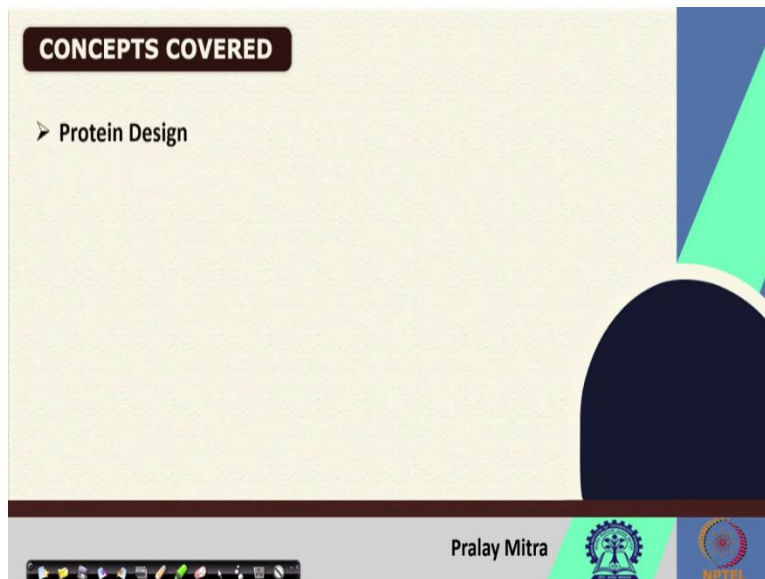


Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture: 42
Computational Protein Design (CPD) (Contd.)

Welcome back. So, we are continuing our discussion on computational protein design on the last lecture, I demonstrated the need for the existence of a computational framework for protein design otherwise theoretical possibilities are enormous it is astronomical in nature. So, for 20 different amino acids and if the length of the protein sequence is n then 20 to the power n number of theoretical possibilities are there. So, that is the bad news, but the good news is also that not all possibilities exist. So, if we apply our knowledge of biology, then we will see that only few possibilities are in existence. So, we will discuss those in this lecture.

(Refer Slide Time: 01:00)



The image shows a screenshot of a presentation slide. At the top left, there is a dark brown box with the word 'KEYWORDS' in white. Below it, a right-pointing arrow is followed by the text 'Protein Design'. The main body of the slide is light beige. In the center, the title 'Plausible Heuristics' is written in blue. Below the title is a bulleted list of six items: 'Branch and Bound', 'Dead-End Elimination', 'Genetic Algorithms', 'Simulated Annealing', 'Monte Carlo / Replica Exchange Monte Carlo', and 'Tweaked DEE or MC'. At the bottom right of the slide, there is a circular inset video of a man with a mustache, wearing a white shirt, speaking. The slide has a decorative border on the right side with blue and green geometric shapes. At the bottom, there is a grey bar containing the name 'Pralay Mitra', a logo of a tree inside a gear, and the NPTEL logo.

So, the concept we will call the protein design keyword is also protein design. So, while designing the core of the protein design problem, the core of the protein design algorithm or the core algorithm for protein design. So, people have tried n number of different approaches starting from Branch and Bound, Dead-End Elimination or insert DEE, Genetic Algorithms, Simulated Annealing, Monte Carlo or Replica Exchange Monte Carlo and other better variation of the Monte Carlo or Tweaked DEE or MC.

So, people have tried all as of now, we discussed in the context of algorithms for protein modeling and engineering the Genetic Algorithms we discussed we discussed Monte Carlo we discussed Replica Exchange Monte Carlo. So, for our purpose in protein design, we will exploit

Monte Carlo and then also we will extend that with the Replica Exchange Monte Carlo and we will also look for the opportunity of parallel implementation of the algorithm. So, that fur thus speed up can be enhanced such that for large protein sequences or if I wish to explore the solution space for several number of proteins, then it can be done very quickly or easily.

(Refer Slide Time: 02:34)

The slide is titled "Energy function" in blue text. It contains two main bullet points. The first is "Physiochemical parameters" (note the typo "Physio" instead of "Physico") with a sub-list: "van der ^{Waal} interaction" (note the red handwritten correction from "waal" to "Waal"), "electrostatic interaction", "hydrogen bonding", "solvation environment", and "..... many more". The second bullet point is "Knowledge based energy functions". To the right of the list, the text "Physics-based" is written. In the bottom right corner, there is a circular video inset showing a man in a white shirt. At the bottom of the slide, there is a navigation bar with icons and the name "Pralay Mitra" next to two logos.

