

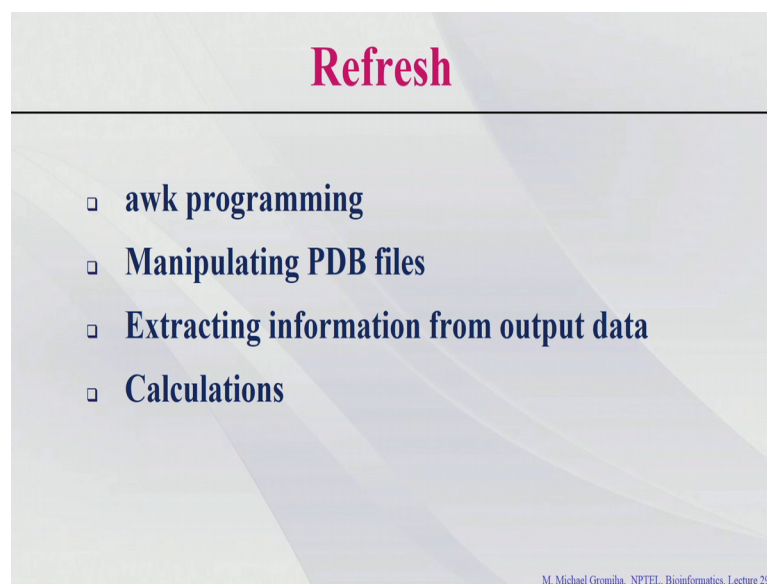
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 29 a
Development of Algorithms I

In this lecture, I will discuss the various steps or what are the various aspects one as to consider for development of any algorithm. So, last class we discussed about the scripting language that is pattern action statement that is awk programming. So, what is awk programming?

Student: So, it is scripting language.

(Refer Slide Time: 00:29)



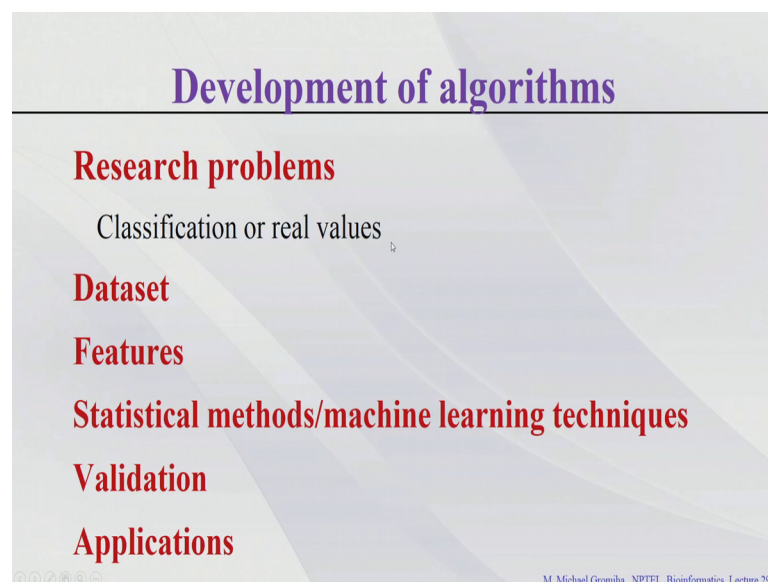
It's scripting. Data driven language, this has a lot of potential applications mainly when we handle data, several data sets or you can manipulate the outputs given from any of these programs. Mainly for this massive data if you want to handle. So, there are various options available in awk programming. Either you can use one line commands or you can make a kind of program ready to run in Unix environment to manipulate several data.

Also specifically we discussed about the usage of this awk commands to manipulate PDB files for example, how to get the C alpha atoms or how to extract the sequence information right, also the how to get the data for atom records and so on.

Then also we discussed little bit about the calculations, we can also use for various arithmetic operations also make a proper outputs we can use this awk in an efficient way. So, now, the question is, we discussed about various aspects of bioinformatics and databases and the tools right. So, if you want to develop your own algorithm on different aspects then what is various aspects we need to consider and what are the available problems. How we can tackle in bioinformatics that I will explain now. First we need to identify the research problem.

So, we discussed various types, the two major types of research problem is one is a classification problem; that is, yes or no either yes or no. So, various aspects whether the particular residue can bind or no they stabilize or not right. This is a kind of classification problem.

(Refer Slide Time: 02:20)



And second one is a kind of exact real values then for example, is 4.5 right. So, what is the temperature this is 21.2 exact we can tell this is the actual value. So, two types of problems, this is generally occurring in various fields of bioinformatics right. So, I will explain some of the aspects in this lecture.

So, once we would define or once you identify your research problem then how to proceed right. There are various aspects we need to consider when you develop any algorithm. First I will, for example, the dataset that is very important because your, for the final results and the model mainly depends on your dataset. Earlier we discussed about the data redundancy how to construct non-redundant dataset and what is the importance of non-redundant data set right.

