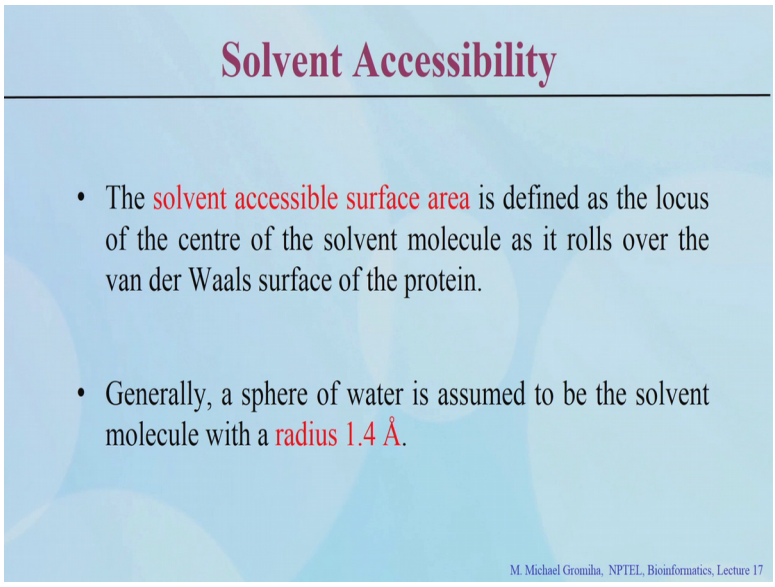
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 17b
Protein Structure Analysis II

Then another aspect this is very important aspect in protein structures, is solvent accessibility.

(Refer Slide Time: 00:19)



That name itself tells this is divided into two solvent and accessible what is the accessible?

Student: Reachable.

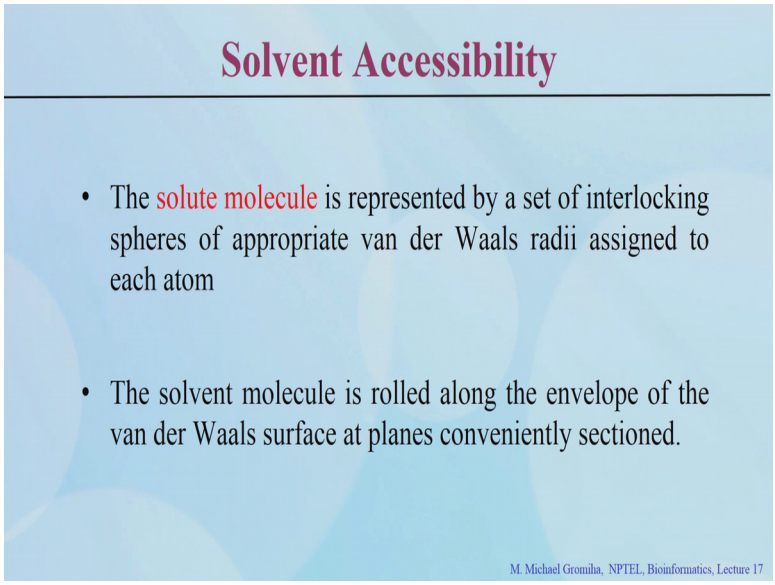
Reachable like how far we can reach whether it is very close. So, you can reach more or how far we can in contact right. So, what is the solvent; with the respect to a solvent we have a protein structures if there is a solvent how far this solvent can reach each atoms are each residues in a particular protein right. So, based on that you can define the solvent accessible surface area, because the area we tell because how much area they can overlap, this we call the accessible surface area.

So, you can define this as the locus of the centre of the solvent molecule, as it rolls over the van der Waals surface of the protein. So, we have the protein different residues and they take a different atoms, we can represent these atoms in the over lapping spheres and if you have a sphere, then we take a solvent molecule and you could roll the solvent molecule on this protein and see how far they are in contact. There contact is called the contact area and how far the center of the solvent molecule rolls on the van der Waals surface then that is called the accessible surface area. Usually we take a solvent, water is in which environment.

Student: Water.

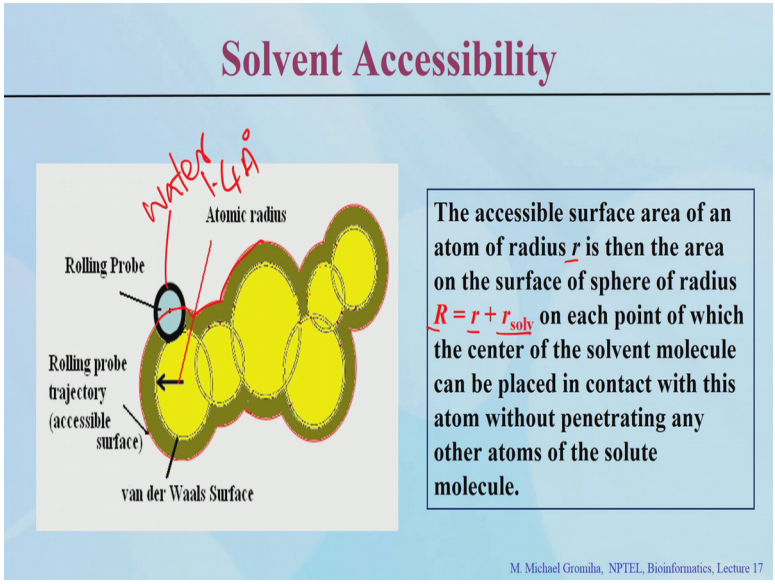
Water environment the protein used water environment. So, if it consider you proteins environment, then we take a solvent as water right. So, you can take the water and you can see the radius of 1.4 Å because of the environment of any proteins.

