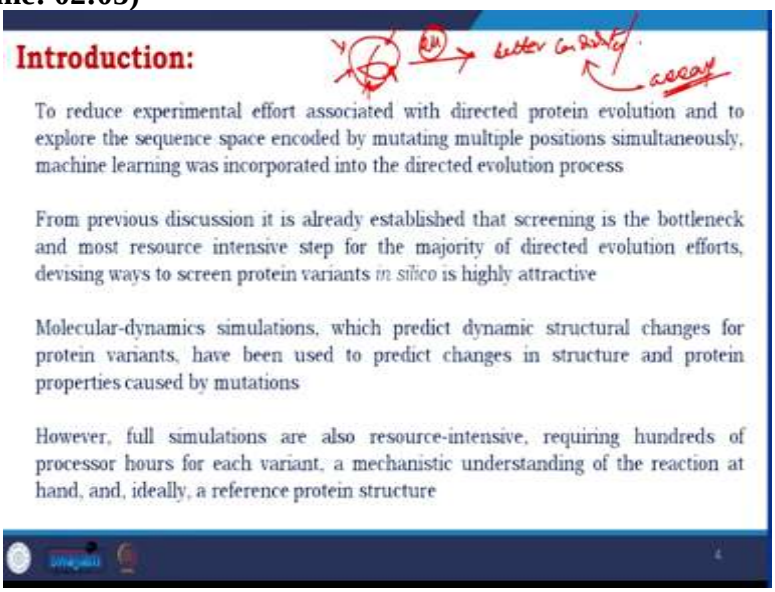
Structural Biology
Prof. Saugata Hazra
Department of Biotechnology
Indian Institute of Technology – Roorkee

Lecture – 55
Achievements, Challenges and Future Direction in the Field of Protein Engineering

Hi everyone, welcome again to the course of structural biology, we are at the end of our 11th module protein engineering, this is the last class. In this class, because before we have discussed about what, which and how about protein engineering, the processes, rational designing, directed evolution, there details, the associated work, how the flow and all. Today, I will discuss machine learning approach along with the success stories which help us understanding.

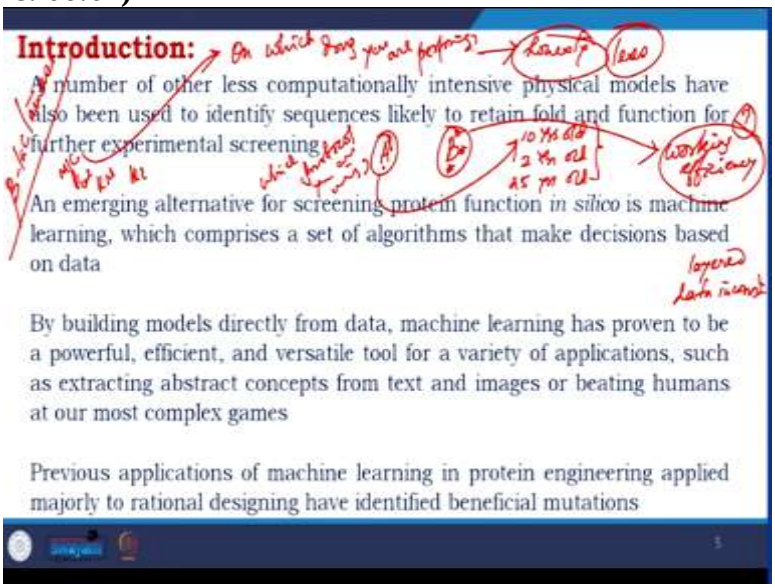
(Refer Slide Time: 02:05)



To reduce the experimental effort associated with directed protein evolution. In addition, we are to explore the sequence space encoded by mutating multiple positions simultaneously, more complex designing of mutational like spectrum in each generation and that is why we want to introduce machine learning. From previous discussion, it is already established that screening is the bottleneck and most resource intensive step for the majority of directed evolution effort. So that is why we told that our previous discussion already established that screening is the bottleneck and most resource intensive step for the majority of the directed evolution efforts. Now devising ways to screen protein variants in silico is highly attractive. Molecular-dynamic simulation, which we have talked about in the previous module, predict the dynamic structural changes for protein variants have been used to predict changes in structure and protein properties caused by mutations.

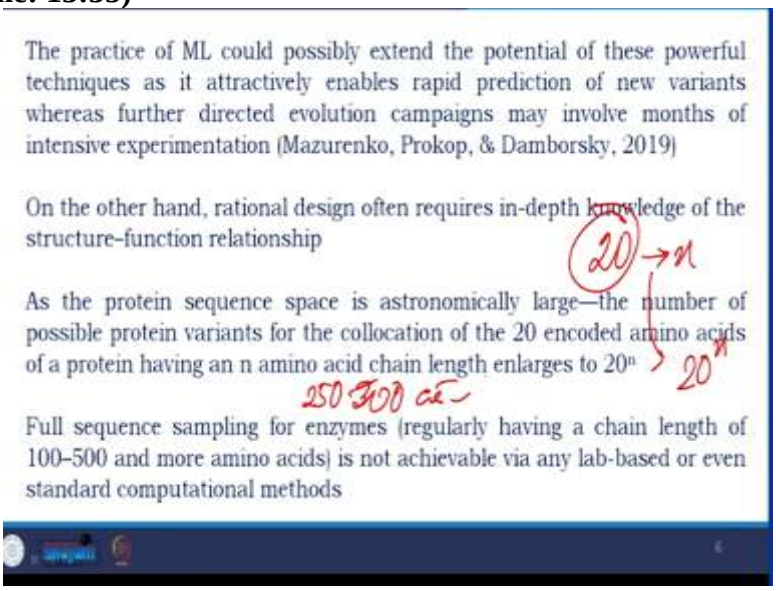
But, if you want to perform a full simulation of a protein that is also computed intensive, and you already know the huge number we have to target, so that is definitely computer intensive. So, it requires 100's of processors hours for each variant. And a mechanistic understanding of the reaction at the hand and a reference protein structure you needs them to perform simulation.

(Refer Slide Time: 06:07)



A number of other less computationally intensive physical models have also been used to identify sequences likely to retain fold and function for further experimental screening. An emerging alternative for screening protein function in silico is machine learning, which comprises a set of algorithms that make the decision based on data. They could correlate that with the working efficiency. It is true that in many fields other than biology, machine learning is now getting tremendous success in biology, the success is not in that measure the reason is also understandable. But, when we come in case of protein, the structural biologists, the biochemist bio physicist, because they work on fundamental principles, the relativeness is much less here. So, you could expect a lot data consistency here. And that is why it tremendous attention is now coming, because we already know the fundamentals of protein structure. And those information are making AI based or machine learning based prediction more efficient day by day. By building models directly from data, machine learning has proven to be a powerful, efficient and versatile tool for a variety of applications such as extracting abstract concept from text and images or beating humans at our most complex games. Previous applications of machine learning in protein engineering applied majorly to rational designing have identified beneficial mutations.

(Refer Slide Time: 13:35)

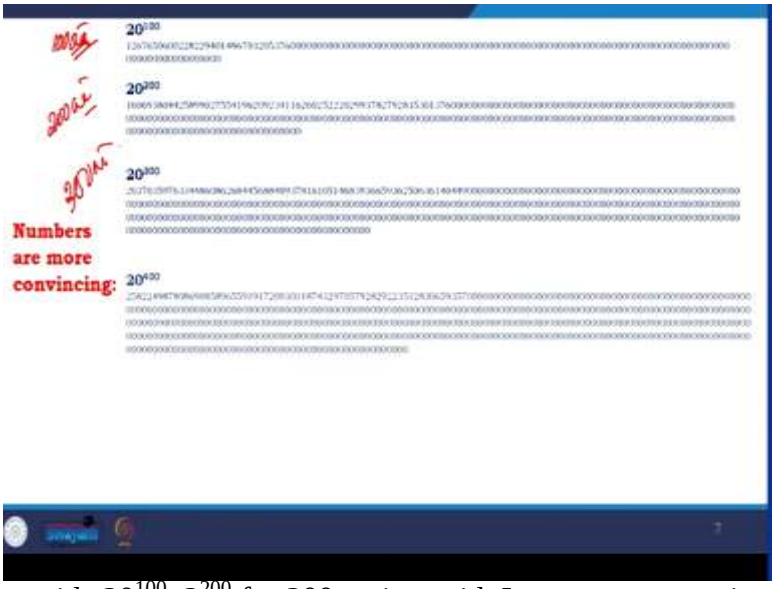


So, we already have some success in a little and it was applied to rational designing because logic is there. So, we have the logic, we take those logic in the data set, we incorporate them and develop the algorithm. The practice of machine learning could possibly extend the potential of these powerful techniques as it attractively enables rapid prediction of new variants, whereas further directed evolution campaigns may involve months of intensive experimentation.

On the other hand, rational design often requires in depth knowledge of the structure function relationship. De novo design or spatially rational designs need logistics, it needs the fundamentals.

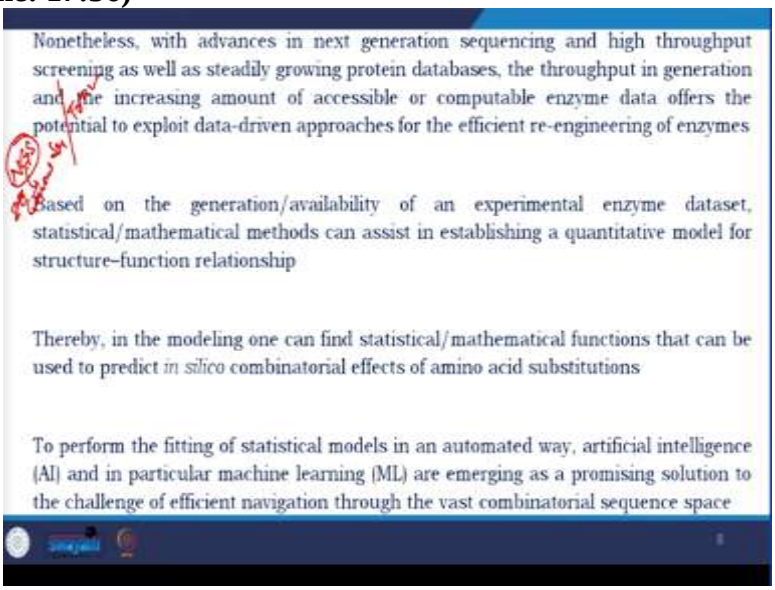
As the protein sequence space is astronomically large, the number of possible protein variant for the total collection of 20 encoded amino acids of a protein having 'n' amino acid chain length enlarges to 20^n. Full sequence samplings of enzymes were regular 100 to 500, because we know 250 to 300 amino acid is average. So, you could say 100 to 500 is not achievable by any lab based on even standard computational method.

(Refer Slide Time: 16:39)



So, for 100 amino acids 20^{100}, 2^{200} for 200 amino acid, I am not even saying the number, you look at the number, you look at the 0 and you will convince this is not possible for human brain, this is not possible for simple computing. So, you need prediction, you need logistics, you need biasness and that is why the importance of machine learning.