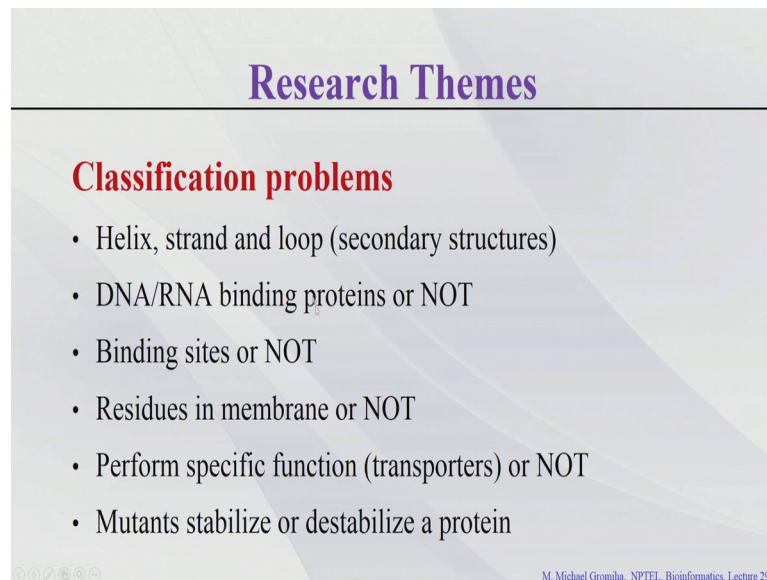
I will explain the details about all these features. Then once you make a dataset, then we need to identify the important parameters or important properties what we can learn from this dataset. Anything is very specific? For example, if you consider the classification problems for example, A or B. So, what is peculiar or what is very important in A and what is very special in B. So, we could extract this information then you can use this information to classify these two groups A and B.

So, then we use some methods, there are various methods like statistical methods or machine learning techniques, you can use this, subjectively you can see, put some conditions and we can classify. If this particular value is positive or negative then this is for class A or class B and so on. Or you can use some machine learning techniques, here this we can, if the machines if we give all the input data and output data and your features. The machine will train the data and it will give you the most probable output right. Then you can use that model for predicting any new set of data.

Once we get this method and get the result then we need to validate. First, we need to assess the performance once your performance is good then you need to validate whether this is valid for different sets of data that we will explain and the applications. So, what are the potential applications for your method and what will be the various aspects you need to consider to get more visibility of your particular program.

So, I tell you some of these research themes I just now discussed the two different types of problems one is classification problems.

(Refer Slide Time: 04:58)



Research Themes

Classification problems

- Helix, strand and loop (secondary structures)
- DNA/RNA binding proteins or NOT
- Binding sites or NOT
- Residues in membrane or NOT
- Perform specific function (transporters) or NOT
- Mutants stabilize or destabilize a protein

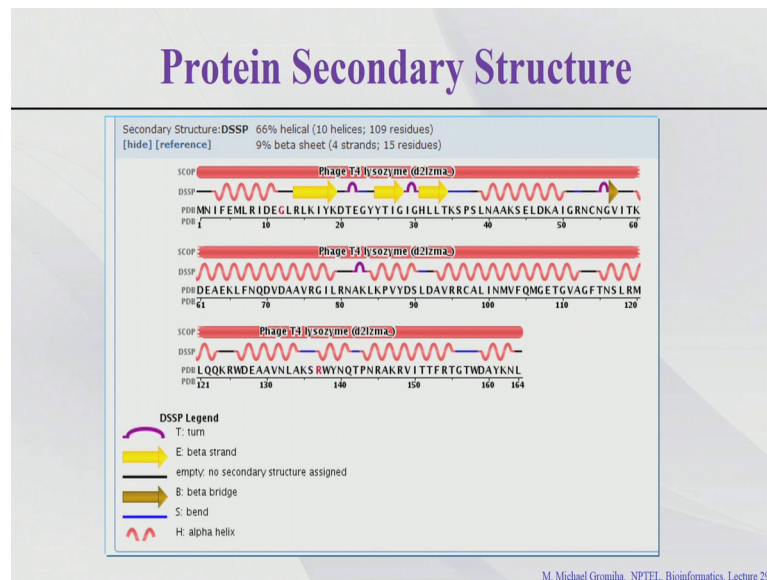
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 29

Another one is exactly predicting the real value. Classification problems will tell you some of the important areas. So, one is whether you can predict the secondary structures right. So, we discussed about secondary structures what area various secondary structure

Student: (Refer Time: 05:08).

Helix, strand, coil turn right whether have can we predict? Then we discussed about the DNA binding proteins or the RNA binding proteins whether you can identify the sites this is binding or not. Likewise we can see the membrane, whether where the residues are located right. So, then you can see some functions. So, if you have a sequence whether this can perform the transport or not might you can do that then you can see the mutations that also we discussed in the earlier classes whether this will stabilize or destabilize?

(Refer Slide Time: 05:39)



I will tell you more details. So, if we took the secondary structure. So, if we see some region they are helix, some region strand and also different types right. So, if you have this sequence and if you do not know about this secondary structure, you can predict whether in particular segment this can form helix or can beta strand or it is a coil. If you want to have two state problem you can see this is from this from helix or not or strand or not that ways you can see in the classification problems we can make two state classification or you can make three state classification or we can see the go for the multi state classifications

So, in this case depending your secondary structures either you make two states of you can make three state if you want to three state you can take helix, strand and coil, two states means you can take helix or not or strand or not or you can have the four set classification helix or strand or coil or the turn you can do that then you go with the accessibility right. So, how to get the accessibility?