(Refer Slide Time: 01:56)



So, then what to do? So, you can see the solute molecule; solute molecule what is solute molecule here? Is a protein right. So, you can see a protein this is a solute molecule, you can see the interlocking spheres. So, you can see this is the interlocking spheres here.

(Refer Slide Time: 02:12)



And we can assign with the different van der Waals radii; what are atoms present in water in protein; carbon?

Student: Oxygen nitrogen.

Nitrogen.

Student: Oxygen sulfur.

Oxygen sulfur right.

Student: sulphur

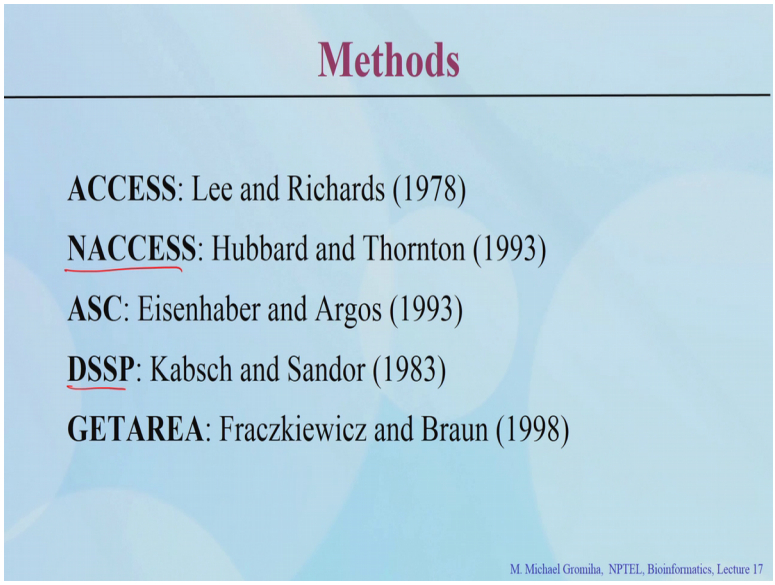
So, based on the atoms we have the van der Waals radii. So, each item you can have the different van der Waals radius. So, they are interlocking spheres now what to do you can now roll the solvent molecule, on the along the envelop of this van der Waals surface you can see the van der Waals surface here right. So, this is the water molecule this is the water, of 1.4 Å. So, you roll water molecule on the van der Waals surface. From that now you can see the radius of the surface is equal to R plus r the solvent molecule, and you can calculate accessible surface area of atom of radius R you can see these r or this is a area of the surface of a sphere of radius r which is r equal to r plus r the solvent molecule, on each point with the center of the molecule are in contact these solute

molecule in contact with there is solute molecule without penetrating the other atoms right.

So, if we have this water molecule which are rolling around, how far they can penetrate into each atoms depend based on the van der Waals radius of this each atom in a protein. So, that area we call as solvent accessible surface area right. So, we have the water molecule now you roll on this molecule. So, this is the contact, how much you can contact this is the contact area and how much you can penetrate into the van der Waals surface that you can get the accessible surface area, that can be that some accessible surface area you can accessible by the solvent with the atoms.

So, there are several methods; this is the principle how they estimate the solvent accessibility, there are various methods to calculate the solvent accessible surface area of each atoms to based on the different algorithms.

(Refer Slide Time: 04:08)



So, I show some of the algorithms one is the ACCESS this is the one first introduced the literature in 1978, actually started 1971 Lee and Richards they proposed the concept of solvent accessibility, and the program they developed in 1978.

So, they use this information how the protein can be accessible by a solvent molecule. Later on they refined in 1993 and they developed NACCESS to get the accessible surface area of all atoms. Likewise there are several other methods like ASC or DSSP or

GETAREA, ealier we discussed the DSSP for the second structural proteins they combine the second structure as well as solvent accessibility, at the one file you can get all the data.

(Refer Slide Time: 04:58)

Comparison index	ACCESS	DSSP	NACCESS	ASC	GETAREA
Standalone executable availability	Yes	Yes	Licensed	Yes	No
Online calculations/ database	No	Yes	No	Yes	Yes
Polar and nonpolar area	No	No	Yes	No	Yes
Atom-wise surface area	Yes	No	Yes	Yes	Yes
Source code availability	No	Yes	No	Yes	No
Choice of probe radius	Yes	No	Yes	Yes	No
Choice of van der Waals and other parameters	Yes	No	Yes	Yes	By Manual editing
Secondary structure	No	Yes	No	No	No
Reference	and Richards 1978	Kabsch and Sander, 1983	Hubbard and Thornton, 1993	Eisenhaber and Argos, 1993	Fraczkiewicz and Braun, 1998

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 17

So, these are the some of the aspects of different methods, why do we need different methods? There are various aspects. So, there are several algorithms available to calculate the accessible surface area of all atoms and residues in a protein. So, the major programs where ACCESS is the one which are develop early in the literature may be 1970s and followed by DSSP NACCESS ASC GETAREA and so on, currently we have several other algorithms are also available in the literature.