(Refer Slide Time: 17:36)

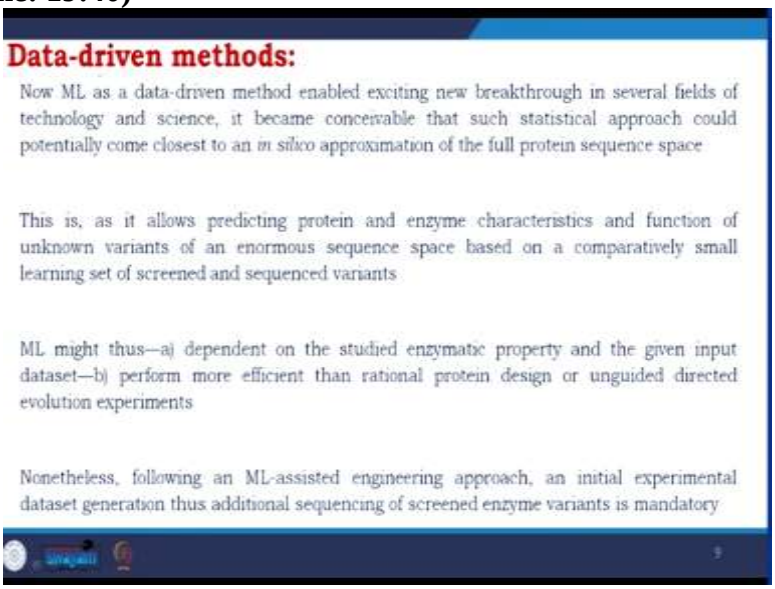


Nonetheless, with advances in next generation sequencing and high throughput screening, as well as steadily growing protein databases, the throughput in generation and the increasing amount of accessible or computable enzyme data offers the potential to exploit data driven approaches for the efficient re-engineering of enzymes.

Based on the generation/availability of an experimental enzyme dataset, statistical/mathematical methods can assist in establishing a quantitative model for structure function relationship. Thereby, in the modelling one can find statistical mathematical functions that can be used to predict in silico combinatorial effects of amino acid

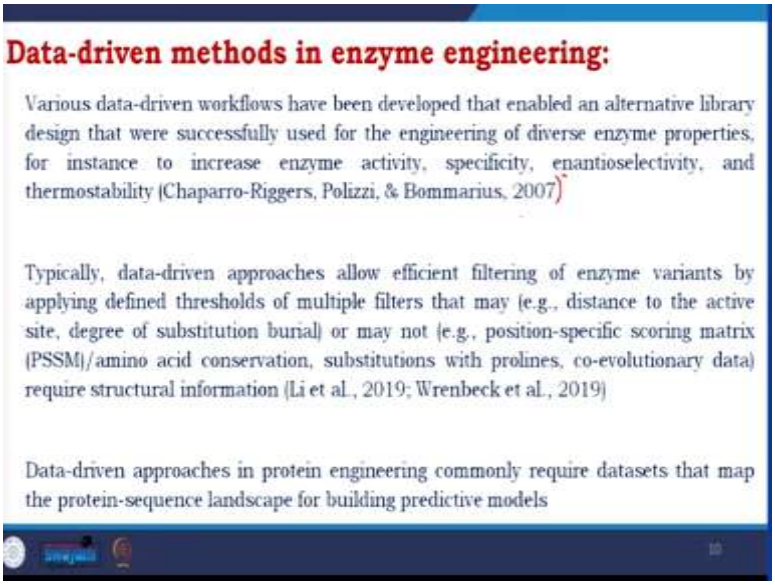
substitutions. To perform the fitting of statistical models in an automated way, artificial intelligence and in particular machine learning are emerging as a promising solution to the challenge of efficient navigation through the first combinatorial sequence space.

(Refer Slide Time: 19:40)



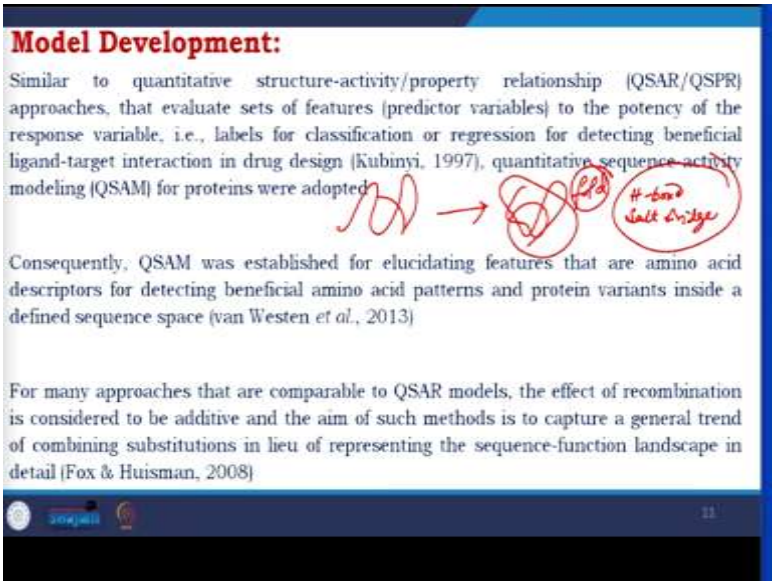
Now machine learning, as a data driven method enable exciting new breakthroughs in several fields of technology and science, it becomes conceivable that such statistical approach could potentially come closest to an in silico approximation of the full protein sequence space. This is, as it allows predicting protein and enzyme characteristics and function of unknown variant of an enormous sequence is based on a comparatively small learning set of screens and sequence variants. Machine learning works on algorithm, where datasets are there, more and more data are coming, more and more machines learning program would be bettering them. Machine learning might thus dependent on the studied enzymatic property in the given input data set, and perform more efficient than rational protein design or unguided directed evolution experiment. Nonetheless, following an MLS system engineering approach, an initial experimental data set generation thus additional sequencing of screen enzyme variant is mandatory.

(Refer Slide Time: 22:17)



Various data driven workflows have been developed that enabled an alternative library design that was successfully used for engineering of diverse enzyme properties, for instance, to increase enzyme activity specificity enhanced your selectivity and thermal stability. Typically, data driven approaches allow efficient filtering of enzyme variants by applying defined threshold of multiple filters that may. For example, distance to the active side degree of substitution do real or they may not like they could not do the position specific scoring matrix or PSSM amino acid conservation substitution with prolines, co evolutionary data require structural information. Data driven approaches in protein engineering commonly required data sets that map the protein sequence landscape for building predictive models.

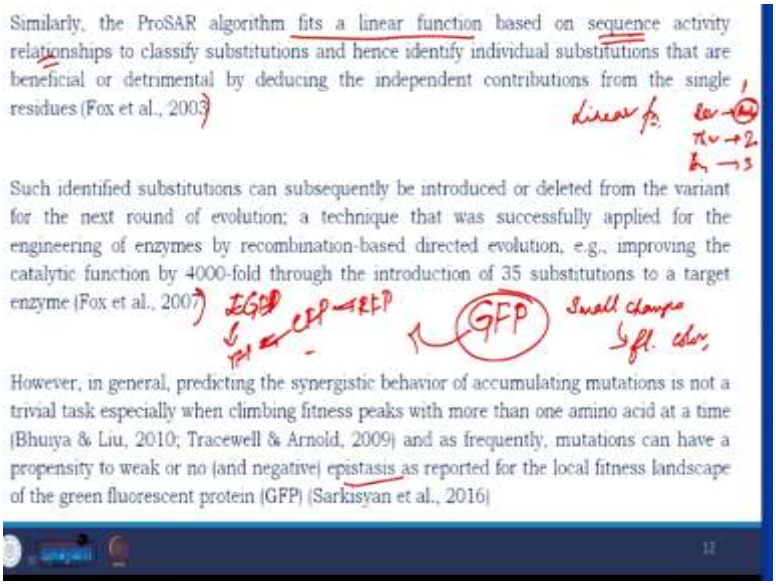
(Refer Slide Time: 23:13)



Similar to quantitative structure activity property relationship, which is QSAR and QSPR which we use in drug development in pharmaceutical industry. So, similar to that approaches which evaluate sets of pictures or predictor variables to the potency of the response variable levels for classification or regression for detecting beneficial ligand target interaction in drug designing, quantitative sequence activity modelling or QSAM proteins were adopted.

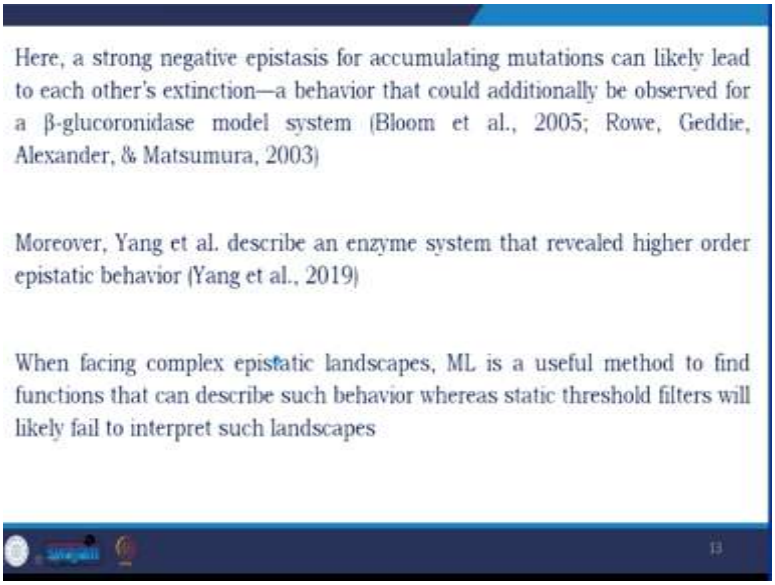
So, if you know QSAR, QSPR, they are known method where we look at the structure function of small molecule, the structure property and then correlate those with activity. So, you could say we kind of look at features and provide a scoring or a number to them. Similar approach develops, which could alternatively say as quantitative sequence activity modelling QSAM. Consequently, QSAM was established for elucidating features that are amino acid descriptors for detecting beneficial amino acid patterns and protein variants inside a defined sequence space. So, what types of properties are there, let us say you are looking at a protein, the challenge here is to understand how the proteins fold, you develop a network in the protein using hydrogen bond and salt bridge, they would help you enable to make introduction of the alternative amino acids. And that is what QASM used to do. For many approaches that are comparable to QSAR models the effect of recombination is considered to be additive. And the aim of such method is to capture a general trend of combining substitutions in lieu of representing the sequence function landscape in detail.

(Refer Slide Time: 26:16)



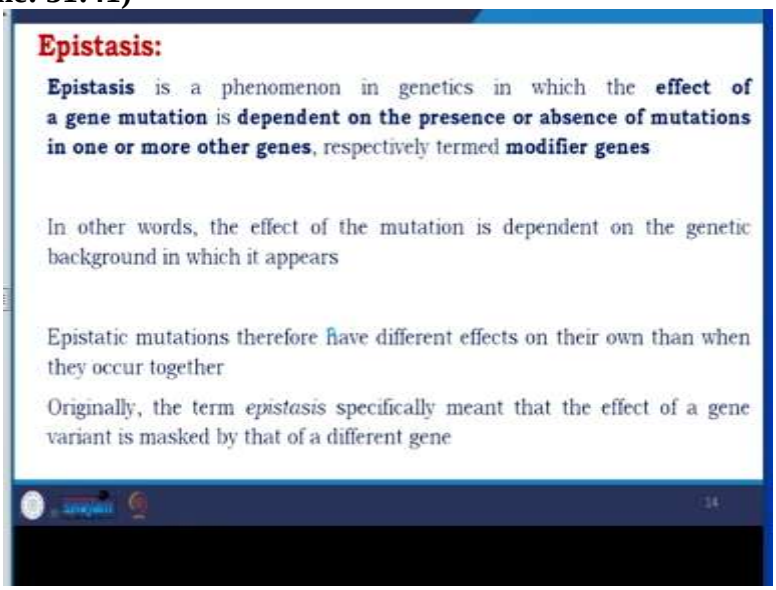
Similarly, the ProSAR algorithm fits a linear function based on sequence activity relationship to classify substitution, and hence identify individual substitutions that are beneficial or detrimental by deducing the independent contribution from the single residues. Such identified so you get improved versions, and you get the combination of the improved data set. You said those identified substitutions which are doing better, can subsequently be introduced or deleted from the variant of the next round of evolution. A technique that was successfully applied for the engineering of enzymes by recombination based directed evolution for example, improving the catalytic function by 4000 fold through the introduction of 35 substitutions to a target enzyme. However, in general, predicting the synergistic behaviour of accumulating mutation is not a trivial task, especially when climbing fitness peaks with more than one amino acid at a time. And as frequently mutation can have a propensity to weak or no epistasis, as reported for the local fitness landscape of the green fluorescent protein. I talked about the green fluorescent protein or GFP which is a around 200 plus amino acid protein. You make small changes and it changed its fluorescent colour. But in some case, there might be a possibility that a change of amino acid could provide a new flow since, but the whole enzyme become unstable, which is not normally the case for GFP protein because if you remember this is a beta barrel protein with extreme stability.