Student: by probing (Refer Time: 06:46).

(Refer Time: 04:46) 1.4 angstrom radius water. So, you can get the residues (Refer Time: 06:50) which are accessible to solvent that you can get the data. So, there are various programs to get the accessibility. What are the various programs to get the accessibility?

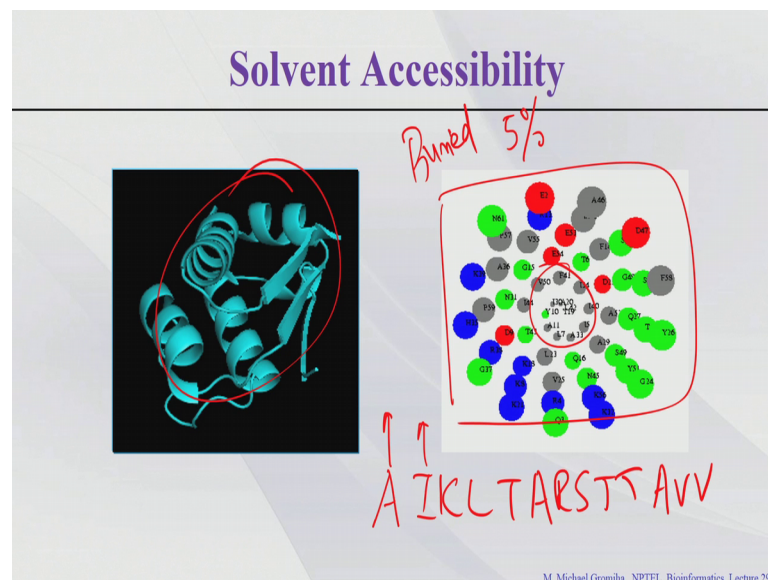
Student: DSSP

DSSP, we can get. NACCESS and many programs you will get the accessibility. Likewise the secondary structure whatever the programs we use to get this secondary structure.

Student: Again DSSP.

Right, you can say DSSP to get the secondary structure assignment also (Refer Time: 07:10) depending on the (Refer Time: 07:10) hydrogen bonding pattern you can see in experimental data. So, here we have a solvent accessibility with this is the protein. So, you can see the protein. So, we can use the concept of solvent accessibility and finally, you get this figure right.

(Refer Slide Time: 07:14)

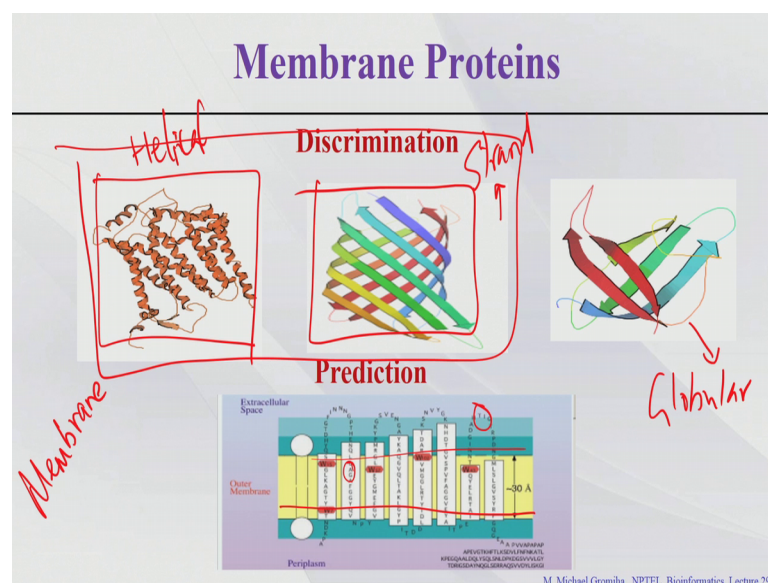


So, here there are two different aspects one aspect you can see the residues which are in buried which are the residues buried especially you can see these residues they are in buried now we define a particular sequence for example, this is a sequence. So, take this one this is buried or exposed. So, this is buried or exposed you can see this two state classification. We can make any cut off and you can make developed model. So, to see in this proteins this residues are buried and which residues are exposed two set classification or you can make three different classes you can take buried you can take exposed and you can make a cut off that is partially buried or partially exposed you can see that.

So, we can take any cut off for example, if we take the buried you can take a cut off of 5 percent accessibility and exposed we can put more than 50 percent and in between you can put partially exposed right. If we take two set classification you can put a cut off. For example, it is 20 percent less than 20 percent you get buried and more than 20 percent take it as exposed. Likewise for any amino acid sequence you are classifying these residues based on solvent accessibility. Why do you need this solvent accessibility? Once you predict the accessibility then you can see the location. So, which residues are interior and which residues are at the surface and which type of interactions these residues can make.

So, if you have a sequence you have predict. There are various methods to predict the accessibility you can also develop a model with a better accuracy right. So, if you have the sequence you can use some parameters and you can predict these two state problem.

(Refer Slide Time: 09:07)



So, now for example, this is a kind of membrane protein. So, for example, if you are interested in this type of membrane proteins, then what is this type of membrane proteins?

Student: (Refer Time: 09:16).