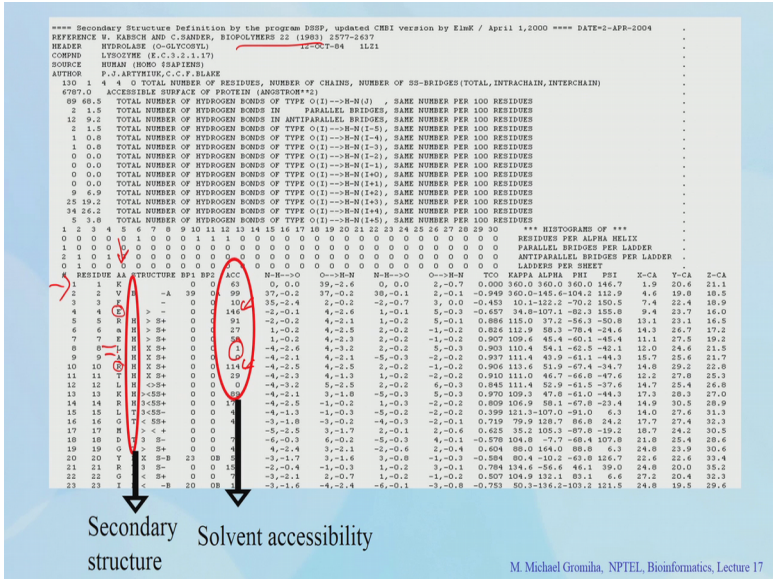
So, if you look into these different algorithms there are some advantages and some disadvantages. For example, if you look into DSSP. So, you can easily use standalone program is available, but it can give you the accessible surface area for only the residues. So, we cannot get the data for the atoms say if you take the atom wise. So, it is no, but if you look at this NACCESS you can get the atom wise one right, but here there is no online calculations or database, but NACCESS you need to get the license you have to write the copyright agreement and then you have to send, and get the license and then you can use it.

And here this will give you only the secondary structure though information regard only the accessibility, but no secondary structure. Look the DSSP it will give you the ASA

plus secondary structure. So, there is advantage of using DSSP, you can get the secondary structure and solvent accessibility simultaneously when you run the program. So, ASC standard program, in this case you need to edit or manipulate this files depending upon your input files right. So, you can edit, here you can see the you can change the probe radius, you can change the van der Waals radius and source code is also available. So, this for free they will give you the source code.

So, in the GETAREA. So, standalone is not available, but we can get the online. So, if you give the PDB then you can get all the data. It give the polar nonpolar area I can give atom wise surface area right, but you can also use by manual editing you can change the van der Waals radius and other parameters. So, some program you cannot change anything some of case only you can get this standalone, some case only you can do the online some cases you can manipulate the change your default parameters and DSSP provides both secondary structure as well as the solvent accessibility, this are the references this is Lee and Richards right.

(Refer Slide Time: 07:41)



So, this is the output obtained from the DSSP, example I show the output from the DSSP. DSSP is developed by Kabsch and Sander they published in 1983 *Biopolymers* right. So, here if you see the data here the data starts from here, this is the residue number the two numbers one is based on the PDB file one is actual residue number from the UniProt. So, you have the two numbers. So, here you get amino acid sequence KVFE and so on, and

here this is the assignment of secondary structure and this case ACC this is the accessible surface area. When you look into the accessible surface area can you see these numbers and tell something from the accessible surface area, that which residues are buried which residues are exposed?

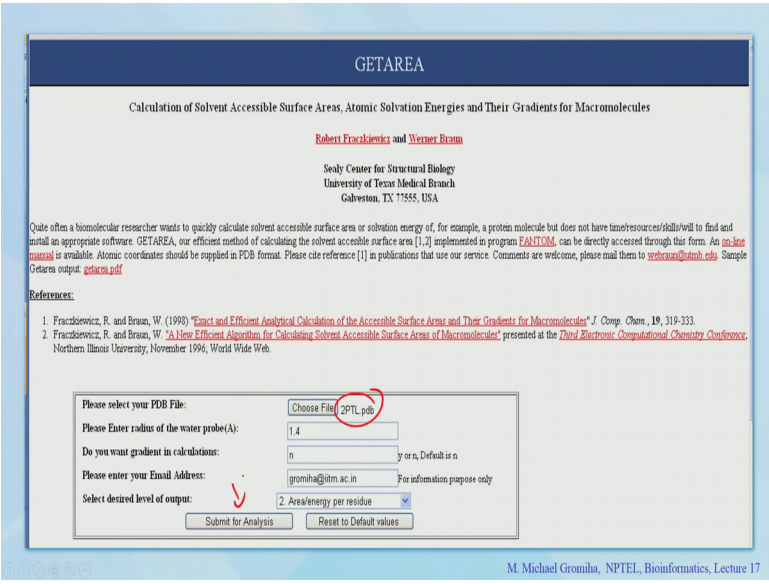
Student: Alanine

If you see the numbers some of them are very high for example, here this is 146 and so, 114 these residues are exposed they are at the surface. So, for example, these residues you see this arginine, and you can see these the glutamine acid these residues are at the exposed then some residues for example, they are 0 and 1; what is the meaning of the 0 and 1.

