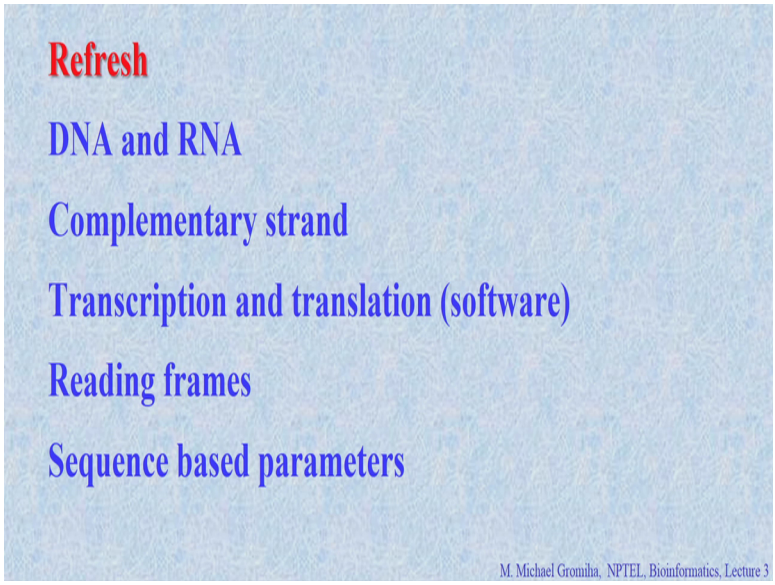
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 3a
Databases

In this class we discuss about Databases. So do you remember what did we discuss in the last class?

Student: About DNA.

(Refer Slide Time: 00:26)



About the DNA and RNA what is the difference between DNA and RNA.

Student: (Refer Time: 00:33).

(Refer Time: 00:34) groups at the two prime side and then the beside. So, we can see the time land.

Student: Russell.

Russell then we discussed about the complimentary stand, if you have a DNA sequence and RNA sequence you can get a device complementarity right, that that will do that and then we discussed about transcription and translation. So, to convert the DNA sequence

to protein sequence; and also discussed about a software which module or which suite we discussed?

Student: Emboss, right.

So, emboss software we discussed and there are lot of options available in emboss right. So, we (Refer Time: 01:07) seek for this complementarity as well as the protein sequence; then we discussed about reading frames. So, how many reading frames?

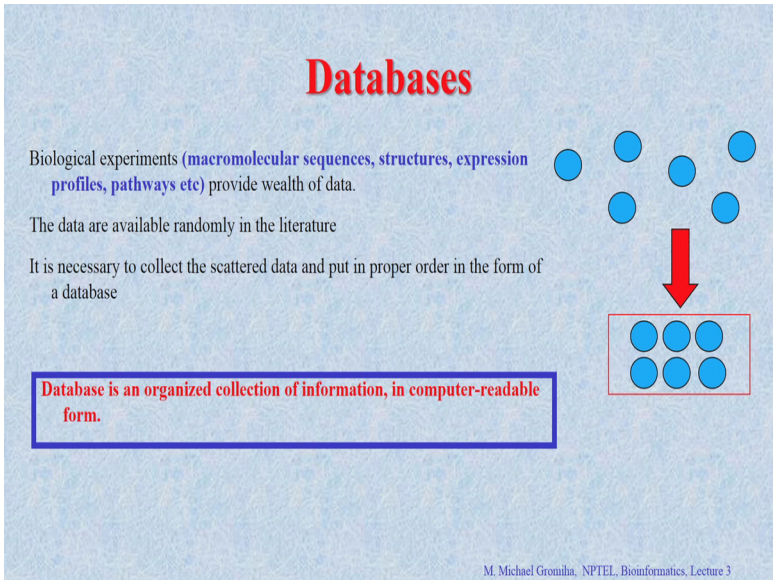
Student: 6.

6 3 forward and three reading frames and we discussed about few of the sequence based parameters, specifically on the stiffness or the flexibility or the based stacking energy right. So, if you discuss about the major aspects of bioinformatics. So, the first class we discussed about 5 major aspects of bioinformatics what are the 5 major aspects of bioinformatics?

Student: (Refer Time: 01:40) databases.

And then get the hypothesis and development of tools around lens hours, and virtual screening of compounds and currently bioinformatics is used in the next generation sequence analysis. So, the first one is databases. So, in this class we will mainly discuss about the databases. So, what is a database?