(Refer Slide Time: 30:44)



Here a strong negative epistaxis for accumulating mutation can likely lead to each other's extinction a behaviour that could additionally be observed for a beta glucoronidase model system. Moreover, Yang et al describe an enzyme system that revealed higher order epistatic behaviour. When facing complex epistatic landscapes machine learning is a useful method to find functions that can describe such behaviour whereas static threshold filters will likely fail to interpret such landscapes.

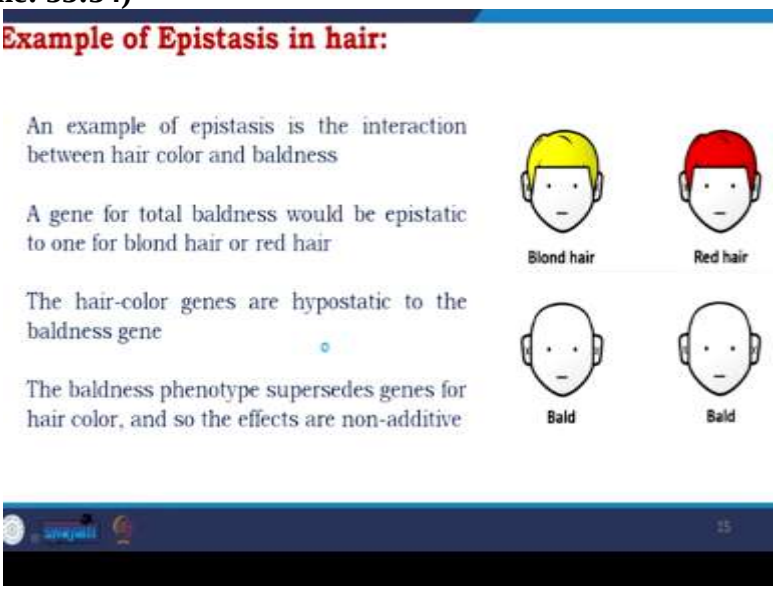
(Refer Slide Time: 31:41)



Epistasis is a phenomenon in genetics in which the effect of a gene mutation is dependent on the presence or absence of mutations in one or more other genes, respectively, termed modifier gene. So, a mutation in one gene, that effect is depend on a modifier gene that is called epistasis. In other words, the effect of the mutation of gene A is dependent on the genetic background. Epistatic mutations therefore have different effects on their own than

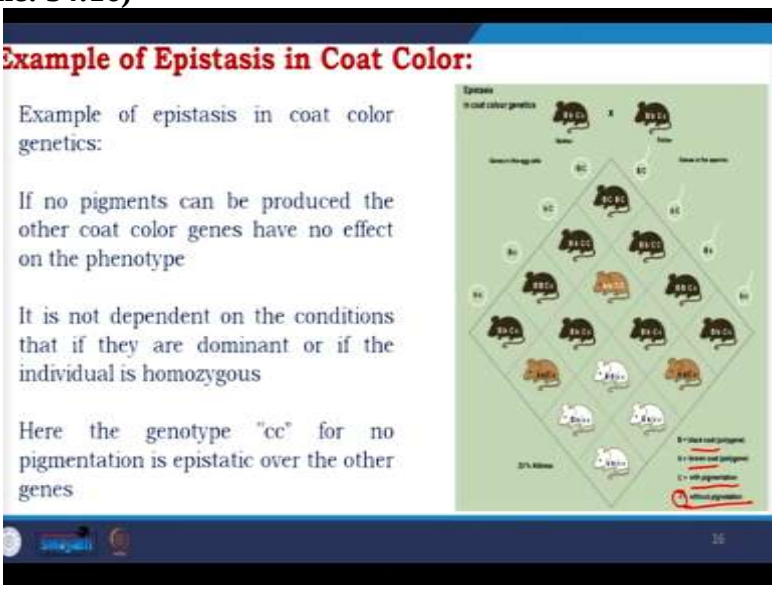
when they occurred together. Originally, the term epistasis specifically meant that the effect of a gene variant is masked by a different gene.

(Refer Slide Time: 33:34)



So, I am giving some example, an example of epistasis is the interaction between hair colour and baldness. Here you see two people, one with blond hair and other with red hair. A gene for total baldness would be epistatic to one for blond hair or red hair. The hair colour genes are hypostatic to the baldness gene. The baldness phenotype supersedes genes for hair colour and so the effects are non-additive.

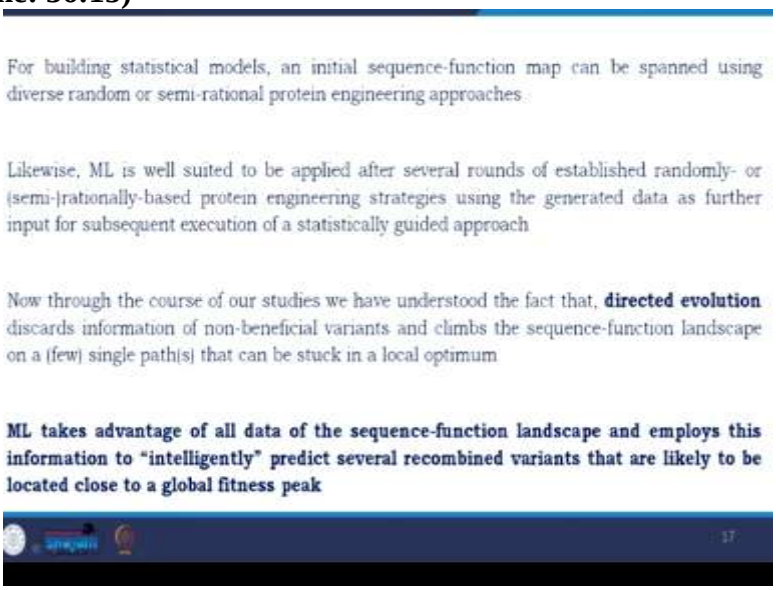
(Refer Slide Time: 34:16)



Another example you will see it is a coat colour genetics, mother is Bb cc and father is also BB so B dominant b recessive C dominant c recessive. And then you see different combinations B dominant B, so this is when I said dominant its means this B is for black coat. The small b is for brown coat these big C is with pigmentation small c without

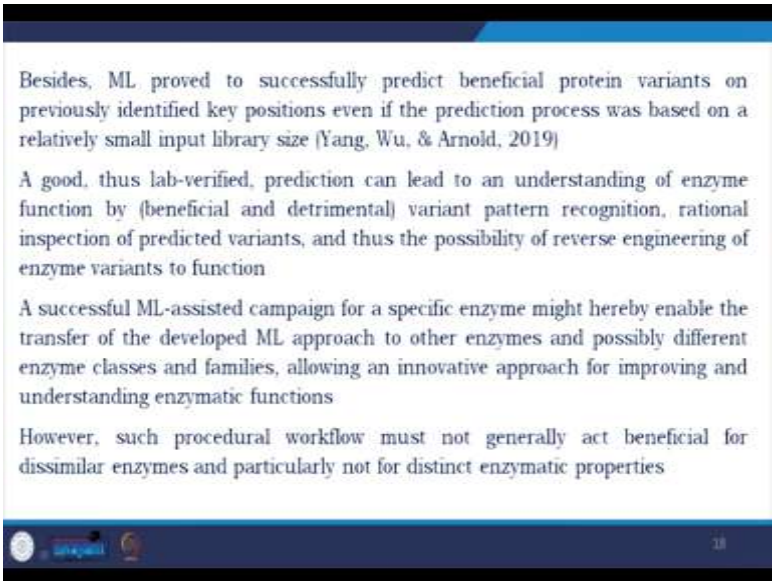
pigmentation which you will see here. So, if no pigment can be produced the other coat colour genes have no effect on the phenotype. So, if you have cc, without pigmentation from both the side you will not see the effect of other colour black brown anything else you will get them I will be albino. It is not dependent on the conditions that if they are dominant or if the individual is homozygous. Here the genotype cc for no pigmentation is epistatic over the other genes. If you see here BB that is albino which is for brown, when you are black and brown, they are also albino when to be for black coat they are also all we know and do the same. So, all the combinations when there is no pigmentation dominate over and make them colourless, that is epistatic effect.

(Refer Slide Time: 36:15)



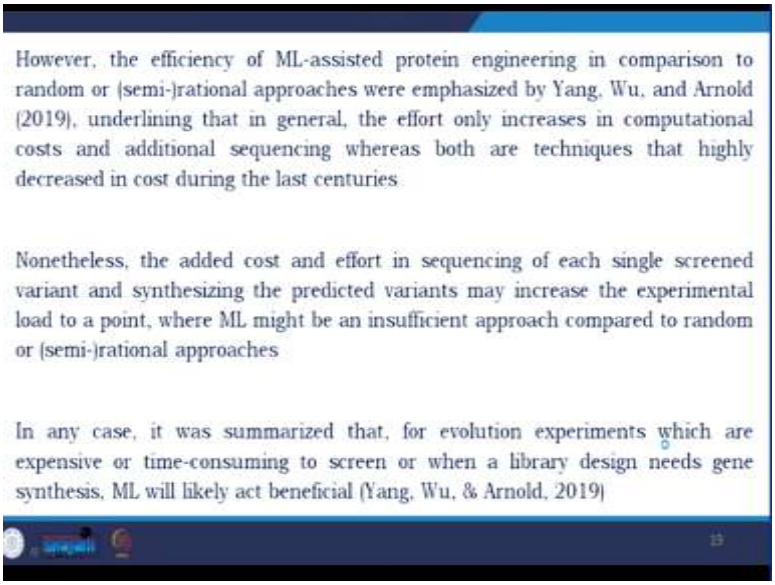
For building statistical models, an initial sequence function map can be spanned using diverse random or semi rational protein engineering approaches. Likewise machine learning is well suited to be applied after several rounds of establish randomly or semi rationally based protein engineering strategies using the generated data as further input for subsequent execution of a statistically guided approach. Now, through the course of our studies, we have understood the fact that directed evolution discards information of non-beneficial variants and climbs the sequence function landscape on a single path that can be stuck in a local optimum. ML takes advantage of all data of the sequence function landscape and employs this information to intelligently predict several recombined variants that are likely to be located close to a global fitness peak.

(Refer Slide Time: 38:22)



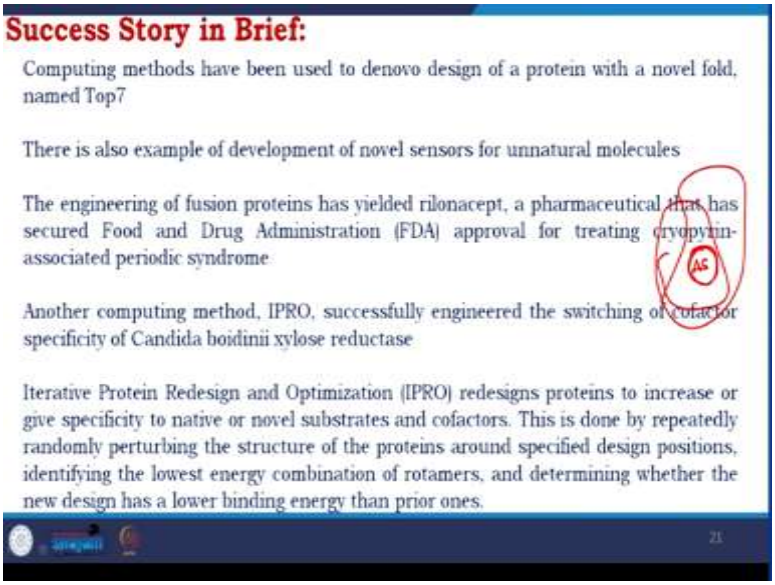
Beside, this machine learning proved to successfully predict beneficial protein variants on previously identified key position even if the prediction process was based on a relatively small input library size. A good, thus lab verified prediction can lead to an understanding of enzyme function by beneficial and detrimental variant pattern recognition, rational inspection of predicted variants, and thus the possibility of reverse engineering of enzyme variants to function. A successful ML assisted campaign for a specific enzyme might here by enable the transfer of the develop machine learning approach to other enzymes and possibly different enzyme classes and families allowing an innovative approach for improving an understanding enzymatic function. However, such procedural workflow must not generally act beneficial for dissimilar. So, as I say the term individualization when the enzymes are totally dissimilar, then this individualization is not easy and the procedural workflow would not work and particularly not for distinct enzymatic properties.