Student: They are buried.

They are buried right. So, in this case this is leucine alanine. So, these residues are buried in the interior fine. So, if you have any PDB code, you can download DSSP you can standalone program is available, you can execute easily just one line command if you give the DSSP and you give the your PDB ID, and if you give you the output file name you will get it. It just a online command you can get the data for the DSSP right, but only disadvantage is it will give only the residue wise accessible surface area, otherwise it is it is very fast.

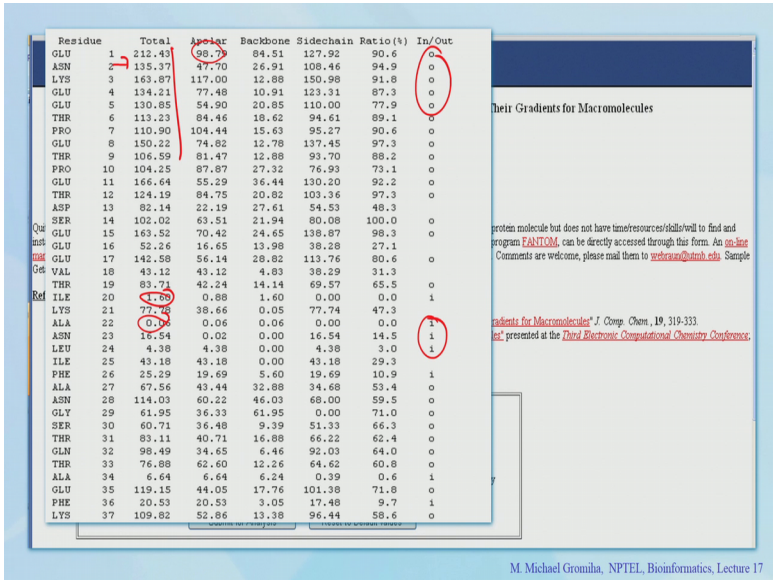
(Refer Slide Time: 09:28)



This is another program is called the GETAREA, here you can use the online version where standalone not available, but you can manipulate the probe radius and other parameters and you can get the data for all atoms. For example, if you want to use the PDB file you can give the PDB here I give 2PTL the PDB right.

So, water molecule there is 1.4, here if you give the email address and if get the output, and submit for analysis, then you will get the data.

(Refer Slide Time: 09:59)

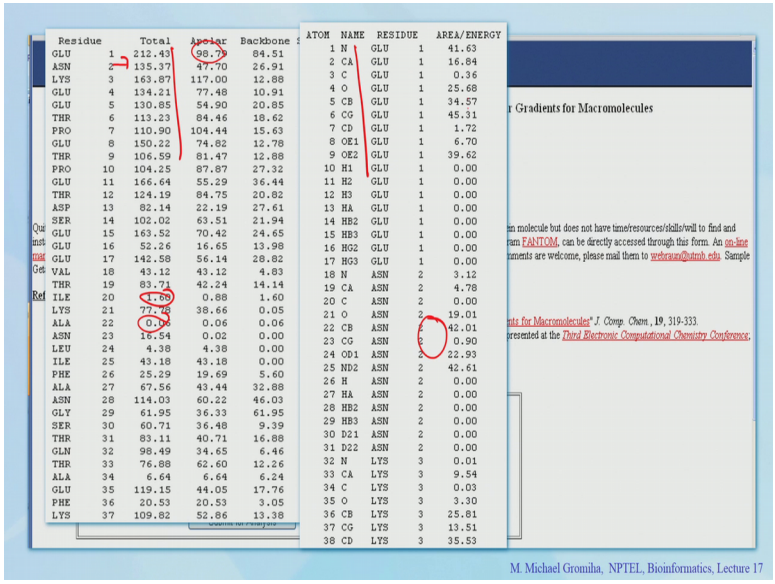


So, if we look into this data. So, each residue this is a total surface area and they group into different atoms based on the apolar and the polar. So, among total 1 2 accessible surface area $Å^2$, apolar is about 98. Then also they classify into backbone and side chain, this is the backbone and this is a side chain as well as this ratio. Then based on this area the classify this is out or in, what it is a meaning of out?

Student: exposed.