On the other hand, from energy function point of view also. So, several variations exist. So, first of all people mostly rely on the physicochemical parameters like van der waals interaction. So, I am sorry there is a small typo, this is waal. So, van der waal interaction electrostatic interaction, hydrogen bonding, solvation environment and many more and usually, this kind of parameters.

When combined in order to generate one scoring function that is called Physics-based scoring function apart from that one knowledge based or say evolutionary information best scoring functions are also in place. So, we will see which one is better say physical chemical parameter and physics based or knowledge based like evolution information or a combination of both of them, we will discuss also that.

(Refer Slide Time: 03:42)

The image displays two slides from a presentation titled "Guide the search through profile".

Top Slide: Shows an "Input Protein Structure" (a red and green ribbon model) being compared to "Homologous Structures" (three similar ribbon models). Handwritten red notes include "RMSD" and "TM Score" with a checkmark, and "70.7" and "70.99". A red circle highlights a specific region of the input structure. A red arrow points to a sequence alignment at the bottom. A small video inset shows the presenter, Pralay Mitra.

Bottom Slide: Shows the same "Input Protein Structure" and "Homologous Structures". Handwritten blue notes include "Conserved regions" with a blue circle around a region in the input structure. A blue arrow points to a sequence alignment at the bottom. A small video inset shows the presenter, Pralay Mitra.

Now, let us start with the algorithm construction for this first of all, let us have an observation from the biology. So, when one structure is given to you, that is your input protein structure, then, look, what you are trying to do given one protein structure and also sequences are known to you, you wish to see that at which positions if you perform some mutation or changes in the amino acid.

Then that sequence or that new sequence will also lead to that given structure in order to do that one, so, first biological information that you can exploit for the given protein structure, why not loop for all those structures, which are similarly looking with this particular given input

structure. Let us call that as a homologous structure. Borrowing the key word of homology from the sequence in this case, we are using that key word homologue. In the context of the structure, we will declare two protein structures are homologous in nature. If say, either RMSD or TM score these two measures you remember we have discussed in the context of aligning two protein structures and the knowing whether those two structures are similar or not.

So, in that context we discussed RMSD and TM score. Now, the same definition now we are going to utilize that based upon that RMSD or TM score which means that given one protein input structure, then what we will do, we will look for the existence of other protein structures which are structurally similar to the given or to the input protein structure.

If they are structurally similar to the given our input protein structure then those will be called as the homologous structures. Now, if I ask you that. So, give a input protein structure and I am also interested to identify other similar protein structures which are homologous to this where to look at or what to search. The simple answer is protein databank because protein databank houses a number of protein structures.

So, why not with the given protein structure, you go for aligning the structures which are existing in protein databank and provide a threshold value either on the RMSD or on the TM score so, that if do structures after the alignment is below some threshold RMSD or TM score, then we will declare that as a homologous protein structure.

Now, the implementation could be very simple and easy if you also consider the length of the protein sequence because in protein structure as I mentioned several protein structures are there say more than 100,000 now, if you are comparing then the comparison even if you do computationally will take some time.

So, in order to reduce that one, one suggestion could be that you first look for the length of the protein sequence. Now, that length of the protein sequence you can grab from the PDB file itself. So, that information will be stored or just you do one scanning of the PDB file and by looking at the number of C-alpha atom corresponding to each chain.

So, definitely for each PDB ID you have to parts that protein structure file in for the PDB and you have to identify each subunits or each chain separately once you will identify that one then

you look for what is the length now, what is the length of your protein structure with some deviations or with some say plus minus addition and subtraction of the length then you can consider that structure or it is reject. So, for example, if your protein is a plane, say 200 amino acids, then you can readily eliminate all those protein sequences whose length is say less than say 50 or say greater than or say 300 or 400 something like that.

Now, after quickly doing that one you can align those two structures again using either a rotation about arbitrary axis or using TM align. After aligning this you compute either RMSD or TM score then you decide on some threshold value that threshold value I will come shortly which could be a very good threshold value use that says threshold value declare whether the structure are same or not I mean homologous or not.

After identifying, this homologous structures, what he will do I got this structure, this structure, this structure then I will go for multiple sequence alignment of those structures. In short MSA Multiple Sequence Alignment of those structures. Here is one sematic given to you once he will do that one.