Alpha helical proteins. This the transmembrane helical proteins. This is strand proteins. If you are interested in this one say we have a set of sequences and we have two

sequence we can say the sequence belongs to the strand type or not, or this protein belongs to this helical type or not because these proteins perform different functions. In this case it is very important to identify the proteins in any genome either this is helix or this is strand you can do that or you can say this is membrane proteins or not if we take this as whole these are membrane proteins.

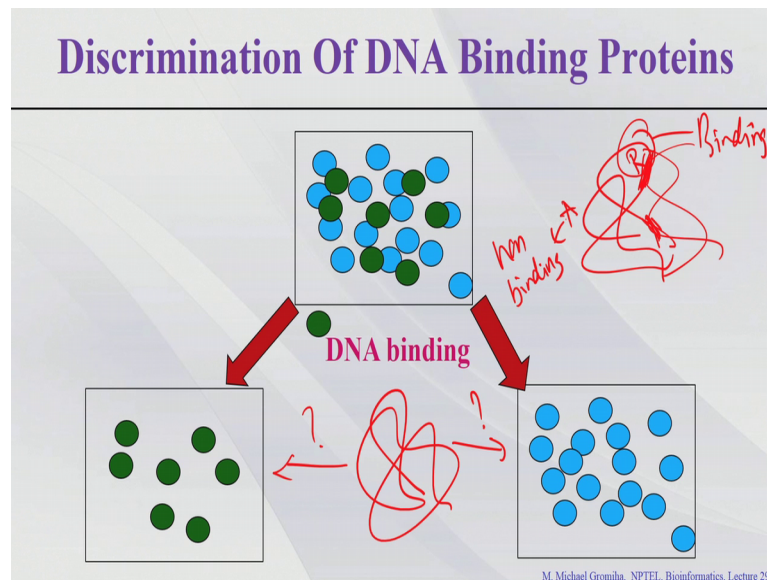
Here if we see this is the globular protein. So, if we see can get a sequence you can identify where is the globular proteins or it is a membrane proteins you can see that, right. So, now, if the case if the proteins membrane proteins. So, you identify a protein as membrane proteins what is a membrane proteins the proteins which are.

Student: (Refer Time: 10:24).

(Refer Time: 10:24) embedded in biological membrane. So, we have this is a membrane then we can see this is a membrane right. So, the proteins there are some several residues which are embedded into membranes. So, the question is if we take an particular residue this is inside the membrane or not.

So, take another residue here this is in the membrane or outside the membrane because this will have functional implications because to understand the functions right. So, this is very important to see whether any residue is located within the membrane or in a surface. This a kind of classification problem. So, go in step further again. So, here we are talking about the binding. Till now we discussed about the single features based on the structure, either helix or strand accessible or not accessible are the proteins membrane or not right. So, now, it get a little complicated. So, we gave the composition for example, here DNA binding proteins.

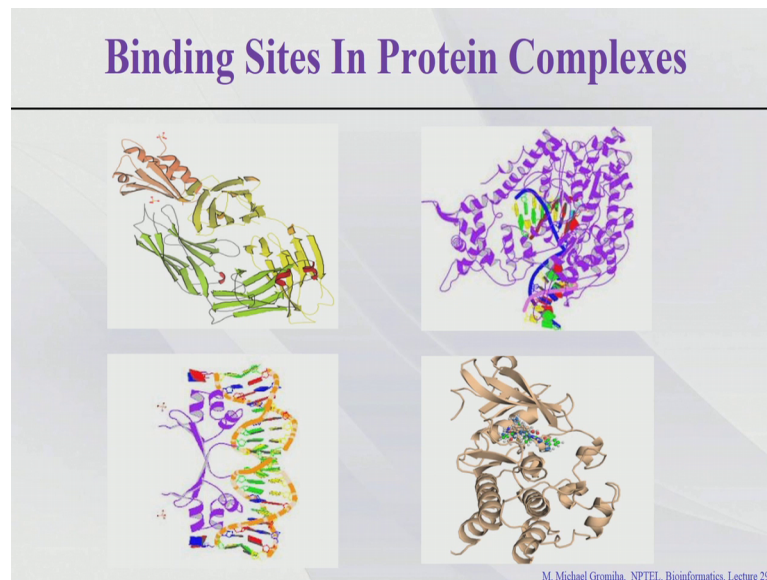
(Refer Slide Time: 11:18)



Let us say we have proteins here and we have the DNA right. So, we can see some part you can see this interacting right.

So, protein as one sequence and DNA as a sequence and which is the sequence the question is whether it is binding or not then we can see such a classification problem whether where it we can bind. Also we can see whether a particular protein is binding a DNA or not. For example, if we see a different types of proteins. The green one is the DNA binding and the blue one is the non-DNA binding, then if you take any protein whether this belongs to this group or this belong to this group. Is binding with the DNA or not? If it binds with the DNA, then the question is for any residue that residue binds or not. For example, this residue is A. This residue R then which one binds R binds. So, this is binding and this is non-binding. So, we can do that.

(Refer Slide Time: 12:37)



So, likewise if we see the DNA you can see protein-protein complexes, protein-RNA complexes, protein-ligand complexes because these are very important. In the. Because they performed various functions we discussed in the previous classes and structure based drug design they are important to identify the binding sites and the pockets right. So, in this case if you have a sequence you can develop methods to identify the residues which are binding this is a classification problem.