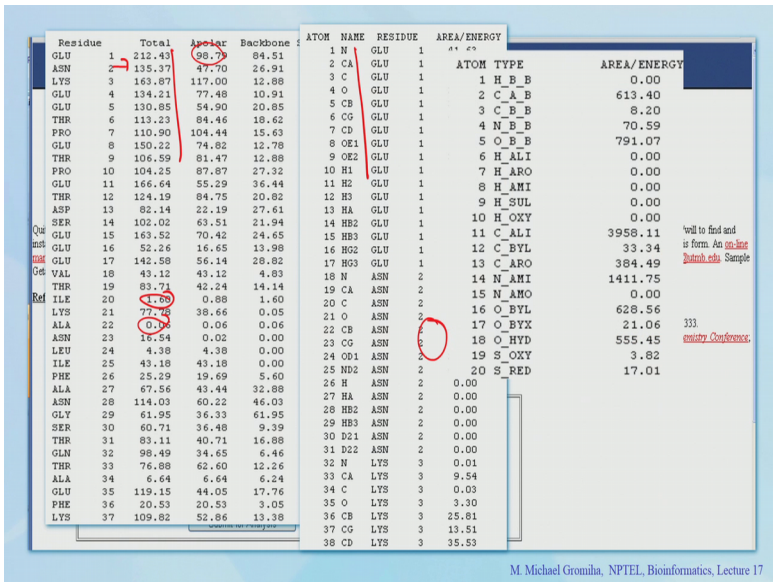
This is a exposed. So, values are very high. So, they are out. So, here this is less here this is less and they are the interior of the protein, they can classified into either interior the protein are at the surface fine.

(Refer Slide Time: 10:46)



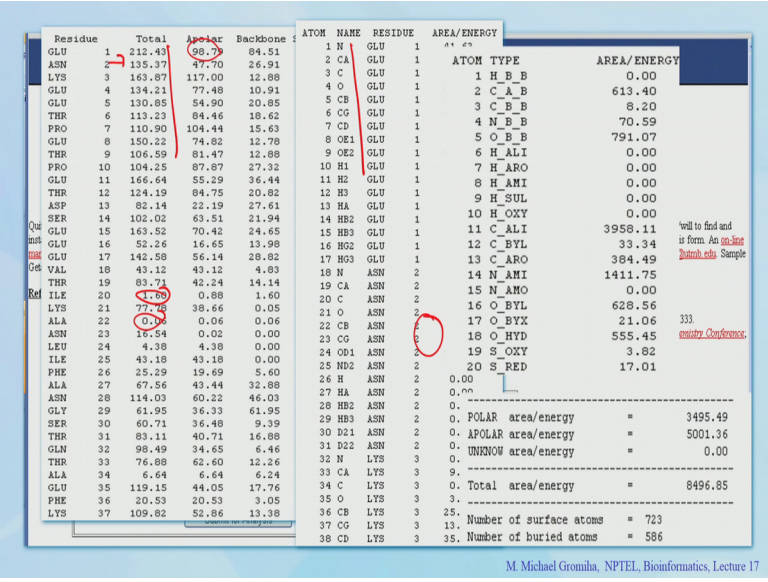
So, now we can also get the this area for each atoms, here this is for each residue, now here you can see even for these glutamic acid. So, they have for each atoms you can see the area for each atom in a residue.

(Refer Slide Time: 11:04)



So, this each atom type where this is CAB or CBB this is the backbone, and you can see the different atom types and you can see whether this sulfur and the aromatic aliphatic. So, you can see the area based on the different atom types.

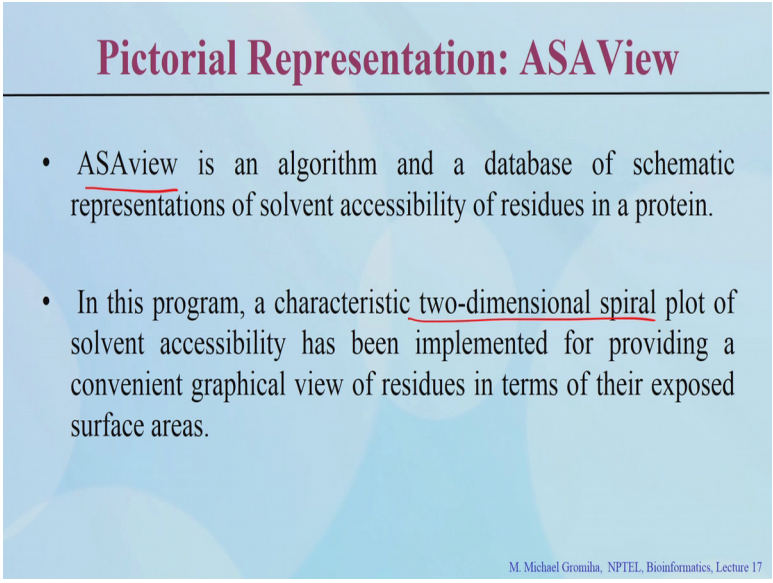
(Refer Slide Time: 11:21)



These are overall you can say overall polar area and the apolar area, and this is the total area also there are number of atoms which are the surface which are at the buried.

So, you get all the information because will give you the data for each atom is easy to use, but now they issue is most of the biologists, they find it difficult to interpret the data. So, in this case instead of giving numbers, if you provide a kind of figures so that, they can understand which residues which are at the surface and which residues are the buried then will be very helpful to interpret the results.