(Refer Slide Time: 02:01)



Structural collection of information in organized form; so if you look into this literature, there are many databases available for example, if you use the general search which one you use?

Student: Google.

Google you go of whatever information you want, immediately go to a Google and type and then you will get the information right. So, likewise in the case of the bioinformatics, what is the major field in bioinformatics? Biology; biology will produce lot of experiment data. So, you will look into the biology we can get the experiment data on various aspects, specifically in macromolecular sequences and the structures expression profiles pathways and so on. So, if you have a wealth of data available in the literature. Next few lists they create the data and they publish in the journals right. So, this data are available in the literature right.

So, it is very important and necessary to collect all the information, and put it together in a computer readable form. That will be if earlier days we used the collect the data and keep by ourselves; currently due the advancements in the computing and the transferring files and all, it is possible to share the data to public. So, likewise if you see the data which are scattered here and there that means, publish in different journals articles. So, it is pain to collect all the data otherwise, those who want to do the analysis they have to collect the data. So, it is time consuming and everyone has to do that right. So, if you have a database.

On the collection of data in one place, then it is easy to extract the data and use it for the further analysis or the development of any servers. So, here it is scattered in the literature. So, you put it in a order form and in few years ago they put together in the form of book. So, they collect the data for example, thermo dynamic data fail collected more than 22000 data and publish a book. So, if a book then again we need to look into all the details and we had to make into the computer readable form, then there will there are lot of errors, possibility of errors. So, it is important to help the organized collection of this information and a computer readable form.

Because in this case we can reduce the errors and easily you can use the information for further applications right. So, for example, if you want to develop a database right, what are the major characteristics you like to do?

Student: First we have to readable content.

First you have to fix about the contents right. So, who will use the contents how shall we get the contents right, what are the; who are the end users right. First you need to decide about the contents and the importance as well as applications if you develop a database and nobody is using then there is not use to develop such database and putting some much of pains right. So, if you develop a database and that is useful to many others right, then it is very important. So, there will be a very helpful for many researches fine. So, what are the contents then what is other aspects you is to consider?

Student: How to link the how to search you (Refer Time: 05:11).

We need to think about the users perspectives whether a users like to get the data on which aspects for example, if you type something in Google and if you want get in information in Google, if you type will you get the first attempt or you will take many attempts.

Student: Most of the (Refer Time: 05:33).

Most of the first attempts, sometimes it will take time it will what to search again and again and , but what if your search 10 times and when 10 times you do not get the data what will you data will use it again we will not use right. So, the thing is when we search the users perspectives, the user should get the reliable data and the required data that is important. And the second thing is the time; now the world is moving very fast right. So, even they do not wait for even 10 seconds, but if a Google in any of the information if a Google takes 5 minutes right. So, will you use that the Google again you do not use it, that you expect the result in one second.

As soon as we put enter then you should get the result this was the near right. So, the time that is very important and what else we need to think about?

Student: How to details (Refer Time: 06:26).

How to get the display the data because as for your requirements; then the second thing is you need to get the reliable information right. Now if you get lot of whatsapp messages right. So, most of more or 75 percent are not reliable whatever people think, they make and then they send as a viral. But if you get the required information and

reliable information, then you can trust. If you do not get the reliable information if you Google it and if you get lot of lot of information then it is not useful because you cannot trust. So, whatever with the database you develop, the data should be reliable that should be user friendly and this should be reliable information then what else you need to think about?

Student: Well it should be rest ridden then.

Well it should be rest ridden then because if you have the same data if you repeat again and again, then if you going to the users because they are to do lot of works right. So, there should be less redundant. So, you have to give the high integration and let us redundant then what are other aspects you use to think about.

Student: Understandable.