So, this is another important one. So, some a cases if you have a protein, some regions are missing in the structures. For example here some region these residues in red. So, they are missing in the protein site, but when it makes a complex with other molecules like other proteins or the DNA or the RNA in this case they get the structure. So, the partner one, they promote the folding right. So, in this case they get the structures. So, if we give a sequence right.

(Refer Slide Time: 13:15)

Disordered Regions

Ex : 1 Free protein : 1BAM:A Complex : 3BAM:A

```

MEVEKEFITDEAKELLSKDKLIQQAYNEVKTSGSPWPATSKTFTINNTKNCNGVPI
KELCYTLLEDYNYWYREKPLDILKLEKKGGPIDVYKEFIENSELKRVGMEFETGNISS
AHRSMNKLGLGLKHGEIDLAIILMPIQLAYLTDRTVFEELEPYFELTEGQPFIFIG
FNAEAYNSNVPLIPKSGDGMKRSIKKWKDKVENK
          
```


Ex : 2 Free protein : 1ES8:A Complex : 1DFM:B

```


MKIDITDYNHADEILNPQLWKEIETLLKMPHLVKASDQASKVGSILFDPVGTNQYIKD
ELVPKHWNKIPIPKRFDLGTIDFGKRDITLVEQFSNYPLNNTVRSELFHKSND
IDDEGMKVAILITKGHMFASNSLYEQAQNLNLAEYNVFDVPIRLVGLIEDFETD
IDIVSTTYADKRYRTITKRDVKGKVIDTNPTRRRRKRGITVY
          
```

The regions which are transformed from disordered to ordered state during complex formation are indicated in red color which are encircled in the figures

3BAM



1DFM

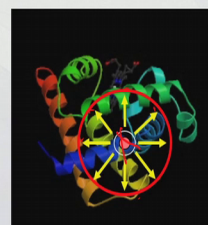


M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 29

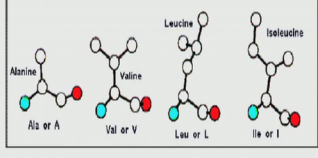
So, If you give the sequence, can we predict which regions are disordered regions from sequence. If we know the structure we can find here the density map is missing. So, here structure is missing. So, this are disordered regions, but if we have the sequence is it possible to predict whether this residues a disordered or not that we can do that for example, here this is the region. These are the ones that has structure in the complexe, but this is missing in the case of free proteins.

(Refer Slide Time: 14:07)

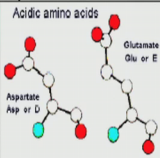
Protein Mutant Stability



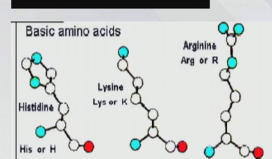
~~A~~ \rightarrow \rightarrow V



Acidic amino acids



Basic amino acids



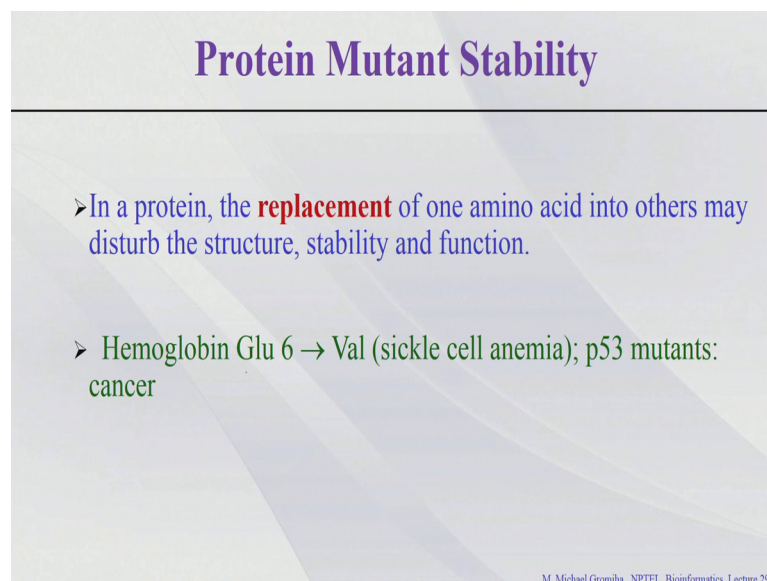
Ala \rightarrow Val: collide OR well packed
Val \rightarrow Ala: cavity OR freely move
Ala \rightarrow Asp: electrostatic
Lys \rightarrow Leu: break ion pairs

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 29

Then the next aspect is. So, we can see stability. We discussed about stability for example, if you mutate a particular residue here. This residue is in contact with various other residues. These are the residues which are in contact. We mutated this residue either it may stabilize more contacts or it may break some of the contacts this may increase the stability or decrease the stability. For example, we discussed earlier alanine to valine depending upon the place where we mutate this can either a well packed which increase the stability or it can have the steric and decrease the stability.