(Refer Slide Time: 11:57)



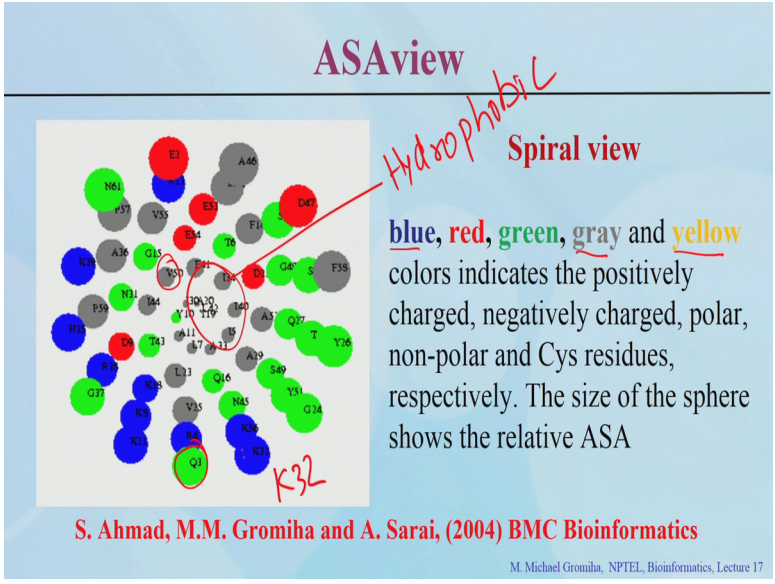
So, we develop data the database called ASAView, we developed database as well as algorithm to represents solvent accessible surface area. It is a characteristic two dimensional plot kind of spiral plot, which will give you some information so that, you can view the figures and then get the information right. So, how to do that? Here the various aspects one is different types of atoms, different types of residues right.

So, here we considered only the residues not different atoms right. So, what are different types of residues in a protein?

Student: Polar.

Polar residues, non polar residues, charge residues right. So, we can show specifically these are charge residues or the hydrophobic residues polar residues and so on. Then they these residues they have different surface area, some of them are less are some of them are more. You can give this circles with the different sizes based on the surface area. So, we want to give the plots in the form of the spirals as well as the bar chart just give you the actual values for each residue in a particular protein. You can use this online plots for any PDB ID or you can upload your PDB you can get the plots ok.

(Refer Slide Time: 13:16)



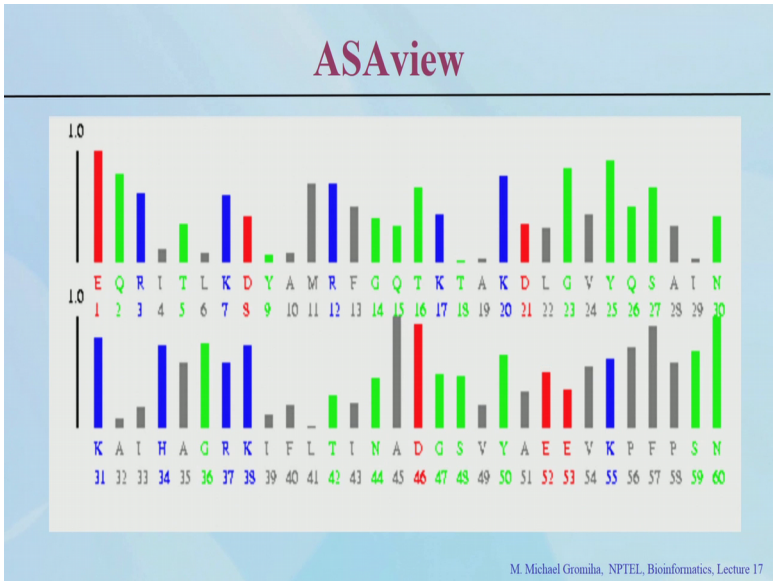
Now, I will show you the plot. So, here it is a plot for a particular protein. We got different colors and the different colors represent for the different types of amino acids likewise there is the blue right. So, you can see the blue this is the K 32, these are

positive charged residues and the green. So, here you can see the green this is mainly for the polar residues based on the notations, and you can see the gray, gray is mainly for the hydrophobic residues valine alanine and the yellow is mainly for these cysteine residues I do not find any cysteine here. And you can see the size, some of them are very small and some of them are very big they represent the relative accessible surface area of each residues. From this one can we guess which residues prefer to be the interior which residues are in the surface.

Student: Non polar.

Mainly we see the non polar residues see the stretch of these gray colors inside the core. So, most of the hydrophobic residues this is the hydrophobic residues, which are mainly at the interior the core in the surface if you see mainly the polar residues and you can see the charged residues positive charge and the negative charged residues. You can see the distribution of this residues at the interior of the protein or at the surface.

(Refer Slide Time: 14:41).