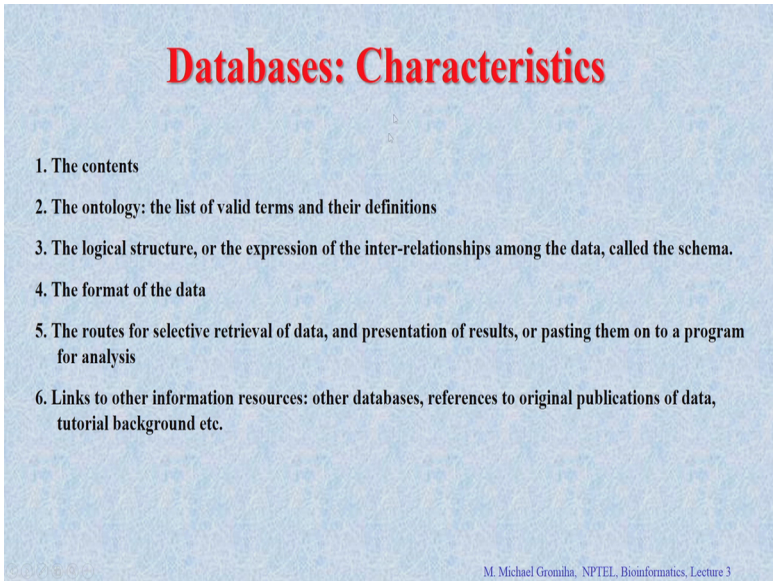
It should be understandable, that you have to give some information you are familiar with the all this information, but the users are not. So, in this case you have to make in such a way that, the user should understand your data that is very important then it what else?

Student: Which should you have come in (Refer Time: 07:48).

Some sort of applications and references as well as you should give option to download the data. You otherwise if you give display lot of information, there will be difficult for the case of the literature it is fine. But in the case of the biological data it should be downloadable. We guess we have riddle plenty of data. So, editing is again it is a issue. So, we can give it to download the data, then several users they use that data database fine.

So, likewise if you look to develop a database, you have to think about various aspects, on these aspects if you consider and develop a database then you will get lot of users when if you will it and then there will be a famous one right.

(Refer Slide Time: 08:31)



Ok. So, now, we will discuss some of the important aspects and what are different characteristics of database. Some of the valid discuss some of them I just I will go through some points. First of all is the contents and that we discussed earlier. So, contents is very important right; there are several database available in literature. So, could you list some of the biological databases?

Student: PDB(Refer Time: 08:52).

PDB protein data bank if you anybody wants to (Refer Time: 08:54) the structures immediately they have to look into PDB, we have that is the unique source and you will have a label information then what else.

Student: NCBI.

NCBI like at the uniprote for the sequence database, thermodynamics (Refer Time: 09:06) our lab also, we developed several databases for the for example, approximate the bending affinity of protein protein complexes right. So, the in all the databases, the contents are very reliable. So, the contents are mainly the experimental data and give you the sufficient information to the users this is for the users required right. So, condense is very important; and the second one is the ontology in case in all databases we get they have to develop several constraints. So, they use various keywords they use various terms and conditions right. So, whatever the biological term they use right.

So, in this case they have took the details for example, you put secondary structure you know secondary structure, but we know we cannot assume that everybody you will know the secondary structure. So, you have to explain. At least in one line you have to write what is secondary structure, what a different secondary structures you consider in your database right. Several accessibility right; the biology is some of them at familiar with the term some of them are not familiar, in this case you have to explain what is that and what are the various terms we consider to define the several accessibility fine. So, all the terms we used you have to give the definition and the third one I put the logical structure right.

So, what is the main aim of a database? For example, if you take the proximate, it is a binding affinity of protein protein complexes that is the major goal. If you take about the PDB putting the main aspect is in the 3D structures, the coordinates (Refer Time: 10:41) coordinates if you have take the proteome. So, in major aspect a thermodynamic data for the proteins (Refer Time: 10:48) folding to unfolding as well as other mutations that is a major one. Now we have the supplement the data with the other information if you take about the thermo dynamics ability, stability changes with the experimental conditions.