(Refer Slide Time: 40:35)



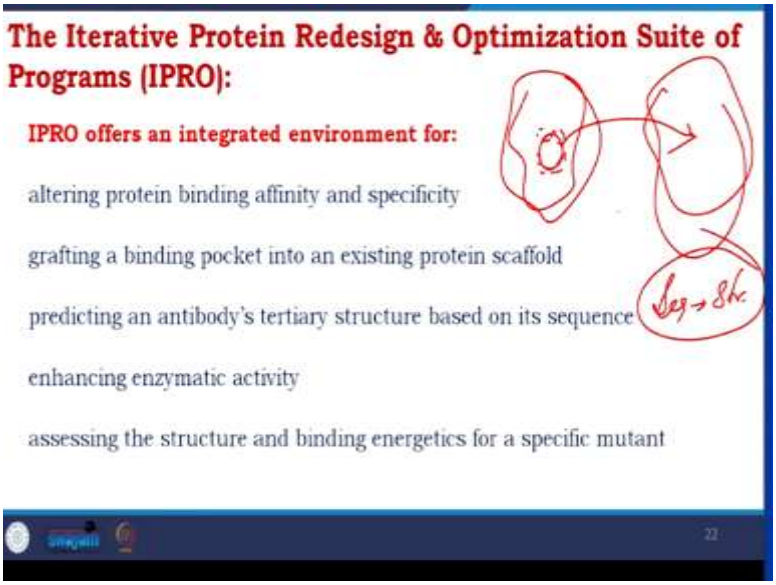
However, the efficiency of ML assisted protein engineering in comparison to random or semi rational approaches were emphasized by Yang, Wu Arnold in 2019, underlining that, in general, the effort only increases in computational costs and additional sequencing, whereas both are techniques that highly decreased in cost during the last centuries. Nonetheless, the added cost and effort in sequencing of each single screened variant and synthesizing the predicted variant may increase the experimental load to a point. Where machine learning might be an insufficient approach compared to random or semi rational approaches. In any case, it was summarized that for evolution experiments, which are expensive or time consuming to screen, or when a library design needs, gene synthesis, ML will likely act beneficial.

(Refer Slide Time: 41:54)



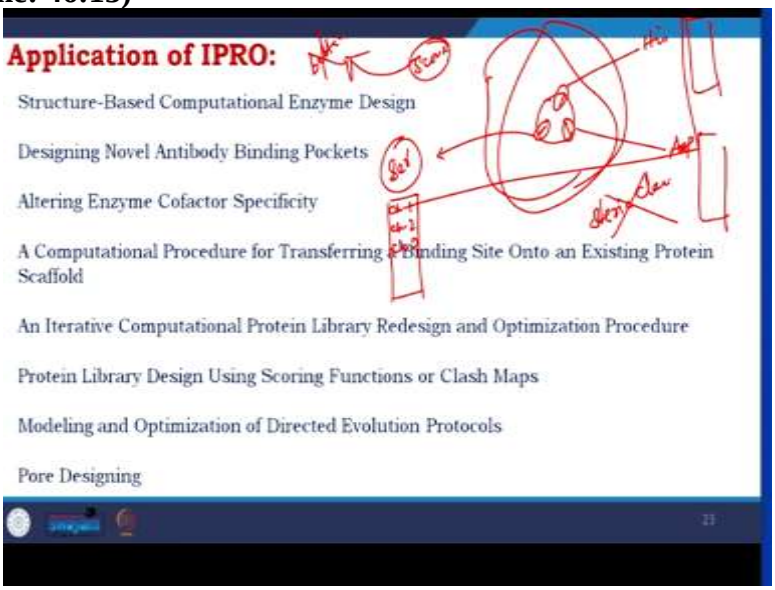
Computing methods have been used to denovo design of a protein with a novel fold named top 7, this is the first protein which was expressed and the determined structure clearly established the fact that in nature, we do not have such fold this is a landmark. There is also example of development of novel sensors for unnatural molecules, I have already shown that how alpha helices are taken and then dimer, trimer, tetramer are made and again the structure solution clearly source that they have adopted very different confirmations. The engineering of fusion proteins has yield rilonacept, a pharmaceutical that has secured FDA approval for treating cryopyrin associated periodic syndrome. Another computing method IPRO successfully engineered the switching of cofactors specificity of Candida boidinii xylose reductase. IPRO (Iterative Protein Redesign and Optimization) redesigns protein to increase or give specificity to native or novel substitutes and cofactors. This is done by repeatedly randomly perturbing the structure of the protein around specified design position identifying the lowest energy combination of rotamers, and determining whether the new design has a lower binding energy than the prior one.

(Refer Slide Time: 45:16)



IPRO offers an integrated environment for altering protein binding affinity and specificity, grafting a binding pocket into an existing protein scaffold. IPRO Predicts an anybody's tertiary structure based on its sequence, enhancing enzymatic activity, the same way as assessing the structure and binding energetics for a specific mutant.

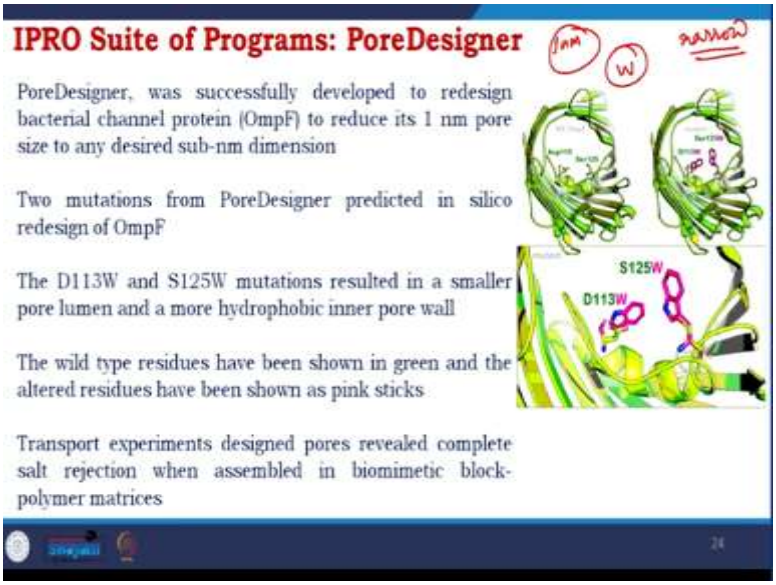
(Refer Slide Time: 46:13)



What are the applications of IPRO?

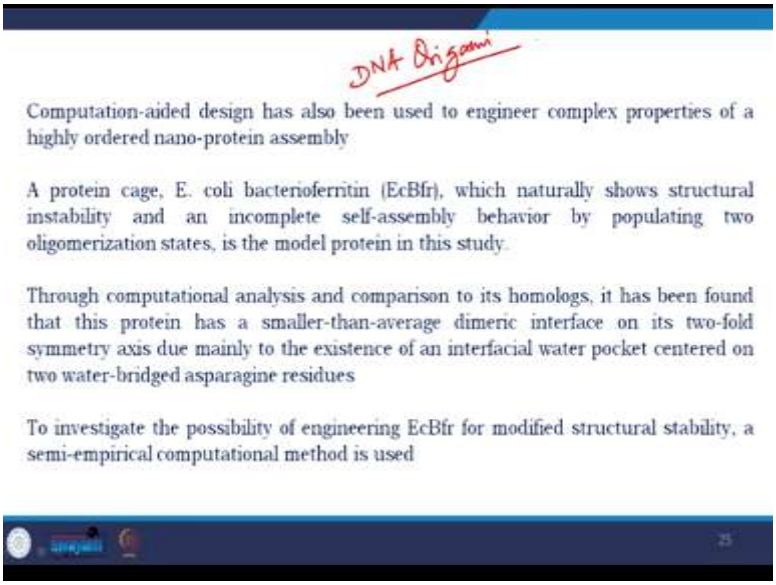
Structure based computational enzyme design, designing novel antibody binding pockets, altering enzyme cofactor specificity. A computational procedure for transferring a binding site onto an existing protein scaffold. An iterative computational protein library designs an optimization procedure, protein library design using scoring function or clash maps. Modelling and optimization of directed evolution protocols designing.

(Refer Slide Time: 48:28)



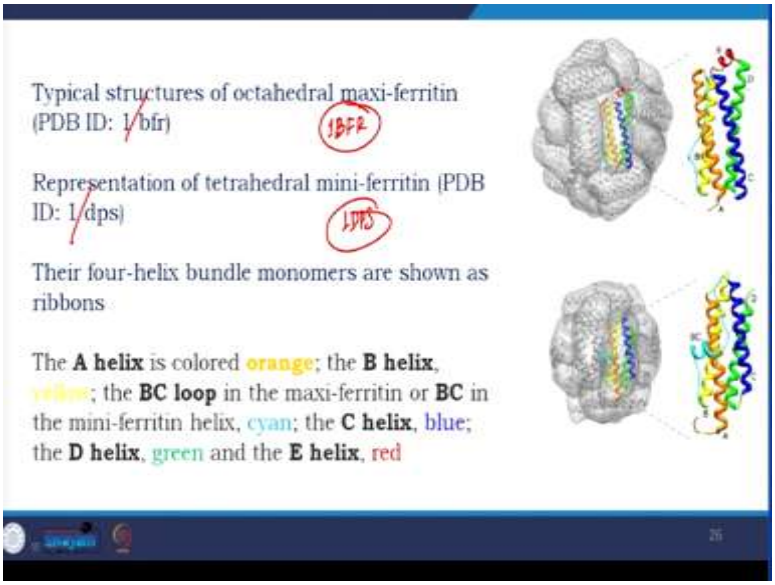
PoreDesigner (IPRO suite program) is a subset of the IPRO program, was successfully developed to redesign bacterial channel protein OmpF (channel protein), to reduce the one nanometer pore to sub nanometer dimensions. Two mutations from PoreDesigner predicted in silico redesign of OmpF protein. The D113W and S125W mutations resulted in a smaller pore lumen and more hydrophobic inner pore wall. The wild type residues have been shown in green and the altered residues have been shown in pink sticks as you could see here. Transport experiments designed pores revealed complete salt rejection, when assembled in biomimetic block polymer matrices row.