So, since structurally they are same then it is very much possible that at each position there are identical residues and also there may be some gaps which will be introduced this hyphen here here indicates that there are gaps otherwise if you look at then you will see that in some places for first position I can see seeding and methylin both for other positions each column colored with some color value indicating that they are with the same amino acids.

Now, I think it will be easy for you also to understand given a protein structure if I wish to given a protein structure if I wish to design a new sequence which will fold to that structure then what I did I identified all the protein structures which are homologous structures with the input sequence then from there after doing the multiple sequence alignment I got some aligned sequences now, at each position I look at the variations in my amino acids.

Now, from that variations I can compute. So, first of all after looking that variation, if I see that only one particular amino acid is occurring always which means that other 19 amino acids will not be probably mutated at that position or probably substituted or replaced at that position. If that is true, then what you can do that so, that is not allowed to be mutated or placed in that particular position.

So, you rule out those possibilities. So, if in one position you see it is only the valine then probably others will not occur there that way you can guide your search. But be assured that since these structures are homologous in nature and based upon the threshold you will be applying on the RMSD or TM score if you say give very strict threshold that only say aligned.

So, in terms of the TM score which is a normalized value varies from 0 through 1.0 if you say that all those structures which are greater than point 99 will only be used as the homologous structure and the for the alignment then perhaps you will see that all the sequences are identical and there is no scope for any variations during the for the mutation.

But if it is not like that if you consider say greater than, say 0.7 or so then definitely will get some variations at some positions. Also, you should remember that theoretically although all the amino acids are equally probable or at each positions, but it is not the fact that biologically they are going to be stable if in all positions I will change the amino acid or muted that amino acid it is not true in some positions only mutations or changes are allowed.

But not for other positions, some other positions may be very crucial for the fold or structure of the protein. And if I mutate at that position, then I may loose the required structure as well as a function of that protein. So, during that mutation, I have to be careful about that one and to some extent, that support I will get if I consider the homologous structures.

So, homologous structures will tell me which reasons are not going to be changed, that part I can call as the conserved which are not going to be changed during the mutation. So, from here you can see that this these three beta sheets 123, 123, 123, 123 these three beta sheets are going to be conserved, but this yellow, so, I am using blue color.

So, this region you see, so this is the for this you will see may see some variations. You may see some variations, for this loop region also you can see some variations. So, some for some reasons, you may see some variations for other you may not and you have to follow that one if you follow that one then you will understand that it is not going to be 20 to the power N but very less number compared to that 20 to the power N. So, that is about the guidance.

(Refer Slide Time: 15:16)

PSSM

$M(p, a) = \sum_{x=1}^{20} w(p, x) \cdot B(x, a)$

The Algorithm

Algorithm 14:

Input: A protein structure

Output: Novel protein sequences that will fold to that given protein structure.

Steps:

1. For the given protein structure identify the homologous experimental structures.
2. Compute the PSSM from the homologous structure.
3. Perform Monte Carlo (MC) simulation starting from a random sequence.
4. Analyze the accepted new sequences generated during MC simulation to output best candidate solution.

Ab initio protein Design

Metropolis's criterion

Once you will have the multiple sequence alignment then for each position you can compute one score value that score value we can consider as the probability. So, if you have say N number of amino acids and here you have 20 different amino acids. So, this N number of amino acid by this I used to say then what do you can do you can compute one N cross 20 matrix in that N cross 20 matrix at each position.

So, if I consider N cross 20 matrix. So, here say let us assume position 12345 up to say N and as for the column ACDE these are the amino acids up to Y what I can do that at first position what is the probability of having Elenin that I can compute and stay keep here what is the probability

of computing cysteine I will compute and keep here that way I can feel this matrix. The same thing is done here. Now, how can I compute the probability very simple you have the multiple sequence alignment now, we discussed about the henikoff weight not the weight of the sequence, but at each position what is the contribution you remember that equations 1 divided by r multiplied with s where r and s indicates that at each position how many different amino acid exists and what are the multiple copies of one particular amino acids exists.

So, based upon that one I am computing that and if I compute then taking some for one position is going to be one and here also I need that some is going to be the one and I am getting that one. So, that probability I can compute using the same concept of that henikoff weight and I can put it here.