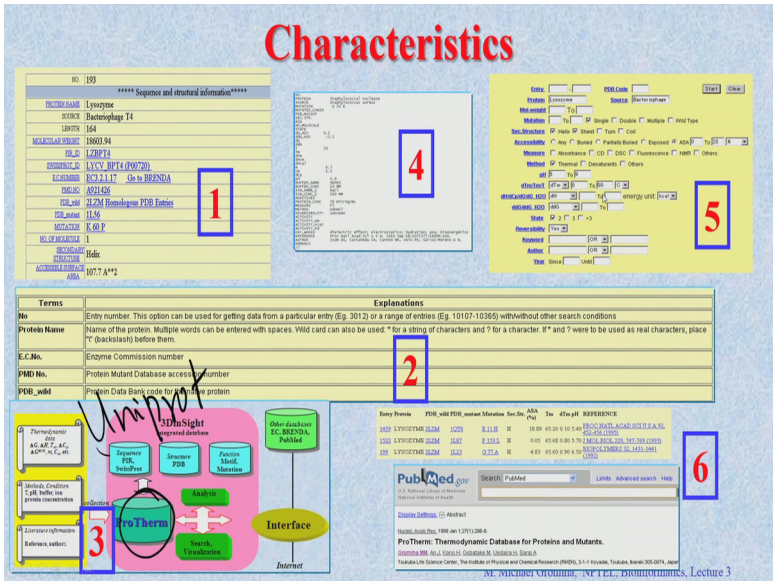
So, you have to provide the conditions, what is the temperature what is the ph and about the buffers ions and the concentration so on. And again we can give more information regarding the structures how the protein or more detail about the proteins as well as the mutations and vary a structure which features all these things we need to give right. So, in this case we have to frame a logical structure, the major thing is your data and how to supplement with other information right. So, in the fourth one is the format of the data because for each entry you give some information if you use various format, then the users will be confused. So, you have to follow uniform format if you look into the Uniprot right.

So, if you see the first line is the name and here is anonymous means and all these this order is the same if you go to the protein databank. So, you to give the compound and the protein aim and the resolution, we see the particular order right. So, in this case you have to give your data in specific format that is very important. I will show some of the details in the later slides, then now we have the overall view of the data what you will have put in and what the terms you use, and how to design your database and what the how do format your data right.

Now, you have the data then put into the website then you should give some sort of sources to retrieve the data. So, you can give some search options this will be new flexibility to the users right, because all the users they do not want the same data some of the users they require some data, some of requires they think about the different types of data. So, in this case you to give a route to select the data from a database and then you have to present the results; that if we present everything together there will be a case you have to present in a way that it is very legible and understandable to the users as a requirement of the users this is fine.

So, then you the data is there now the issue is your database, your contains various other information in this case we have to provide the information where you link the data for example, your main aim is thermodynamics, but few link with the structures or the sequences, then you have to give the link of the structural data or the sequence data like we have to give all the information. Now I will explain with some examples.

(Refer Slide Time: 13:33)



Now, this is why database called thermodynamic database for the proteins and mutants that we developed few years ago right. So, first we give the contents. So, this is the one we do the details about the contents. So, we give sequence structural information if we go to the website, you will get all the details and the thermodynamic data and the experimental conditions literature and so on. And the second one I show the first one is the contents and the second what is second aspect.

Student: Terms.

Terms and condition ontology the terms and conditions. So, if we see these are the various terms we use right. So, for all the terms we use we need to give the details for example, this is a PMD numbers. So, PMD number means we know what is PMD, but the users are not familiar the PMD. So, you can this is protein mutant database number if there is database called protein mutant database right. So, here we give the expansion now for example, PDB wild. So, what is PDB wild? PDB wild means protein databank for the native protein we get this a data database is for the mutants. So, here we give a mutation information. So, for this case you have to give the information, what is the first one what is the second one what is a number right.

Here in this case lies in K is replaced to Proline at residy number 60 right. So, we give all the information in the ontology, then the users now. So, what is mutated to what otherwise this is difficult to understand and what is number 3? Because the schema right. So, here I give a schema now this is our central thing theme is thermodynamic data. So, proteome is here right. So, this is the central theme and here we give the data on the experimental conditions. So, I give thermodynamic methods in conditions and reattach information, they and the data watch the data you put in the proteome.

So, we put the delta G pre and the change delta H and delta TM and the delta G G H 2 and so on this is the free energy change due to the thermo denatuaration or the denatuaration denatuaration melting temperature and so on fine. Now we give the sequence information such a information. So, I put the data here, we get the sequence information if (Refer Time: 15:52) prot currently they emerge to Uniprot right. So, you give a structure as the PDB and the function precise the motif imitations so on. Now we give the other data basis try to for to link with this one, like as the brenda database (Refer Time: 16:09) and the (Refer Time: 16:10) database right.

Sort of we do the enterface so that you connect to the internet. So, you will get the data and you can do download the data and you will visit for the further applications this is what we do. This is the one is the content second the ontology that is schema and the fourth one.