(Refer Slide Time: 51:01)



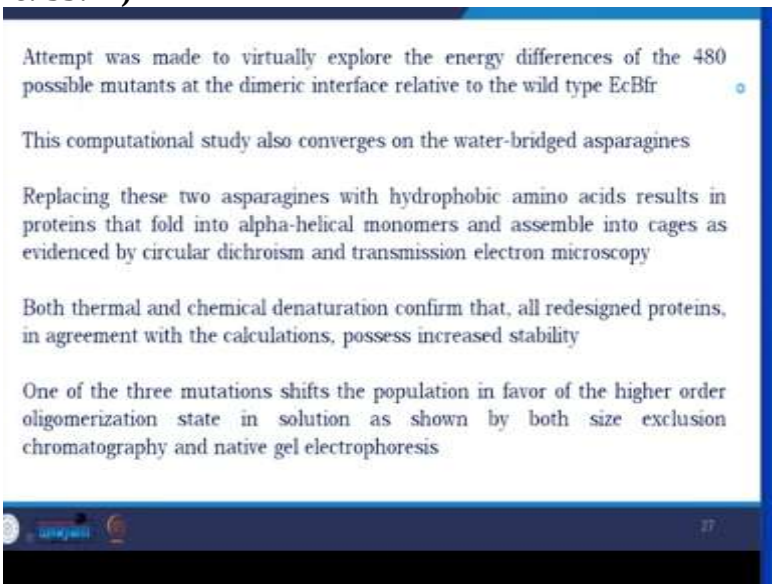
Computer aided design has also been used to engineer complex properties of highly ordered nano protein assemblies. Nano-protein assembly, it comes from the inspiration of DNA origami, DNA origami is a field where a lot of DNA is fold together to develop a kind of 2D, 3D structure. A protein cage, *E.coli* bacterioferitin (EcBfr), which naturally show structural instability and an incomplete self-assembly behaviour by populating to oligomerization states, is the model protein for the study. Through computational analysis and comparison to this homologues it has been found that this protein has a smaller than average dimeric interface on its 2 fold symmetry axis due mainly to the existence of an interfacial water pocket centered on that two water-bridge asparagine residues. To investigate the possibility of engineering EcBfr for modified structural stability is semi empirical computational method is used.

(Refer Slide Time: 52:53)



If you see this is the typical structure of octahedral maxi ferritin and mini ferritin, maxi ferritin is 1BFR (PDB ID) and 1DPS (PDB ID) for mini ferritin. The structure is 4 helix bundle monomer. If you see here that A helix is orange, the B helix is yellow, the BC loop in maxi ferritin in cyan, the C helix is in blue, the D helix is green and E helix is red which is present in maxi ferritin.

(Refer Slide Time: 53:44)



An attempt was made to virtually explore the energy differences of the 480 possible mutants at the dimeric interface relative to the wild type EcBfr. This computational study also converges on the water based asparagine. Replacing these two asparagines with hydrophobic amino acid results in protein that fold into alpha helical monomers and assemble into cage as evidenced by circular dichroism and transmission electron microscopy (TEM). Both thermal and chemical denaturation confirm that all redesigned protein in agreement with the calculation possess increased stability. One of the 3 mutations shifts the population in favour

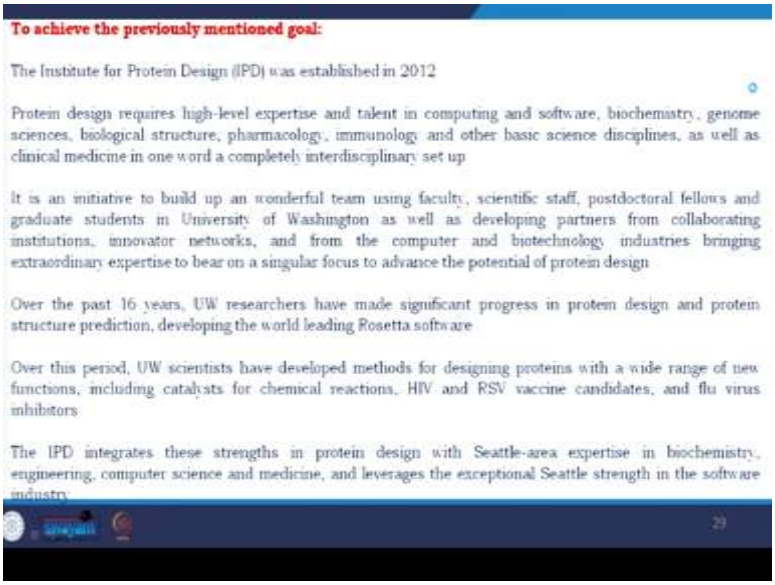
of the higher order oligomerization state in solution as shown by size exclusion chromatography and native gel or trees gel electrophoresis.

(Refer Slide Time: 55:26)

I will talk about Institute for Protein Designing in University of Washington, this is the Institute for protein designing, these guys David Baker, I am constantly talking about, he has taken instrumentals step in developing protein designing principles. Trisha Davis, PhD chair department of biochemistry, David Baker, PhD Director, Lance Stewart PhD MBA, Senior Director of strategy are the part of the organizational structure, but then there are many scientists they are working together. And how big the initiative you will get an idea looking at the funding organizations Bill and Melinda Gates Foundation, Defense Threat Reduction Agency DTRA, Defense Advanced Research Projects Agency DARPA, Department of Energy DOE, Life Science Discovery Fund LSDF, National Institute of Health NIH, National Science Foundation NSF, Takeda pharmaceuticals, Washington Research Foundation and beside them, a lot of individual organization individuals are contributing.

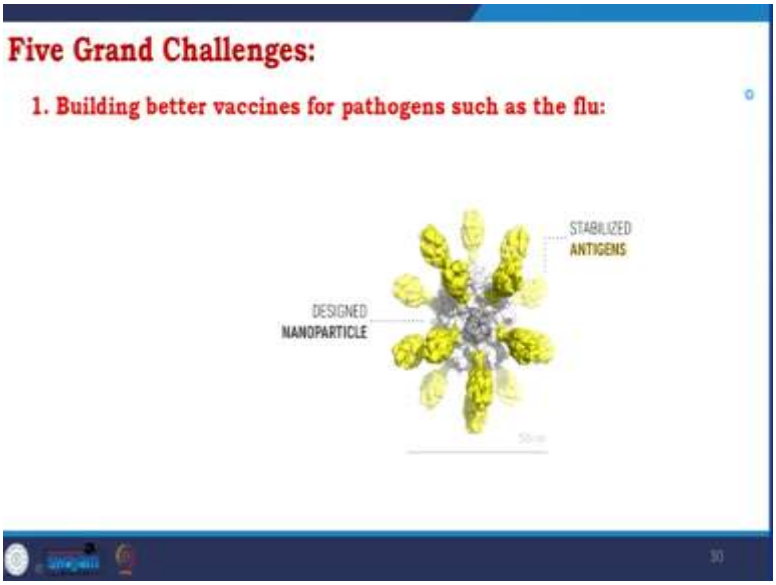
So, the exquisite function of naturally occurring protein is to solve the challenges faced during evolution. However, Institute of protein design or IPD, they face challenges that are not faced during natural evolution because of the current situation. The goal of the Institute of protein design IPD is to develop an apply method for designing a whole new world of synthetic proteins to address these challenges.

(Refer Slide Time: 58:06)



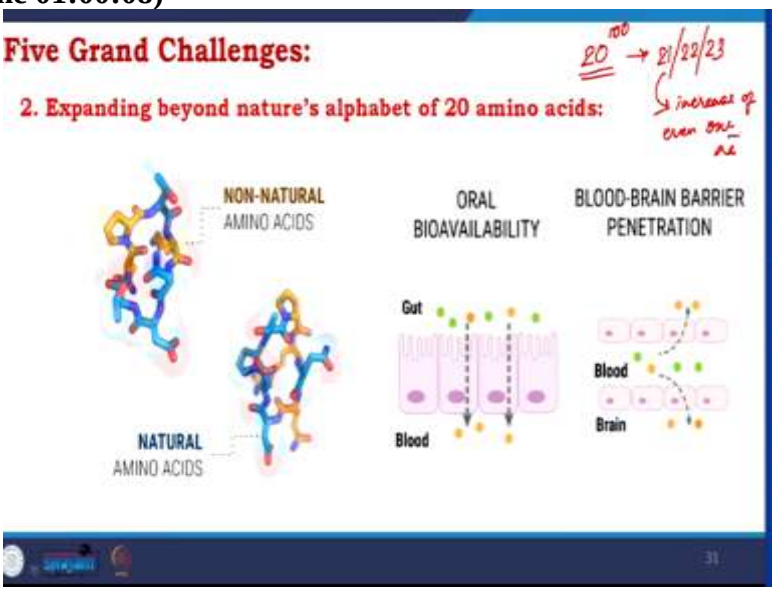
Institute for protein design was established in 2012. Protein design require high level expertise and talent in computing and software, biochemistry, genome science, biological structure, pharmacology immunology and other basic science disciplines as well as clinical medicine in one word, a completely, extremely interdisciplinary set up. It is an initiative to build up a wonderful team using faculty scientific staff, postdoctoral fellows and graduate student in University of Washington as well as developing partners from collaborating institution, innovator networks and from the computer and biotechnology industries, bringing extra ordinary expertise to be or on a singular focus to advance the potential of protein design. Over the past 16 years, University of Washington researchers have made significant progress in protein design and protein structure prediction, developing the world leading Rosetta software. Over this period those scientists have developed methods for designing protein with a wide range of new functions, including catalysts for chemical reactions, HIV and RSV vaccine candidates and flu virus inhibitors. The IPD integrate these strengths in protein design with Seattle-area expertise in biochemistry, engineering, computer science and medicine and leverage the exceptional Seattle strength in the software industry.

(Refer Slide Time: 59:50)



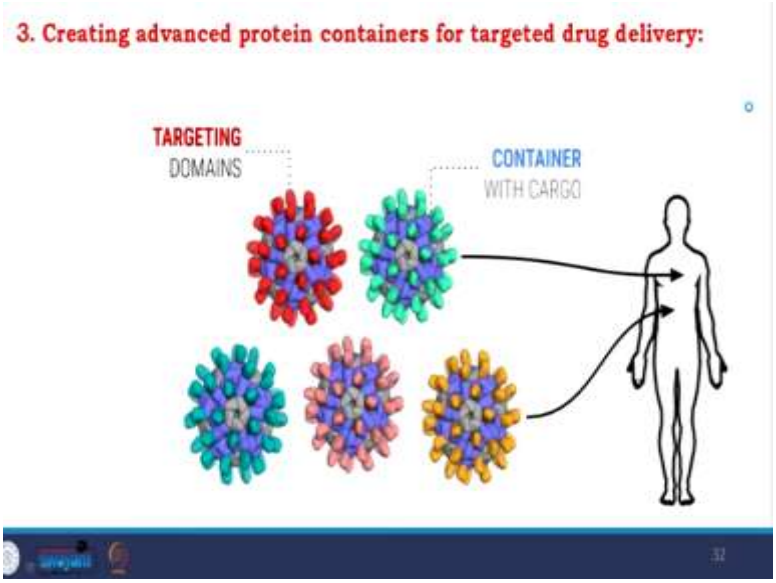
IPD looks for 5 grand challenges; they are working on building better vaccines for pathogens such as the flu. They combine the protein with nanoparticle; make many designing so that they could develop better vaccines.

(Refer Slide Time 01:00:08)



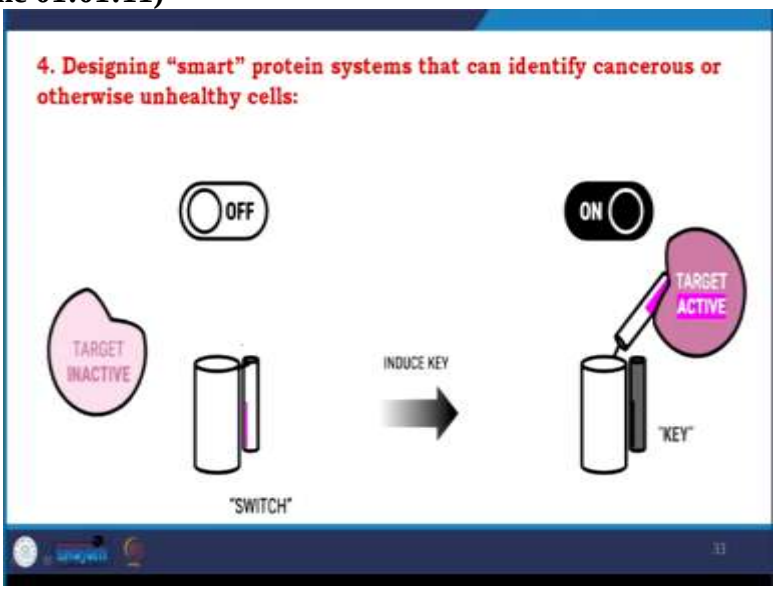
They are going to expand beyond nature's alphabet of 20 amino acids. So, if you remember, you just do the mathematics 20^{100} see the number and then do 21, 22, 23 and you see that increase, with increase of even 1 amino acid that is what they are doing.