Like the alanine to aspartic acid, either in surface you can make the electrostatic interactions with the lysine arginine or if it is in the buried then this will disturb the protein destabilize the proteins right. So, like this if you have any mutation for example, Alanine15 is mutated to valine in a particular protein whether, this stabilize or not, now question is a two state problem this will stabilize or destabilize right. So, you can take this is a problem.

(Refer Slide Time: 15:16)



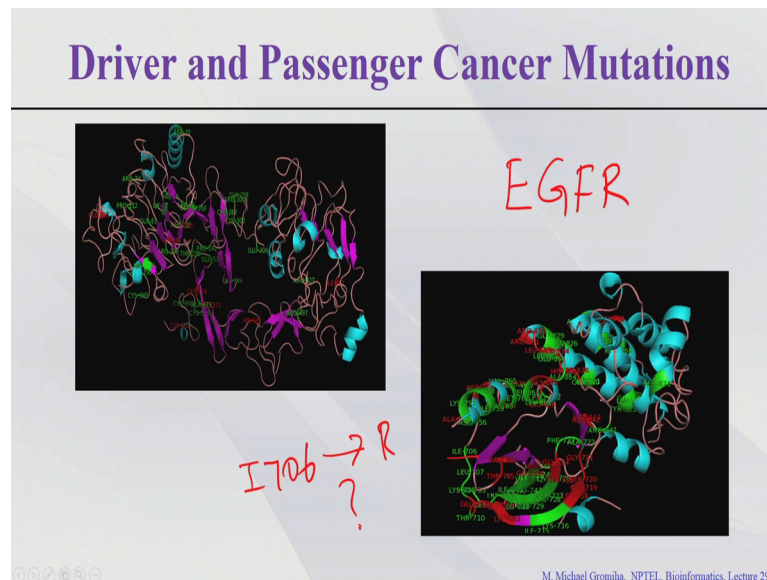
Protein Mutant Stability

- In a protein, the **replacement** of one amino acid into others may disturb the structure, stability and function.
- Hemoglobin Glu 6 → Val (sickle cell anemia); p53 mutants: cancer

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 29

So, like this we have the cell diseases for example, glutamic 6 to valine sickle cell anemia hemoglobin take p53 there are many mutations then which can cancer. So, here also we can see whether this mutation rights it is causing tumour or not, we can this is a kind of two state problem.

(Refer Slide Time: 15:36)

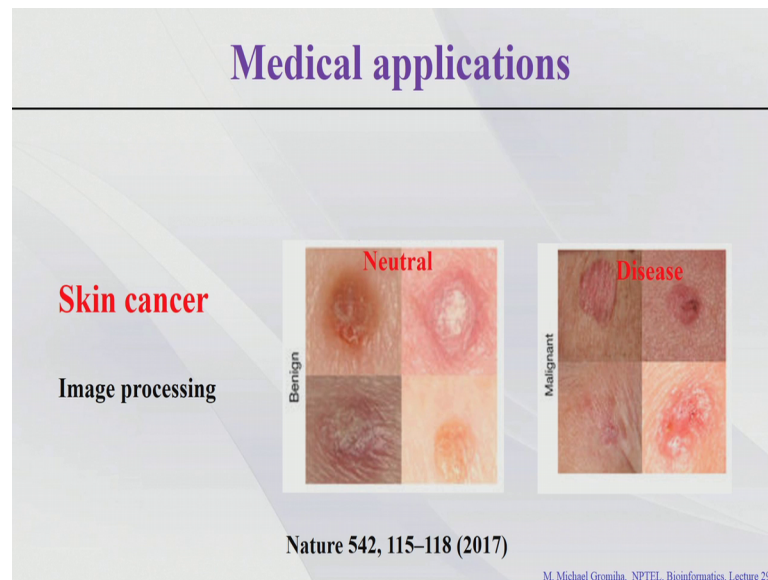


So, here I am going to show one example this is the data for the epidermal growth factor receptors here two different domains and if we see some letters in green as some of them are in red and the green one are the drivers; that means, they cause, they can cause cancer for the progression of the cancer. So, the green ones this is the passenger; that means, even if you have the mutations there is no change.

So, the several type of mutations red one is the driver mutation and the green one is the passenger mutation right. So, you can see now if you have a particular mutation for example, is Isoleucine 706, I706 if we mutate to other residues for example, arginine. So, what will happen this passenger mutation or a driver mutation you can classify, you can take it as a classification problem whether driver or it is passenger.

Likewise if there are several image processing you can do the images we can scan several images here I show two types of images right.

(Refer Slide Time: 16:42)



So, here one is this benign now there is a nothing do they cancer, but here is a malignant right. So, can we see these images and then predict whether this is a cancer prone or this is a neutral one, this we mainly deal with the skin cancer.

So, you have the two types of data either you have the numerical data or you can have the image data, so even both cases, you can have different types of problems like you can classify into different groups. Like ways you can see the several problems in bioinformatics that you can classify where these are there.

Then you go with the real value prediction here there are various problems which can give exact real value.

