can heat this default is 300, but you can increase it to get the all the data at once stretch.

(Refer Slide Time: 27:51)



So, now, we have the results. So, if we based on your conditions finally, you get the results and it is also possible to download the data for the analysis. So, when you download the data then what are the potential applications? How can you apply this ProTherm data. The ProTherm has several applications we can say first and you can see you can estimate the stability of protein structures and we discussed in a previous class because delta G experimental is known. So, you can do the contributions and you can relate and here we discussed about the mutation. So, we can try to understand the factors

which influence the stability upon mutation either different secondary structure or different locations that you can understand.

Then if you have any data then if you predict any data you can compare the experimental data because experimental data is available. At the third one we discussed about stabilizing residues we can also try to compare the residues identified as stabilizing or also experimentally obtained from these experiments the CD or the DSC.

And you can develop models for predicting the stability upon mutation whether the single mutation or the double mutation or a multiple mutations right. So, you can use various potentials or various switches what we relates the stability with the properties and we can use these properties to develop models. Then you can check the additivity for example, if you have the multiple mutations 3-4 mutations whether the single mutations add up together this is similar to the multiple mutations or is it totally different how can we account that right.

Then you can see the extreme stability mainly for thermophilic proteins, if you make few mutations whether the stability is increased, how to account that stability. Then see the mutants which can change the stability at least more than 2 kilocal per mole, are they destabilized or stabilized. So, you can see these mutant test and analyze what factors for the extreme stability. And also you can see the narrow range of stability that there is not much change. So, we can design the proteins with enhanced stability for example, if we know there is some mutations which are important to enhance the stability then try to utilize these type of mutations right.

And then make a protein a design a protein with enhanced stability which can withstands at more temperatures the potential applications in biotechnology. So, ProTherm is very useful now this is we get more than 1000 citations for this one as it is continuously used by a researchers right. So, you can also use this for any of your projects or your any of your results fine. So, till now what did we discuss today?

Student: Stabilizing residues.

Right, the first part we discussed about stabilizing residues. What are the various factors or properties considered for identifying the stabilizing residues?

Student: Hydrophobic character.

Hydrophobic character and long range interactions.

Student: Conservation score

And Conservation score based on these characteristic features we derived 4 parameters surrounding hydrophobicity, long range order, stabilizing center and conservation score.

And conservation score. So, we get the data for all the residues in protein structures and we make a cut off based on the cutoff you can identify stabilizing residues. And we analyzed this stabilizing residues or some choosing properties and compared with the B-factors and as well as some of the experimentally known information. The second part we discussed about the database what is a name of the database.

Student: ProTherm.

ProTherm this thermodynamic database for proteins and mutant this is why they put the name ProTherm this is thermodynamic database for proteins and mutants right. So, if you see the ProTherm I show this figure. So, here this is a folded state, this is unfolded state and this is a curve. Yesterday I showed about a thermal denaturation curve. This is the way you design this logo to show the thermodynamic stability for proteins and mutants. What information we have in a ProTherm database? Experimental data mainly the thermal stability or the stability upon the denaturant denaturation delta G H_2O melting temperature and supported with the experimental conditions which conditions they used to obtain the data.

Then we have adherent additional information regarding a sequence and the structure and the mutation as well as the literature information. Then we have different options to search the data as well as to obtain the data and sort the data. Once we get the results then it is able to download and you can use the data for any of these potential applications because ProTherm has several applications.

In the next class I will discuss one of the applications how to relate different amino acid properties to a stability of the mutations and the models to predict the stability upon mutations.

Thanks for your kind attention.