Now, after putting that here, what you can also think that will it be only the probability value or I will include something more true something more means, I am interested to draw your attention to this part. So, here $B \times a$ is one weight factor you can consider that weight factor is you can consider as a Blosum 62 matrix.

So, why am in using 62 because at most 62 percent similarity exists in the protein sequences from which I computed this one and it is also demonstrated that Blosums 62 performance as a substitution matrix is much more better compared to say Blosum 80 or Blosum 40 or Blosum matrix computed at different redundancy level.

So, this part I can multiply or I can wait so, that will be an added information regarding the evolutionary information of one amino acid. So, when I am trying to muted, then what is the probability. So, that is the additional stuff you may keep to detail, but mainly what we have to do that instead of beating the bush or going for 20 to the power N possibilities with equal probability.

I wish to bias by probability by the occurrences of amino acid at one particular position after doing the multiple sequence alignment of the homologous structures corresponding to the input structure. So, that is the one guidance I am suggesting to you now after that guidance my algorithm for protein design is going to be very simple at the core it is using one Monte Carlo simulation technique replica exchange will come later Monte Carlo simulation technique.

So, input is a protein structure output novel protein sequences that will fold to the given protein structure. So, grossly there are four steps for the given protein structure identify the homologous experimental structures compute the PSSM from the homologous structure perform Monte Carlo simulation starting from random sequence analyze the accepted new sequences generated during Monte Carlo simulation to output best candidate solution.

So, to go one after another for the given protein structures identify the homologous experimental structure what is the reason because we wish to compute the PSSM from homologous structures. So, these two steps we discussed extensively just now. So, in summary what I have to do given one input structure.

So, I will take that input structure use some library function which is developed by me or existing somewhere which will align this input protein structure with all the subunits or chains of the protein structures available in the protein databank after aligning that one I will compute either RMSD or TM score after that alignment between the simple structure and the PDB structure. And if the RMSD or TM score is above some threshold value that threshold value also.

I will mention I will mention keeping into consideration that this particular threshold value will give me similar structures that similar structure I am calling as homologous structures. So, if that it is above that threshold value then what I will do I will take that structure and keep it separate that way once my scanning of the protein databank is over and I will have all the protein structures with me.

Then I will align those selected protein structures which are homologous protein structures. And we will form one multiple sequence alignment from that multiple sequence alignment we will compute position specific scoring matrix either the way we define say henikoff weight or in some way.

So, that image position the probability of the occurrence of each amino acid will be reflected instead of the theoretical probability of theoretical probability of occurrence of all 20 amino acids at that position with equal probability. So, that is the basic purpose. Next, we have to perform a simulation at the core we use Monte Carlo simulation method with Metropolis criteria. Metropolis criteria will be there.

So, this algorithm is Abinitio protein design it will look for a new protein sequence brand new protein sequence that is why it is starting from a random sequence but the random sequence will be of the same length as the input protein structure. So, what you understand when the, before the simulation, I already computed one matrix the size of the matrix is n cross 20 where n is the length of the input protein structure and 20 is the number of amino acids.

So, at each position what is the probability of that amino acid so, I computed that. Now, in the replica exchange Monte Carlo what I will do I will start with some random sequence. So, I will randomly generate some sequence and after randomly generate some sequences as per the algorithm of the Monte Carlo simulation method.

So, it will test whether that sequence is a valid one or not based upon some scoring function. So, let us assume there exists one scoring function the exact definition and the declaration of the scoring function I will defer for the timing we will discuss that on the next lecture. But if such scoring function exists then what it will do so far the first sequence definitely since there is no previous so, it will accept for the second sequence.

So, how it will generate up randomly it will pick some position and randomly also it will mutate some position so, after mutating that position it will have a new sequence now, the energy between these two sequences will be compared after comparing that one then it will either select or reject. So, let us focus on this third step that is perform Monte Carlo simulation.

(Refer Slide Time: 25:38)

Random Seq

1 11 23 29 51 N

C → T
h → A
seqL E → E
C → Y

$E(seqL) < E_1$
 $E(seqL) > E_2$

② For each position pick one amino acid that will replace/replace the existing one.

① Random numbers to identify the positions of the sequence for mutation

$\Delta E = E_2 - E_1$
if $\Delta E < 0$ accept E_2
else $q \rightarrow U(0,1)$ if $q > e^{-\Delta E}$ accept for mutation

The Algorithm

Algorithm 14:

Input: A protein structure

Output: Novel protein sequences that will fold to that given protein structure.