(Refer Slide Time 01:00:48)



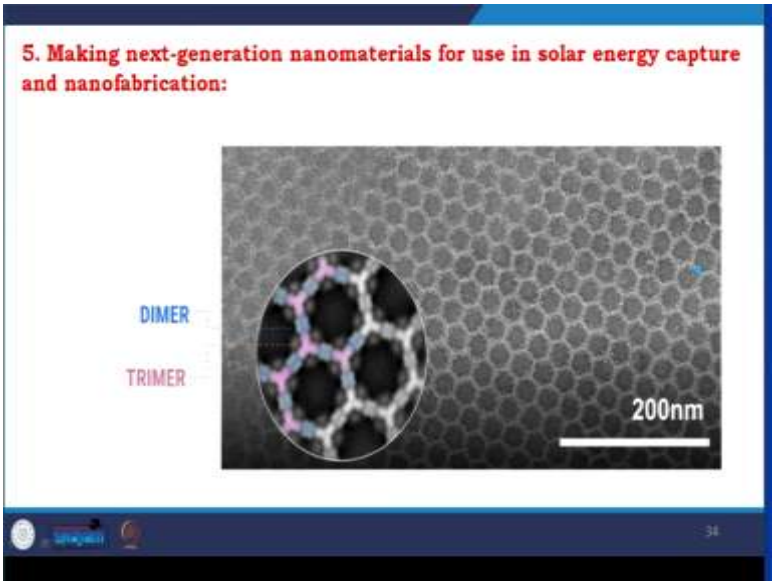
They are creating advanced protein containers for targeted drug, so you have target domains, they take them, so that they could target a particular region and then fill the inner with drugs with the one to be delivered there.

(Refer Slide Time 01:01:11)



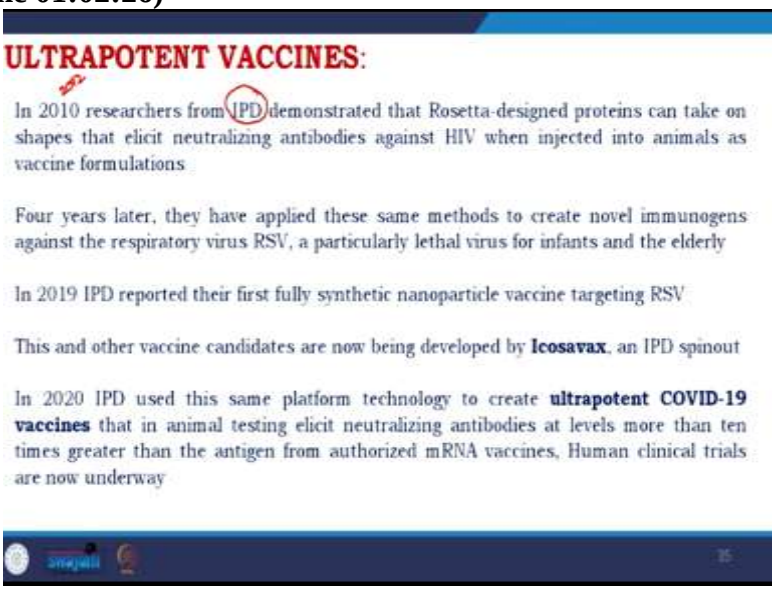
Designing smart protein system that can identify cancerous or otherwise unhealthy cells, they have switches where you keep the interaction. You have the target, which is inactive, then you have the switch with induce key, the switch would be activated. For example, let us say you have interaction and that interaction like a phosphorylation, merge with the kinase activity, this will interact with and then you get a FRET signal in that way.

(Refer slide Time 01:01:59)



Then, another very interesting aspect is making next generation nanomaterial for use in solar energy capture, nano fabrication and anything else. So, you could see that they have already developed, like if you see dimers trimers of proteins are there and they are used for development of those next generation nanomaterial.

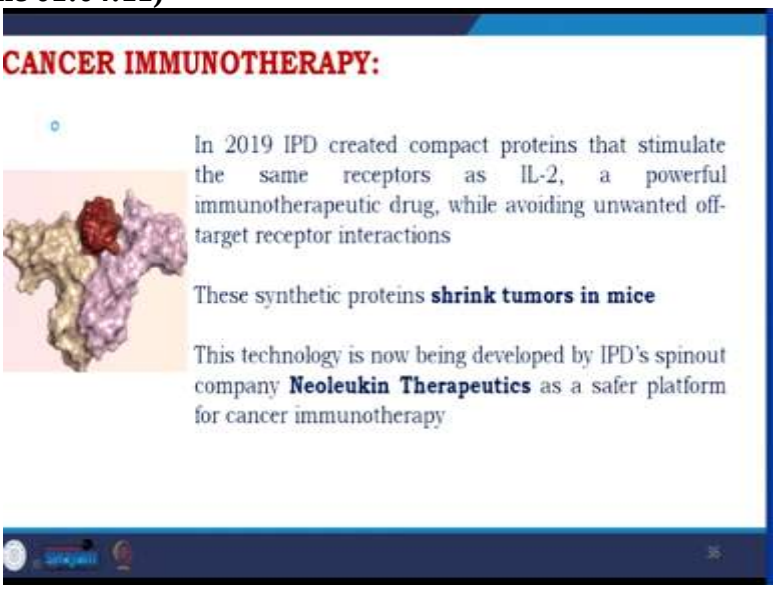
(Refer Slide Time 01:02:26)



Few success stories from David Baker lab and this protein design initiative, ultra-potent vaccines. In 2010, researchers from IPD demonstrated that Rosetta designed protein, can take on shapes that elicit neutralizing antibodies against HIV when injected into animals as vaccine formulation. 4 year later 2014, they have applied the same method to create novel immunogens against the respiratory virus, RSV, a particularly lethal virus for infants and the elderly. In 2019, they reported their first fully synthetic nanoparticle vaccine targeting respiratory virus RSV. This and other effects in candidates are now being developed by Icosavax, an IPD spinout company. In 2020 IPD use this same platform technology to create

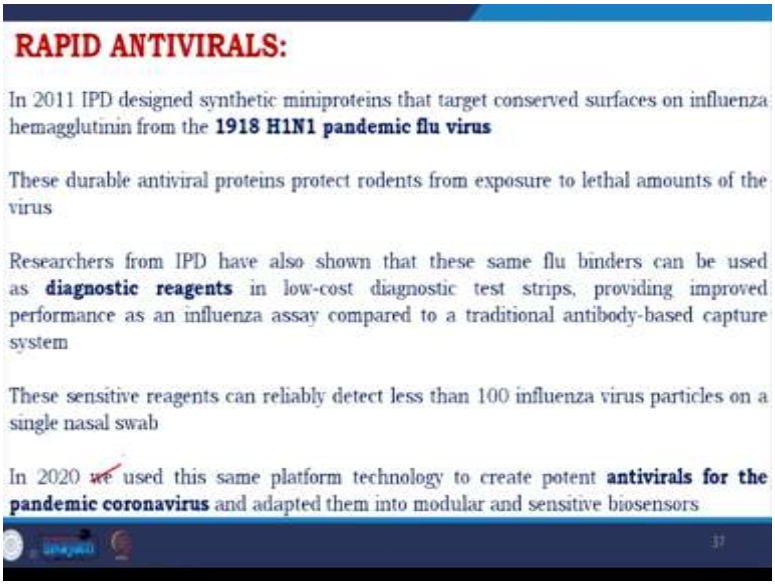
ultra -potent COVID-19 vaccine that is animal testing elicit neutralizing antibodies at levels more than 10 times greater than the antigen from authorized mRNA vaccines human clinical trials are now underway.

(Refer Slide Time 01:04:11)



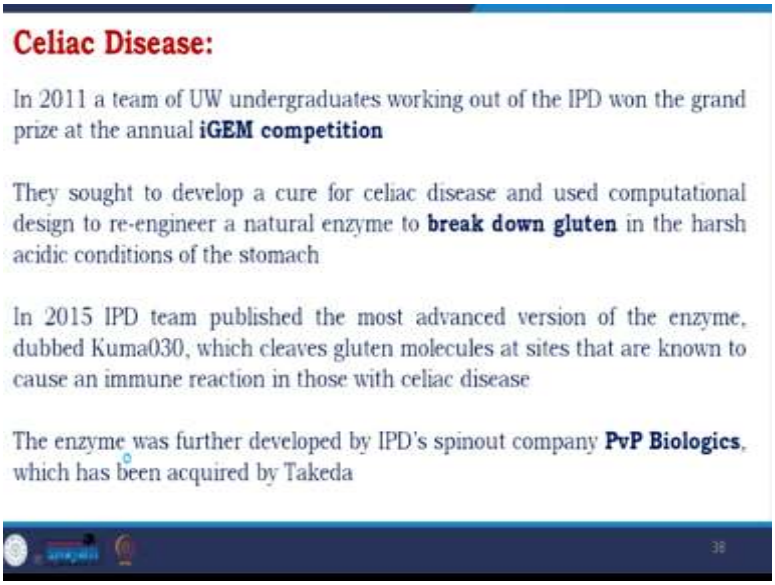
In 2019, IPD created compact proteins that stimulate the same receptor as IL-2, a powerful immunotherapeutic drug while avoiding unwanted of target receptor interactions. These synthetic proteins shrink tumors in mice already. This technology is now being developed by IPD Spinout Company, Neoleukin therapeutics as a safer platform for cancer immunotherapy.

(Refer Slide Time 01:04:51)



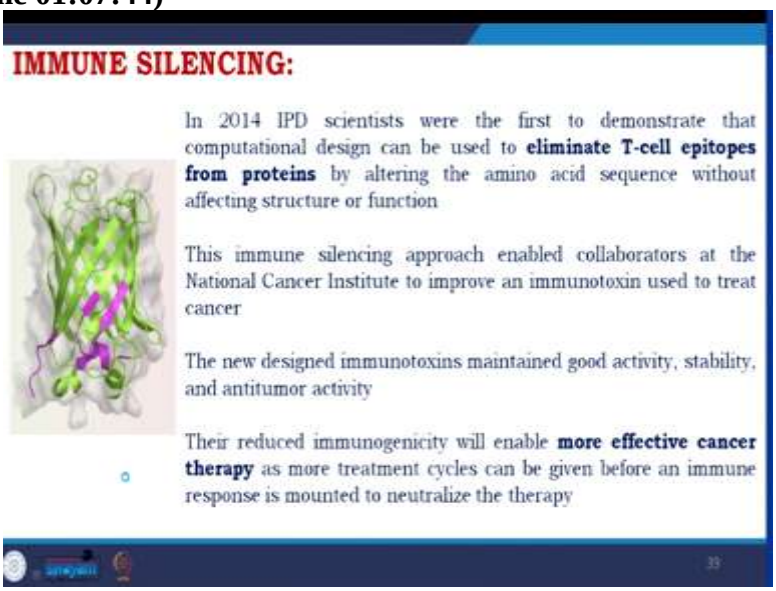
In 2011 IPD designed synthetic mini proteins that target conserved surfaces on influenza hemagglutinin from the 1918 H1N1 pandemic flu virus. These durable anti viral proteins protect rodents from exposure to lethal amounts of the virus. Researchers from IPD have also shown that these same flu binders can be used as diagnostic reagent in low cost diagnostic test strips providing improved performance as an influenza assay compared to a traditional antibody based capture system. These sensitive regions can reliably detect less than 100 influenza virus particles on a single nasal swab. In 2020, they used this same platform technology to create potent antivirals for the pandemic coronavirus and adapted them into modular and sensitive biosensors.