Student: Specific format.

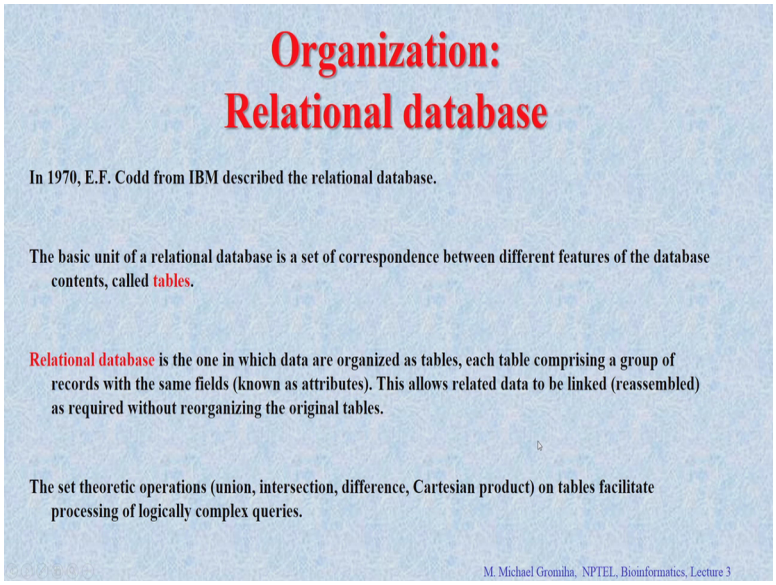
Specific format right; so you give the format. So, so we use a specific format for the a database for example, start with the number this is our proteome number and the protein and the source the mutation. So, they follow the same format for all the data. So, currently we have more than 25000 data, for all the 25000 data we have the unit unified format.

First line means there were condense a number, 10 lines means they have the data G. So, we have the same information because same line say easy to search and interpret a data. Then you provide options because users a have various plug options to obtain the data right. So, here we give the search options. So, users can obtain the data with various such conditions, in the features slight I will explain the details now you have you have lot of conditions. So, user can use any of these conditions with and or to obtain the data. So, when they get the data, they click the start button and they will get the results. Here I tell you with the Lysozyme this. So, we get the data with the Lysozyme.

So, we have it is wire conditions, what are the data which fulfill this conditions display the results. Then we give the link for example, here I link with the reference, which is the article right. So, if you click here, it will get the literature. So, why are you get the data and how they use this fuel conditions all the information you can get from this a proteome database right. So, it is a complete database, which contains the information regarding the thermodynamic data supplementary with other information with proper search option display option and download option. And then other aspects we discussed earlier we should be very fast should be available all the times if you search any of the websites and it is not available for three times then nobody leaves right.

At least if you are server is down, let us should be caution depth currently serve is down it will be available soon that otherwise we will think if they are not maintaining the data base that is very important fine.

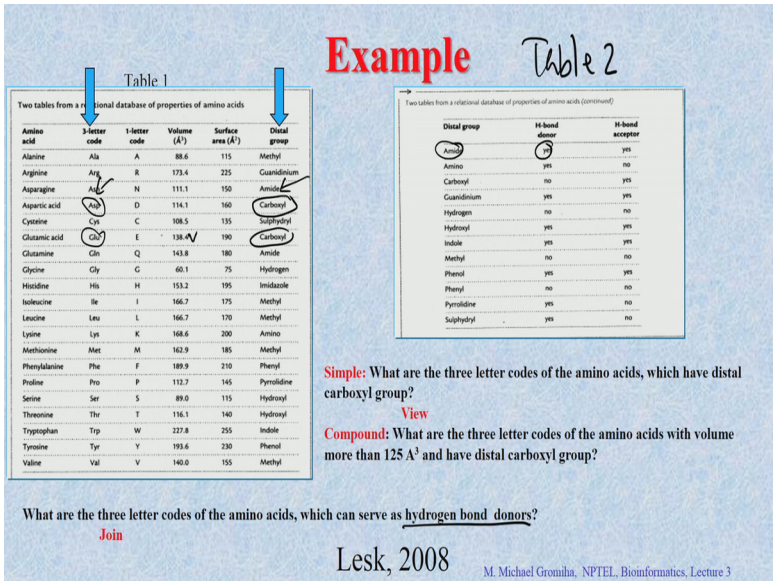
(Refer Slide Time: 18:37)



So, now how to develop a database? There various way should develop a database right. So, in 1970; so E.F Codd from IBM he described over relational database right. So, what is a relational database? It what is a relational database it is a database which gives corresponds between different features in the database right. So, this basic units called tables right. So, what they do they put different information in a table.