(Refer Slide Time: 17:30)

Research Themes

Real value prediction

- Stability change upon mutation
- Binding affinity of protein complexes
- Protein folding rates
- Solvent accessibility

Handwritten notes:

- $A15 \rightarrow V$ (with arrow pointing to "Stabilize")
- $A15 \rightarrow V$ (with arrow pointing to "Destabilize")
- 1.2 kcal/mol
- $A1A \rightarrow 12A$ (with arrow pointing to "Buried")
- $5A^2$ (with arrow pointing to "Exposed")

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 29

For example, a stability change upon mutation for example, A15 is mutated to valine. Earlier we discussed whether this will stabilize or destabilize. This is a classification problem, but you can see also this will decrease a stability of 1.2 kcal per mole. So, exactly real value, if you mutate any specific residue you can measure or you can predict the exact value this will change the stability by 1.2 kilo cal per mole this is a real value you can do that.

Like binding affinity of proteins complexes if you have one protein this is another proteins, if you compare this proteins sites when they form the complex the affinity what is affinity how strong these two proteins can interact. Like proteins folding rates, we discussed the folding rates how fast and how slow a protein can fold right. So, you can see whether a if you have a protein sequence, can you predict this is the folding rate then you can predict.

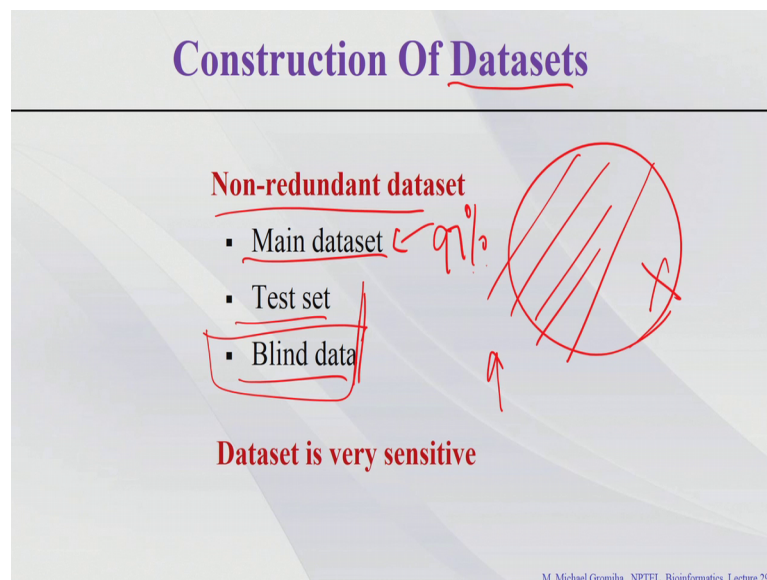
Now, solvent accessibility earlier we discussed two type of classification they buried or exposed for example, we have ASA value of 12 angstrom square. So, you take cut off of 5 angstrom square then this is predicted as exposed predicted or exposed we (Refer Time: 19:04) different cutoff we put 20 angstrom is cutoff that is predicted as buried.

So, now we can exactly predict the values this value is 12.9, 12 angstrom square you can predict this stability of particular residue is 12 angstrom square affinity angstrom square exactly can predict real values. So, now, we discussed about the different types of

classification problems some of them I discussed, there are many many many problems available in the literature with the applications to bioinformatics.

So, now the question is for example, you have identified your problem you want to classify for example, driver mutation passenger mutation and the stabilizing mutants and destabilizing mutants, what you have to do first. This is very important to construct dataset.

(Refer Slide Time: 19:50)



Dataset is very important because we need the experimental data for developing any model if you use any predicted data that is also predicted. So, then you will give the introduce more bias in this case your sensitivity or your reliability will decrease right.

So, you need the reliable data. So, in this case we discussed about various databases, what are the various databases we discussed earlier.

Student: Protein sequence database.

Protein sequence for example, UniProt (Refer Time: 20:20).

Student: Proteins structure.

Structures PDB.

Student: DNA sequence.

For example, a DNA sequences what are the databases for DNA sequences.

Student: EDBJ.

EDBJ, EMBL, Genbank and proteins stability: protherm database, enzymes: brenda database, several databases are available in the literature for the bioinformatics, reliable databases widely used databases fine.

First if you want to develop any method it is very important to construct a reliable dataset makes all your results depend on your data set because data set is very sensitive as much diversity you can add and you have to get the as much reliable data as possible in the dataset.

So, you can for example, if you are interested in the stability you can go to protherm database and you can get the data when you get the data. The full set of data you get data then you need to reduce the redundancy. The redundancy reduction is very important, otherwise what will happen if you do not reduce redundancy, what will happen.

Student: (Refer Time: 24:26).

We introduced bias. So, in this case model is biased with that particular data because if you only have more data in the same size right so that will be biased with that particular data to improve the performance. Finally, if you have new data then the method will fail in this case it is very important to you have non redundant data set how do you driven non redundant datasets?

Student: By sequence.

Yeah for example, if you have sequences then you can use the CD-HIT, pisces or blastclust to reduce the redundancy with various cut-off values. If we have the mutations you can get the unique mutations for example, A to A, R15 is mutated to valine, five data are available and you can consider the unique ones for setting the analysis otherwise if there is this plus your result will bias through this plus. So, we can make the redundant datasets. So, if we have this is the main dataset and you can get the non redundant this is the non redundant dataset. So, this is the redundancy you remove right. So, we have the non-redundant dataset.