(Refer Slide Time 01:05:53)



Celiac Disease, in 2011, a team of University of Washington undergraduates working out of the IPD on the grand prize at the annual iGEM competition, sought to develop a cure for celiac disease and used computational design to be engineer in natural enzyme to break down gluten in the harsh acidic condition of the stomach. In 2015 IPD team published the most advanced version of the enzyme dubbed Kuma030, which cleaves gluten molecule at site that are known to cause an immune reaction in those with celiac disease. The enzyme was further developed by IPD spinout company PvP Biologics which has been acquired by Takeda.

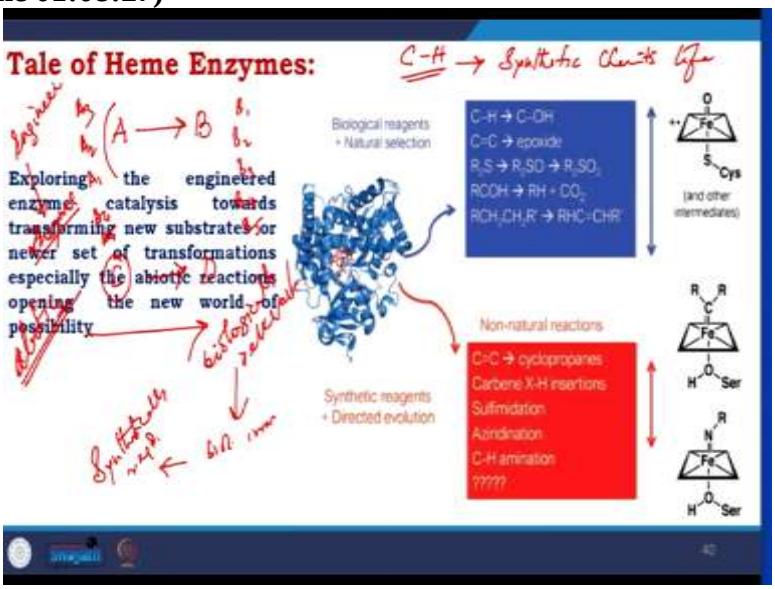
(Refer Slide Time 01:07:44)



Immune silencing, in 2014 IPD scientists were the first to demonstrate that computational design can be used to eliminate T-cell epitopes from proteins by altering the amino acid sequence without affecting structure or function. The immune silencing approach enabled collaborators at the National Cancer Institute to improve an immunotoxin used to treat cancer. The new designed immunotoxin maintained good activity stability and antitumor

activity. Their reduced immunogenicity will enable more effective cancer therapy as more treatment cycles can be given before an immune response is mounted to neutralize the therapy.

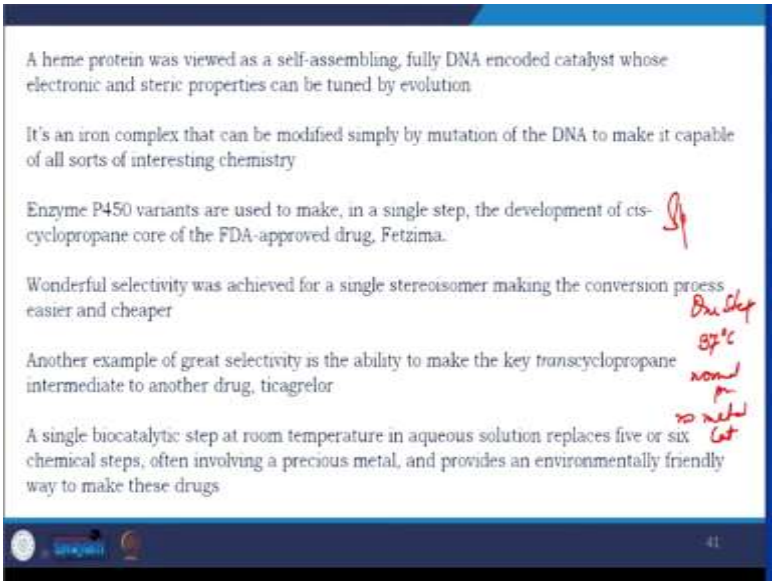
(Refer Slide Time 01:08:27)



Once it was successful, then people were even more imaginative they want to bring a totally different substrate and want to make the transformation taking heme enzymes specially, which are monooxygenase, work on CH bond, one of the most stable bond and one of synthetic chemists life. But if you have heme enzyme, they are very specific.

So, exploring the engineered enzyme catalyst towards transforming new substrates or newer set of transformation is precisely the abiotic reaction opening the new world of possibility.

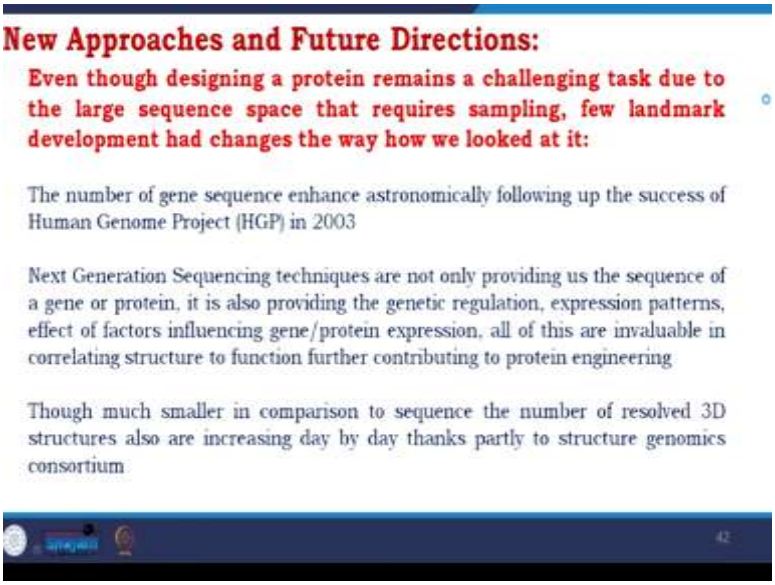
(Refer Slide Time 01:11:37)



A heme protein was viewed as a self-assembling, fully DNA encoded catalyst that's electronic and steric properties can be tuned by the help of evolution. It is an iron complex that can be modified simply by mutation of the DNA to make it capable of all sorts of very interesting chemistry. Enzyme P450 variants are used to make, in a single step, the development of cis cyclopropane which is the core of the approved drug, Fetzima.

Wonderful selectivity was achieved for a single stereoisomer making the conversion process easier and more importantly cheaper. Another example of great selectivity is the ability to make the key transcyclopropane intermediate to another drug, ticagrelor. A single biocatalytic stable at room temperature in aqueous solution replace 5 or 6 chemicals steps, often involving a precious metal and provides an environmentally friendly way to make this drugs.

(Refer Slide Time 01:13:25)



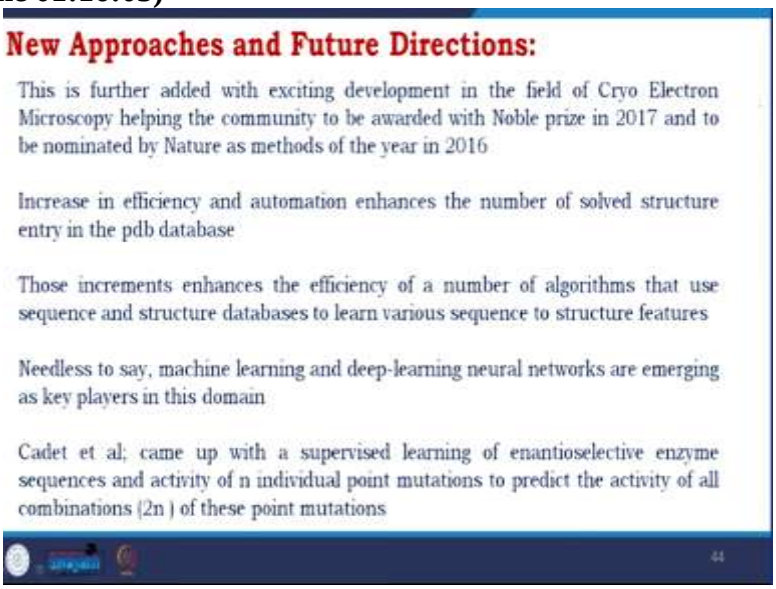
Coming to new approaches and future directions, even though designing a protein remains a challenging task due to the large sequence space that requires sampling, few landmark development changes the way how we looked at it. The numbers of gene sequence enhance astronomically following up the success of Human Genome Project in 2003. Next generation sequencing techniques are not only providing us the sequence of a gene or protein, it also is providing the genetic regulation, expression patterns, effect of factor influencing gene protein expression. All of these are invaluable in correlating structure to function further contributing to protein design. Though much smaller in comparison to sequence the number of results 3D structure also are increasing day by day, thanks partly to structural genomics consortium.

(Refer Slide Time 01:14:24)

Apparently looking at the available data, it is not uncommon to consider that I already have talked about structural genomics. So, it is we all even as a structural biologist, we realize that or we feel that structural genomics project did not go and achieve the success the Human Genome Project have achieved but in reality, this initiative have provided us with many things which help new generation protein crystallographers to work with much larger scale.

To name few, new crystallization screens, automated solution makers, crystallization robots, automated crystal checker, micro fluidics and nano fluidics based, crystallize and setups, have also contributed hugely.

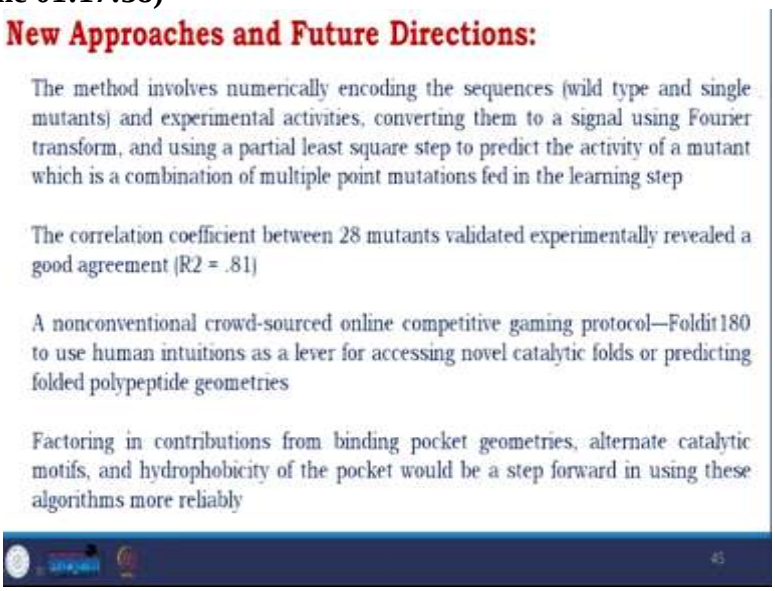
(Refer Slide Time 01:16:03)



This is further added with exciting development in the field of cryo electron microscopy, helping the community to be awarded the Nobel Prize in 2017 and to be nominated by nature as method of the Year in 2016. Cryo electron microscopy has travelled a lot which I have told in the module of electron microscopy. If you see among the structure solution techniques, this

is the most upcoming one, because it has ability to get or solve the structure in more native state. It could work with bigger sites complexes, especially the complex interactions and further protein designing. Increase in efficiency and automation enhances the number of solved structure entry in the PDB database. Those increments enhance the efficiency of a number of algorithms that use sequence and structure database to learn various sequences to structure features. Needless to say, machine learning and deep learning neural networks are emerging as key player in this domain. Cadet et al, came up with a supervised learning of enantioselective enzyme sequences, an activity of individual point mutation to predict the activity of all combinations (2n) of this point mutation.