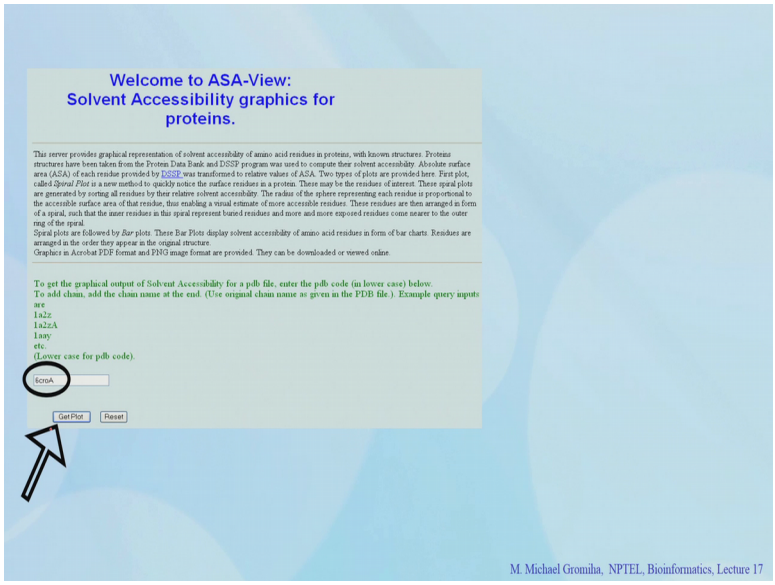


This is another plot is a bar plot which give you the value for each residue. For example, here these starting from the 1 to n, N terminal to the C terminal will you give the data. If we look into these plot here also you can see they continuously not with the high ASA or low ASA; like for example, if you see these residue and here you can see they are distributed differently in the sequence here these residues, but when you look into this number they are buriedly interior.

So, comparing this plot and this and this plot, some residues which are far away in the sequence they are come close to each other and there they placed at the interior of the protein. This also supports this hydrophobic model right. So, that they collapse and then they form the core like hydrophobic residues, and the other polar residues which are at the surrounded by these hydrophobic residues, to make the interactions with the water as well as with the other residues.

So, here in there if you see this hydrophobic residues but they have very less ASA there mainly in the interior of the protein. So, we have the values you can plot, either you can make the parallel plot or you can make the bar chart to see how these residues are distributed based on accessible surface area.

(Refer Slide Time: 15:52)

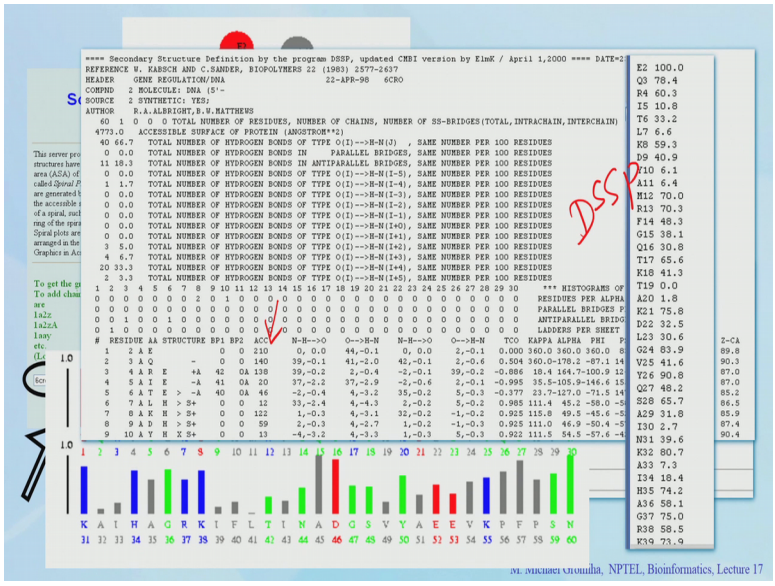


So, we have the web server, we can it can take any PDB ID, this you can get the PDB ID. So, get plot right then you can see that what are types of plots you require.

(Refer Slide Time: 16:01)

So, depending upon which one you want the pdf version or the any relative values whatever you need. So, you can see the graph this parallel plot and you can see the bar chart. In addition we use the actual values if you see this one you get this is actual surface area, this is the output from the DSSP we give the link. So, that they can the users can use and they cannot see the exact values for each residues in their particular protein of interest fine.

(Refer Slide Time: 16:35)



Here this is the value for each residues this only the ASA values.

So, now if you look into this different residues for example, if you compare glycine and the lysine, how many atoms in glycine?

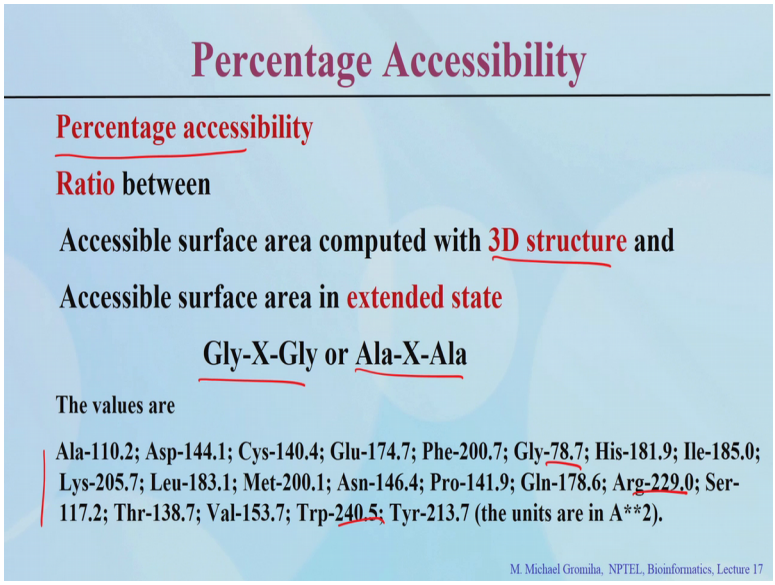
Student: 4

Four there is; only if we are the heavy atoms only 4 only the mainchain atom. If you if you take alanine it is 5.