Now, the next question is when you develop a model if you use all the data then the models will fit with respect to all your data. In this case we do not know the reliability for the new data in this case you classify this into at least two three groups. So, you can take it as a main data sets this you can use it for deriving a features or trying to assess the performance of your method within your data sets whether a method can fit the data that what you have the method is not able to fit then you need to find features which can fit your data right. So, then you keep new data sets for example, to test are completely blind and now you are not taking any information from this data.

Sometimes you take some of the information to derive the features and you can use a leave somewhere information they used as a test set and some cases you are not at all using any of the data to in your data set or doing any features. For example, if you calculate the propensity you can get the propensity with the 100 proteins, then we can use some proteins to train and some proteins you can do test and some cases completely new proteins you do not take any information from that that also we can keep.

In this case in case there are no information you consider from that that you can say as a blind data completely new data. Likewise if you have the different data sets one is your main dataset that you can train and validate the data and you have a test set and also you can use the blind data that is completely new. For example, if you take a new data how do you take this in the, if you take any classification problems you can leave the new data, but take the old data for the new dataset because even if you get some features for example, if we take the hydrophobicity values you take from the literature 20 residues.

But we do not know which proteins they consider to derive the values if you use this test set even your data set is new if you use the hydrophobicity as a features then the values that are included in this test, but we are completely new data so; that means, this data is totally new no information is taken from that particular data right. So, this is how we take it as a blind data set fine. So, why do we need three different types of data sets we check the reliability because otherwise if we check with this data set you can get the accuracy of 97 percent for example, we would be very happy.

So, ninety seven is good number right. So, then you then you see thing that this method is working fine because we do not consider that there is some overfitting in the data. So, here your method is try to accommodate all the your data with some specific features.

So, if you do not know whether this can also work with this a new data that is all the one need to predict otherwise if you do with the same datasets it is a kind of predicting yesterday's weather. You have to predict tomorrow's weather. So, in this case we need to completely blind. So, we do not know with the preset conditions and the other environmental factors we will think tomorrow may rain right.

Likewise we have to get the information from this main dataset and with the available conditions and known information we can predict the blind data this is why we need more dataset. This is very sensitive because dataset is very important. So, even if you take time to derive data sets. So, you have to be careful to make good dataset. So, this is a reason if you make good dataset you will get lot of citations because people do not want to do that because it is very time consuming compared with the other stuffs in developing algorithm.

So, usually they are want to do the developed data sets they try to find if data sets are available readily available right. So, they get the readily available data sets and they try in and they use their own methods on the data sets right. So, if you make your own data sets reliable data sets then they used to be a many many researchers in the literature. So, no one we have the data sets. So, you have the data sets for the training or whatever the blind data sets. Then what can we do next.

So, we will try to understand and see whether any intrinsic properties within your data sets can we identify any important features which can fit your data. For example, if you have the two proteins one is a globular proteins one is a membrane proteins can you find any information, for example, with respect to the ammonia acids any specific residues are biased or over represented in any of this proteins or under represented in this proteins.

Then if you see new ones we can say this residue is over represented here for your test set also it is over represented. So, this could be used for identifying the type of protein. So, that you find the different features, various features we discussed whatever are various sequence base features.

Student: Amino acid composition

amino acid occurrence, composition.

Student: Average and amino acid properties.

Average properties, molecular weights.

Student: Sir.

Right, profiles, conservation scores. So, we can derive various properties using these sequences likewise if we have structures there also you can do various features for example.

Student: Solvent (Refer Time: 28:02).

Accessibility.

Student: Contact potentials.

Potentials right. So, we can see there are various surrounding hydrophobicity we can do we do various features using this structures. Contact order, long range and so on right. So, now, if you have a sequence or if you have a structure we can think of the reliable features which can address your problem for example, if you are interested in binding affinity of folding proteins complexes this is high affinity or low affinity.

So, let the features which are relevant to the affinity for example, if you take the affinity we discussed protein, protein complexes so which interactions are important? Electrostatic and the cation-pi and aromatic-aromatic interactions. So, in this case you can look for this the residues which are responsible for this type of interactions and the second one instead of looking for the whole sequence if you are able to identify binding sites whether you can see any dominance of these residues at the binding interface right.

So, if you do it properly with the features which are relevant to the particular problem then you can perform better in your prediction accuracies. So, eventually some of these features like the physiochemical properties or the specific interactions for example, what are the physiochemical properties for example.

(Refer Slide Time: 29:25)

Feature Selection

Evaluate different features

- Physical/chemical Properties → M.W., shape, Charge
- Specific interactions →
- Amino acid composition
- Residue pair preference
- Motifs

Selecting important features plays a key role

GXXXG motif
DRY
G-Pol-Hy-Pol-...

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 29

Student: Molecular weights.

Molecular weight because the shape.

Student: Charge.

Charge, hydrophobicity and so on. What are the interactions? Electrostatic interactions, van der Waals interactions, hydrophobic interactions so on. So, you can see the composition that we discussed earlier pair preference I can see any specific motifs for example, there is a motif G-X-X-X-G motif, for membrane proteins. For GPCR we have this DRY motif. Also we discussed about the beta-barrel membrane proteins that is glycine, then polar and hydrophobic, polar hydrophobic this motif. The various motifs are present for different types of proteins with specific functions or specific structures. So, we can derive these features or relevant features with your sequence.