This section features conversations with personalities related to science, highlighting the factors and circumstances that guided them in making the career choice to be a scientist.

In-vitro to *In-silico* – How Computers are Arming Biology!

Andrej Sali talks to Geetha Sugumaran and Sushila Rajagopal

The developer of one of the first computational programs for comparative modelling of proteins, Andrej Sali is a researcher of many disciplines. With an early training in chemistry, he has forayed into biology, physics, and statistics to develop many open-source computational algorithms to understand biological systems. Graduating from the University of Ljubljana, Slovenia, Andrej Sali has been mentored by Sir Tom Blundell and Nobel Laureate Martin Karplus – pioneers in the field of cheminformatics and protein structure. Sali has worked as an Assistant Professor and then as an Associate Professor at the Rockefeller University. In 2003, he moved to the University of California, San Francisco (UCSF), as a Professor of computational biology in the Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences (QB3 – where he has also held the Directorship). Andrej Sali is as much an entrepreneur as he is a researcher and has founded Prospect Genomix (that merged with Structural Genomix, finally acquired by E Lilly and Co.) and Global Blood Therapeutics. Presently, his group at UCSF works towards developing integrated computational methods for determining and modulating structures and functions of proteins and their assemblies.

What follows is an interview with Prof. Andrej Sali (during his recent visit to India as the Jubilee Professor, Indian Academy of Sciences), preluded by a short introduction to protein modelling.

1. Biological Modelling – What and Why?

The evolution of biological sciences over the past decades has been a fascinating process. Early biologists tried to understand life through meticulous observation of the natural phenomenon and creative experimentation. The success in these cases many times depended on serendipi-

tous findings. But with the advances in technology, the approach is no longer the same. Today, various specialized branches of science have merged to look at the vast complexity of life in a transdisciplinary manner. It is also true that as the amount of knowledge has increased, newer technologies that aim at deepening the understanding of biological systems have also emerged. In this context, last few decades have seen unprecedented progress in the development of novel computational methods to understand the structure and functions of biological molecules. While biologists are still required to validate the computer-generated models, by subjecting the samples to carefully designed experiments like X-ray crystallography or NMR, today, there are numerous algorithms and software that aid experimental structure determination by predicting the likely structure and functions of a biological molecule saving much time, efforts, and resources. What is more, with the increase in computational power, such modelling is not restricted today to isolated biological models. Investigators are no longer limited to the study of small pieces of the outsized puzzle called life. Robust algorithms and biological databases make it possible to visualize the assembly, structures, and, functions of large biological complexes and even pathways on screen. With Human Genome Project (1990–2003) and the consequent technologies churning out more data than ever, the science of biological modelling is making biology more feasible from the perspectives of both time and cost.

2. Protein Modelling – Strategies

While nucleic acids (DNA and RNA) has been of utmost interest to the life sciences community, recent decades have also seen a surging interest in proteins. As it is said, "In the drama of life on a molecular scale, proteins are where the action is" [1]. With their complex and varied structures and highly specialized functions, protein structure and dynamics pose one of the biggest challenges to the scientific society. It is truly intriguing how mere 20 amino acids can combine to form various hierarchies of protein structure, each a little more complex than the previous (see *Box* 1).

Conventionally, proteins are studied individually. The coding sequence of a protein of interest is identified and cloned into a suitable expression vector. Once the lengthy processes of cloning, expression, and purification are successful, and enough quantities of pure proteins harvested, these proteins are employed in experiments or used to prepare solutions for NMR spectroscopy or to grow crystals for structure determination by X-ray crystallography [2, 3]. However, with advances in both computational power and tools, *in-silico* approaches are aiding *in-vitro* instruments in determining the structure of biological molecules. The eventual goal of protein modelling is to predict the structure of a protein from its sequence with an accuracy that is comparable to the best results achieved experimentally. The rapidly generated *in-silico* protein models can then be effectively utilized in almost all contexts which requires knowledge

of the structure of proteins and also in cases where experimental techniques fail, for example,

Box 1. Protein Structure [2]

1) **Primary Structure**: The primary structure is the first level of structural organization in proteins and refers to the simple sequence of amino acids of each polypeptide chain in a protein.

2) **Secondary Structure**: Secondary structure refers to a 3-dimensional local confirmation formed in a protein chain due to hydrogen bonding. The two most common secondary structural elements are the alpha helices and beta sheets (others include – beta turns and omega loops).