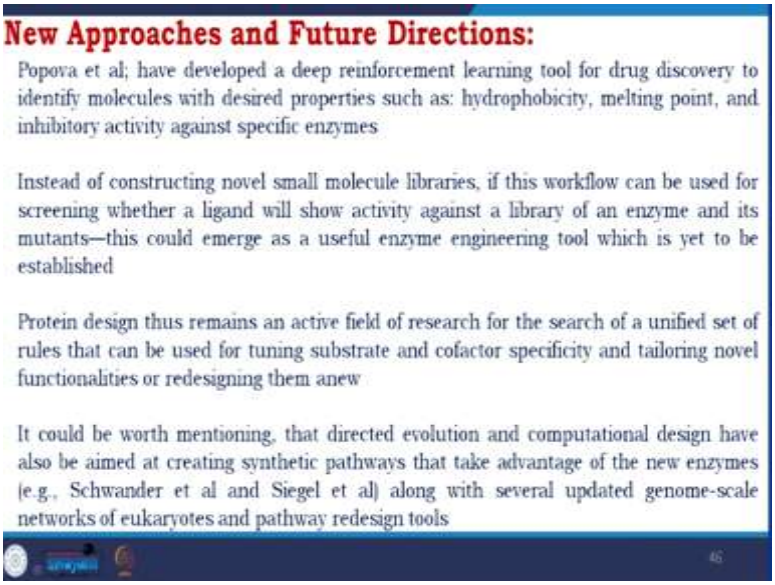
(Refer Slide Time 01:17:38)



The method involves numerically encoding the sequences wild type and single mutants and experimental activities converting them to signal using Fourier transform and using a partial least squares tape to predict the activity of a mutant which is a combination of multiple point mutation fed in the learning step. The correlation coefficient between 28 mutants validated experimentally revealed a good agreement. The non-conventional crowdsourced outline competitive gaming protocol Foldit 180 to use human intuition as a lever for accessing novel catalytic force or predicting folded polypeptide geometry.

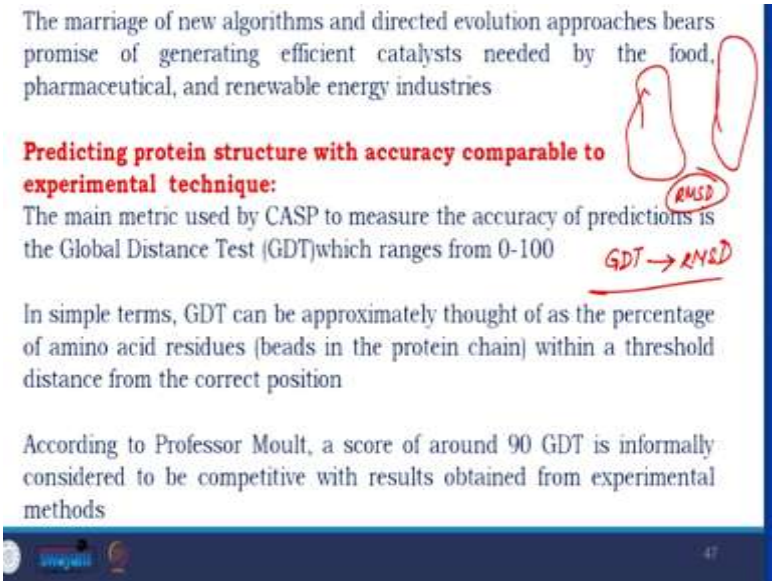
Factoring in contribution from binding pocket geometries, alternate catalytic motifs and hydrophobicity of the pocket would be a step forward in using these algorithms more reliably.

(Refer Slide Time 01:19:06)



Popova at all have developed a deep reinforcement learning tool for drug discovery to identify molecules with desired properties, such as hydrophobicity melting point and inhibitory activity against specific enzymes. Instead of constructing novel small molecule libraries, this workflow can be used for screening whether a ligand will show activity against the library of an enzyme and its mutant this could emerge as a useful enzyme engineering tool, which is yet to be established. Protein design thus remains an active field of research for the search of a unified set of rules that can be used for tuning substrate and cofactors specificity and tailoring novel functionalities or redesigning them a new. It could also be worth mentioning that directed evolution and computational design have also be aimed at creating synthetic pathways that take advantage of the new enzymes along with several updated genome scale networks of eukaryotes and pathway redesign tools. So, increase initially where we are working on enzyme or protein. Now we are working on enzyme pathways, metabolic pathway engineering.

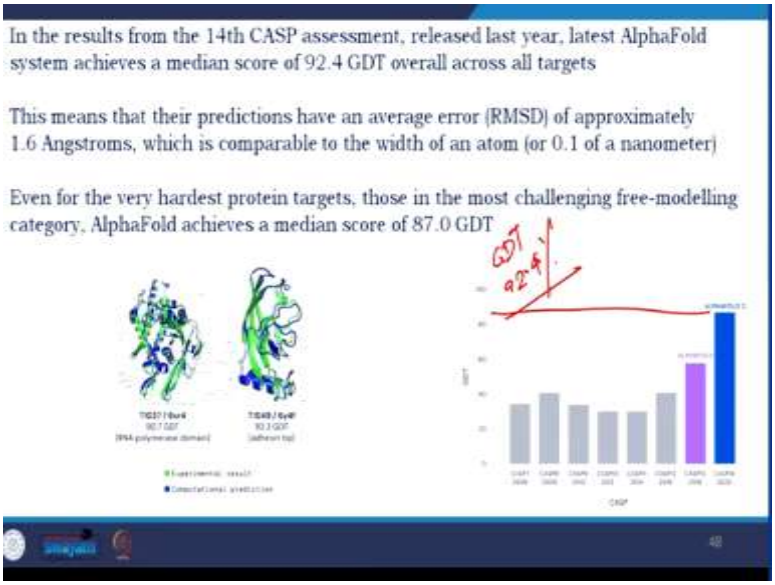
(Refer Slide Time 01:20:30)



The marriage of new algorithms and directed evolution approaches, we our promise of generating efficient catalysts needed by the food, pharmaceutical and renewable energy industries. Predicting protein structure with accuracy comparable to experimental techniques: The main metric used by CASP to measure the accuracy of prediction is Global Distance Test or GDT, which ranges from 0 to 100.

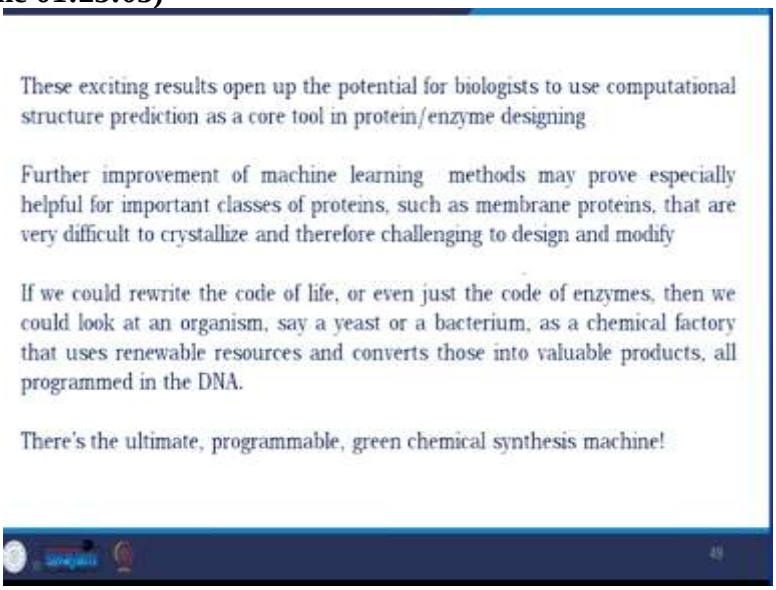
In simple terms, GDT can be approximately thought of as the percentage of amino acid residues beads with the within a solid distance from the correct position. According to Professor Moult is a score of around 90 GDT is informally considered to be competitive with results obtained from experimental methods.

(Refer Slide Time 01:22: 09)



In the results from the 14th CASP assessment, released last year in 2020, latest AlphaFold system achieves a median score of 92.4 GDT, overall, across all targets. This is amazing, because first time a program so that their continuous prediction is comparable to experiment, and it is not for one structure, it is the continuous performance that is why it is very important. This means that the prediction of an average error of approximately 1.6 angstrom, which is comparable to width of an atom. Even for the very hardest protein targets, those in the most challenging free modelling category, AlphaFold achieves, a median score of 87 with a very difficult targets.

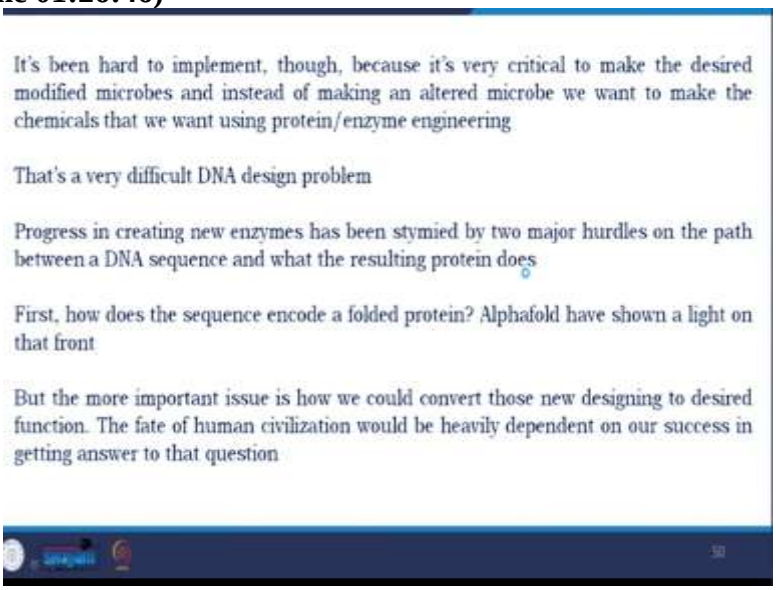
(Refer Slide Time 01:25:05)



But it is just opening up the potential for biologics to use computational structure prediction as a core tool for protein or enzyme designing. Further improvement of machine learning methods may prove especially helpful for important classes of proteins such as membrane proteins that are very difficult to crystallize and therefore challenging to design and modify.

If we could rewrite the code of life, or even just the code of enzymes then we could look at an organism, say yeast or a bacterium as a chemical factory that uses renewable resources and convert those valuable products, all programmed in the DNA. There is the ultimate programmable green chemical synthesis machine. There is no waste, especially no harmful waste every waste will be converted to valuable product.

(Refer Slide Time 01:26:46)



It is been hard to implement though, because it is very critical to make the desired modified microbes if you want to change your entire microbe, because nature have developed a lot of restriction show that it is difficult to modify the entire them. And instead of making an altered microbe, is your initial step that we want to make the chemicals that we want using protein or enzyme. So, instead of making a modified microbe, let us dream about making the enzyme modified enzyme or modified protein. That is a very difficult DNA design problem. Progress in creating new enzyme has been stymied by 2 major hurdles on the path between DNA sequence and what the resulting protein does. The first one, how does the sequence encode a folded protein? Fortunately, standing there in 2021, AlphaFold of shown a light on that front and that brings us to the more important second question, the more important issue is how we could convert those new designing to desired function. The fate of human civilization would be heavily dependent on our success in getting answer to that question. So, with that, I hope you understand the importance of protein designing how the entire learning of this course is coming and applied to this. And that is the beauty of basic science to application protein engineering, enzyme engineering, given us a platform to utilize the knowledge.

We gather through our PhD postdoc training, learning molecular biology, learning how to clone how to synthesize how to understand gene, how to work with the protein, how to characterize the protein, how to make mutants and all these things. And then how to solve a high resolution structure, how to use a lot of techniques, spectroscopy techniques, microscopy techniques, chromatographic techniques and all to get a pure protein and make the characterization and the training of functional acids all of them. Come together to develop or to help us working on a applicative area like protein engineering or something which is structure based drug designing, which I will talk about in the next module. Thank you very much for listening. Keep listening and keep asking question if you have any doubts. Thank you very much.