Student: 5.

If we take arginine or tryptophan try to a more number of atoms. So, when you do they residue wise accessible surface area they add up, they do it for all the atoms and finally, add up then you give for the residue. In this case there is a bias of this residues some of them are small some of them are big in this case we need to normalize, how far your your residue is accessible right.

(Refer Slide Time: 17:24)



In this case you can calculate the percentage accessibility that is the ratio between the accessible surface area, which are computed 3D structures as well extended state what is they meaning of extended state?

Student: Fully accessible.

The maximum it can be accessible, in this case they are two different ways one is you can take this tripeptide, and you can use this concept to get the accessible surface area.

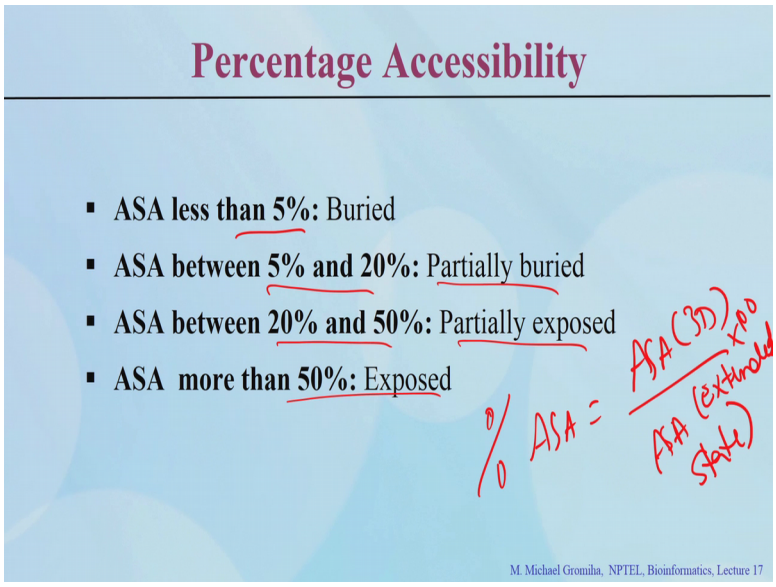
And the most easiest one is they take the conformation glycine-x-glycine or alanine-x-alanine, why they use glycine-x-glycine?

Student: Glycine has no side chain.

Yeah glycine has no side chain. So, in this if it is x it is central residues any residue its highly accessible, because neighboring two residues they do not have the any side chain. You look at the glycine-x-glycine in all the proteins and see the conformation and from that they can take either the highest value or they can take the maximum value with the average value. From that they will see this into the probable accessible surface area for any residue whether extended state. These are the values for the 20 residues and as you see glycine has the lowest one or you can see the tryptophan or the arginine, you can see the highest values.

Now, you can get the ratio, one from the 3D structures you run the program and any of the programs; what are the programs we discussed? NACCESS, ACCESS, ASC GETAREA or DSSP you get it for the folded state and XYZ values we know then you divide that right then will get the percentage accessibility the percentage ASA.

(Refer Slide Time: 18:59)



So, that is ASA in the 3D structure and ASA in the extended state, from this you can calculate the percentage multiply 100.

So, then the value is less than 5 percent, then we call these residues are buried; what is the meaning of buried? They are in the interior core for example, if you see this figure this is the buried and you can put the value of 5 to 20 percent this partially buried, and 20 to 50 as partially exposed and more than 50 as exposed. And this is the general values we use the literature, but this not very strict here also you can change your cutoff based on the type of application you use. Generally if it is less than 5 percent, we use a buried some cases they use 2 percent or that is changeable. Then 5 to 20 as partially buried and 20 to 50 is partially exposed and 50 percent is exposed. So, this how we can get the accessible surface area converted into percentage and we can do it for analysis depending upon the location of each residue.

So, we summarize what are the various aspect we discussed today?

Student: Structure classes.

Yeah different structure classes, what are different structure classes?

Student: SCOP SCOP and CATH.

Right all alpha, all beta, alpha plus beta, alpha with beta what are different databases which can give this information?

Student: CATH and SCOP.

CATH and SCOP, then we constructed contact maps what is the contact map? This a plot connecting the contact between two residues in protein structures; in construct any space you can define a distance and different different atoms and if you have the contact you put a dot. So, this will give you the 2D representation of the contacts between residues in protein structure right. So, from that kind can you can defined mean long range contact, or short range contact or medium is the contact, depending upon the residues with the close in space, but how far they are distant in the sequence.

So, then we discussed about solvent accessibility, how far each residue is accessible to solvent. The different programs to get the solvent accessibility and we can represent in a pictorial view for example, which is the program to get the pictorial view?

Student: ASAView.

ASAView, and to get the percentage accessibility, so you get the surface area at the buried the 3D structures and you can do it from the extended state and get the ratio, then you will get the percentage ASA, based on that we can classify into buried or partially buried or partially exposed or exposed right. So, you can also derive various other parameters say conduct order or long range order, and the buriedness preference have residues to be with the interface, and the free energy, and different types of interactions and the electrostatic or the residues which can form disulphide bonds, disulphide bridges right. So, there various features write the properties you can derive from the protein 3D structures and we will look into details in the following classes.

Thank you for your kind attention.