Steps:

1. For the given protein structure identify the homologous experimental structures.
2. Compute the PSSM from the homologous structure.
3. Perform Monte Carlo (MC) simulation starting from a random sequence.
4. Analyze the accepted new sequences generated during MC simulation to output best candidate solution.

Pralay Mitra

The Algorithm


Algorithm 14:

Input: A protein structure
Output: Novel protein sequences that will fold to that given protein structure.

Steps:

1. For the given protein structure identify the homologous experimental structures.
2. Compute the PSSM from the homologous structure.
3. Perform Monte Carlo (MC) simulation starting from a random sequence.
4. Analyze the accepted new sequences generated during MC simulation to output best candidate solution.

Pralay Mitra



So, what I said that first what I will do that I will have one random sequence the length must be same as the input. So, if it is then let us assume this is my sequence and length is N so, since N then let us start with 1. So, this is my random sequence. And I deferred the discussion regarding the energy function for the time being but let us assume there exists one energy function and that energy function will be computed if I say this is my seq1 then on seq1 it will be computed fine since it is my first sequence.

So, I will accept this one now, let us follow me how I am generating the second sequence from this first sequence. So, what I will do randomly I will choose some locations. So, randomly say four locations I have identified I can do that one after randomly identifying four locations at that positions since I identified those four locations without a loss of generality let us assume this is 11 this is 23 this is 29 and say this is say 51 randomly I picked.

Now, I will consult that N cross 20 matrix PSSM matrix positions specific scoring matrix where the probability has been computed and stored there. So, what I can do that randomly I will generate one number and I will check the probability at that particular matrix if the probability is in that range then I will select that particular amino acid to be replaced at this position by that.

So, initially if say I had cysteine and I picked say (27:57) then I will change C to T here similarly if initially say I have G and now I picked say Elenin then G to A I agree there is a possibility that initially there was say initially there was E and I also picked E. So, I will pick E. So, I replace that is also possible and I can do that one.

Now, for 51 also let us assume I picked C and it is Y. So, two different random numbers one random number to identify the positions of the sequence for mutation and this is one and for each position pick one amino acid that will replace or substitute that existing one. So, this way I modified my random sequence which was initially random to another random sequence.

Let us assume that is my seq2. So, I have initially it was E seq1. Now, I have E seq2 now you know what is Monte Carlo simulation steps. So, since I generated this E seq2 if I assume that this is my E2 and this is my E1. So, ΔE is going to be $E2 - E1$ if ΔE is less than 0 except $E2$ else so, I have to go for Metropolis Monte Carlo simulation technique.

So, you generate one random number $q \in [0, 1]$. So, U generate one random number then if q greater than $e^{-\Delta E / T}$ then accept $E2$ otherwise reject so, this ΔE is here and T is my temperature again this temperature has nothing to do with the environmental temperature of the protein this temperature you have to decide by looking at the distribution of your system or how you are generating this then based upon that one you have to compute this one.

So, this is the steps for performing Monte Carlo simulation starting from a random sequence. Now, the fourth step is analyze the accepted new sequences generated during MC simulation to output based candidate solution. So, definitely that is going to be my final step for Monte Carlo also.

So, this fourth step says that you have generated you have accepted several such sequences say I have again without any loss of generality m number of sequences accepted during MC simulation if m then what I have to do analyze regarding the analysis my hypothesis is if the similar kind of sequences generated frequently then there is a possibility that, that sequence is going to be the best one.

So, what I will do I will cluster them based on during the clustering process what I will do so, I will compute their pairwise distance based upon these blosum 62 matrix use blosum 62 matrix for computing distance between two amino acid sequences. So, that way I compute the distance. Now, that once I computed the distance based upon that one I will cluster. So, once I will cluster then I will have several clusters.

So, cluster 1 cluster 2 cluster 3 each cluster will have some cardinality I will sort based upon their cardinality and with an assumption that the highest cardinality is my best choice. So, that is my first rank solution second best is my next rank solution that way I will rank the clusters what is going to be my pick from that cluster whose distance is less compared to the other sequences or which is the presenting the center of the cluster will be outputted as the sequence representing that cluster and that is going to be my design sequence after this step four. So, we will continue in the next lecture. Thank you very much.