So, this would different records are called attributes, and these attributes are interlinked to each other say without rearranging this information you can fetch the data we can extract to data on any aspect. Also there are several operations, we can use we can use the intersection union and the difference and so on to facilitate the processing of the query complex queries, I will tell you one of the examples ok.

(Refer Slide Time: 19:29)



So, you see the table 1, what are the information available in table 1?

Student: One is say.

Amino acid rename.

Student: Three letter code.

Three letter code.

Student: One letter code.

One letter code.

Student: Volume.

Volume.

Student: Surface area.

Surface area.

Student: And.

Distal group.

Student: Distal group.

Where which groups say middle group or (Refer Time: 19:50) group and so on. Now another question is, what a three letter codes of amino acids which you have distal carboxyl group. So, what to do?

Student: (Refer Time: 20:02).

Go to additional group.

Student: (Refer Time: 20:03).

Right the question is carboxyl group. So, what are carboxyl group is here. So, here is the carboxyl here is the carboxyl. Now other the question is what are three letter codes are (Refer Time: 20:18) then why are you have to look?

Student: (Refer Time: 20:19) three letter codes.

Three letter codes second column right. So, it is aspartic acids and.

Student: Glutamic acid.

Glutamic acid that is fine. Now the second question if you see its a compound question the same question what are three letter codes of the amino acids, with volume more than 125 (Refer Time: 20:37) cube and if I have a distal carboxyl group. So, it is the answer the volume should be.

Student: 125.

125. So, which is the answer?

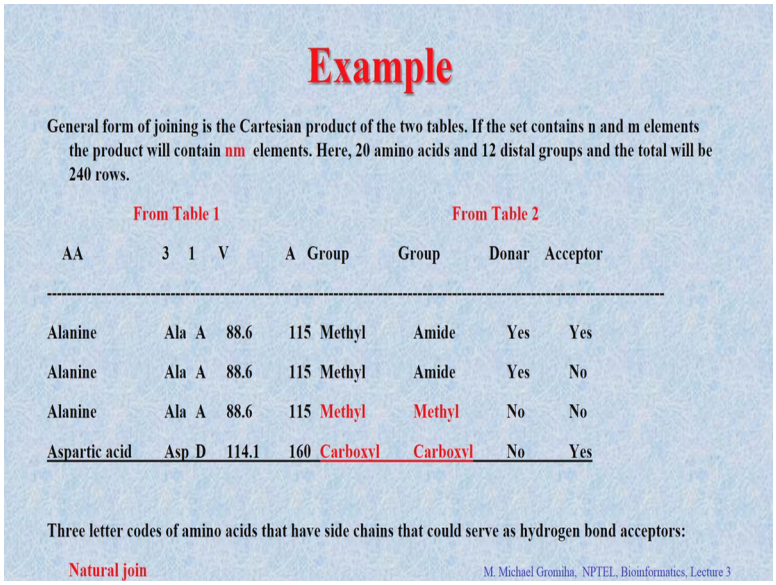
Student: (Refer Time: 20:49).

This one this is 125 fine. So, this is you get the information from table 1. Now next question what are three letter codes of amino acids, which can serve as hydrogen bond donors, this information available in table 1, no this information available in table 2? Donors yes three letter codes.

Student: No.

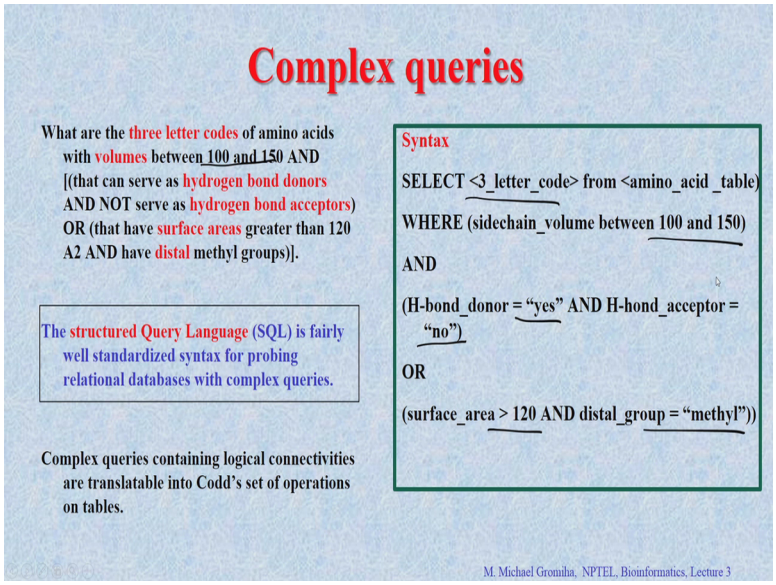
No right. So, in this case you have to use table 1 and table 2 the both the tables you have to use, first go to table 2 this may be consider table 2. So, gets hydrogen bond donors. So, what hydrogen bond donors? For example, take this yes right. So, and the amid then go here with the amid this amid is here and the three letter codes it is one right. So, here you need to joint second.

(Refer Slide Time: 21:53)



If you natural joint you can do table 1 and table 2 if any elements in table 1 and in any element table 2. So, you will get the product say n m elements; so here 20 amino acids and 12 digital groups, so totally around 240 rows. So, for actual alanine, this is the three letter code, this is the one letter code, this is the volume you can see the group and this is from table 1 and this is from table 2 that emerges two and I will get the information fine.

(Refer Slide Time: 22:21)



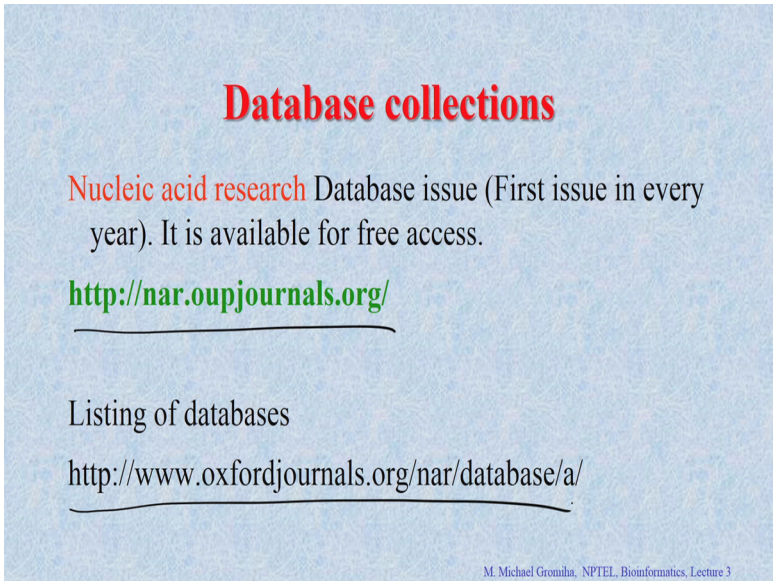
Now, what will happen if you go through the complex queries for example, the question is you have to get three letter code, that fine with volume within 100 and 150. So, this is second question; and the third again that can serve as hydrogen bond donors and not hydrogen bond acceptors or because surface areas greater than 120 and the (Refer Time: 22:44) methyl groups it is very complicated. So, we cannot easily see the tables and then get the data right. So, here we use the structure query language that is called SQL, it is well standardized of the language, we can get the any complex queries we can get the data from the relational database right.

So, here if registration tax first it have three letter code, but it is a coordinate and there a conditions pro wire. Now such an volume within 100 and 150, this will fulfill this conditions. Hydrogen bond donor yes hydrogen bond acceptor? No and then what is next could next condition?

Student: Surface area more than 120.

Surface area more than 120 this will group methyl, now it is easy. You can use these index to fetch the data from this table right. So, you can go to any complex queries, you can write syntax and you can get from any relation database fine.

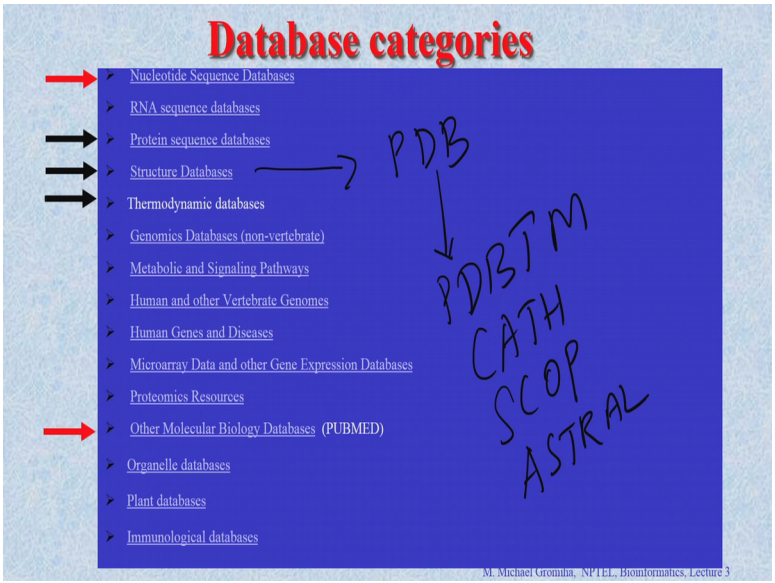
(Refer Slide Time: 23:47)



So, now you show some places you can get the data together. So, there are several database available in the literature, that for if you look into the nucleic acid database issue. So, the nucleic acid research, they publish one special issue every year the first issue of the nucleic acid research then January first issue, the public nucleic the databases. So, if you look at to this a any of websites and then listing of databases.

Currently that are lot of more than 500 database listed in the any database list, and if you go through this database you will get lot of information. Also they classify the data in the various aspects, but these are the different categories right. So, they classify the data based on the nucleic acid sequence databases.

(Refer Slide Time: 24:45)



And RNA sequence databases, protein sequence databases and structure data bases thermodynamics and genomics and metabolic and signaling pathways, and human gene and diseases, proteome databases as well as the other molecular biology databases, plant database, immunological database and so on. So, they have different categories and if you are interested in protein structures, you go to structures and even if get a protein structures right there are various sub classification.

Now what is the image database for protein structure? Protein means a meta database; if you PDB some other image database. Based on the data or in PDB there various sub classifications. So, (Refer Time: 25:2) some classification from PBD.

Student: PDB only (Refer Time: 25:29).

PBDTM what is PBDTM?

Student: Transfer of (Refer Time: 25:34).

Transfer of (Refer Time: 25:35) the database transfer of (Refer Time: 25:36). Also we have cath, write an scope this is for a structural classification of proteins.

Student: Astral.

And astral what is for astral? This one and (Refer Time: 25:48) we can get (Refer Time: 25:49) set of structures using astral. So, you have several databases. So, for each

category for each categories listed in this data base listing, because, let us sub classifications w can search this listing and you can find the data what you want. If you go to the thermodynamics, again you have several classifications plenty of sub classifications and you have go into the details and then see the database available. This will help in different ways for example, if you are looking for any specific data, you can go through the database and check whether the data are available or not and given the data. Second aspect is if you want to develop a database, first you have to check this to be unique there should be a novel. I already if you PDB is available you cannot make another structure database right.

In this case you need a think and check whether such a database is available already in the literature or not; which is available in the literature how far your database will be unique from the existing ones, if not then what are the applications who are end users right. So, you have to think about all the details, even if it is available then you can think who are the users how many is iterations they got, whether still they are continuing or they stopped in between. So, all the information you have to collect and based on the available information you have to develop a new one. In this case this is very important you have to look into the details of the data available in the literature right.

Then this is also help you to design your research problem. For any bioinformatics or computational biology problems you need to have a sufficient number of data. If your data are not available sufficiently then your data base is or not sufficient right, but data not sufficient then we develop a model may not be significant what are hypothesis or the models we make they are not significant. So, you have to that reliable number of data.

So, you can check the literature and the mainly databases and see whether you can get label number of data for a simple if I working on the binding definite of a DNA complexes or putting carbonated complexes. For protein protein we have sufficient number of data you can do it protein nucleic acid, compare to protein protein a very less number of data, carbohydrate even really very less. So check is available or not: if yes, find the data which why are you can find. If not make an new data base and first see whether you can get sufficient number of data. With once you make a database as a discussed earlier so we have sured a reliable information, sufficient number of data, should be when even my users. So, very several aspects you have to think before a development database.