3) **Tertiary Structure**: This is the stable and biologically active 3D structure of single-chain proteins. In this form, the proteins appear globular, and the individual elements of secondary structures are compactly packed against one another to form structures such as α-bundles, α-horseshoes, α-solenoids, β-rolls, β-barrels, β-prisms, α/β barrels, α/β sandwiches, etc.

4) **Quaternary Structure**: This is the highest level of organization in proteins and is assumed by proteins which are composed of multiple chains instead of a single polypeptide chain. Quaternary structure hence refers to the number, arrangement, and interactions of multiple folded protein chains in a multi-chain complex. This includes organizations from simple dimers (2 chains) to large oligomers (8 chains) and larger complexes.

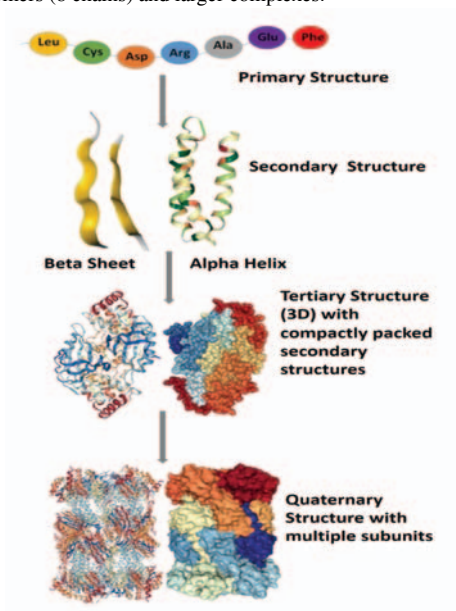

Figure A. The levels of structural organization in proteins.

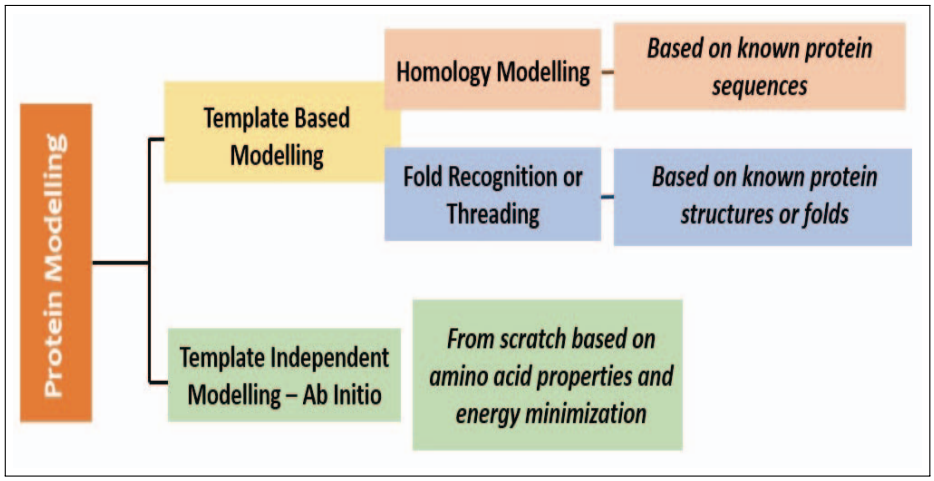

Figure 1. Various approaches to modelling.

proteins that are too large for NMR analysis and crystallization for X-ray diffraction [2].

The major approaches adopted to model and predict the 3D structure of proteins can be classified into template dependent (template based) and template independent approaches (*Figure* 1).

2.1 Template-Dependent or Comparative Modelling

Comparative protein modelling is based on using previously solved structures as reference templates. Hence, the known structural information is extrapolated to the unknown.

Say, we have the sequence of protein 'A' whose structure is unknown and needs to be determined. We might have an inkling that 'A' might be an essential protein and hence might have been conserved during evolution, albeit with minor changes in the sequence. In such a scenario, a simple search of the Protein Data Bank (PDB) using the sequence of our target protein 'A' as a query is enough to dig up homologous (evolutionarily close or related) proteins whose structures have been solved already. Here, it is important to note that the number of matching results depends on the protein comparison and alignment methods used. The target may be subjected to pairwise alignment (the query is aligned only to the best match in the database), where the search is done locally (segment by segment or residue by residue), or globally throughout the sequence using algorithms like BLAST and FASTA. Alternatively, one might perform a global, multiple sequence alignment of the query 'A' to dig out all the related sequences using tools such as PSI-BLAST or CLUSTAL-W. The information from a multiple sequence alignment

is often converted into a sequence-profile. A sequence-profile lists the preference for the 20 standard amino acids at each position in a given multiple sequence alignment using specialized scoring matrices. All the proteins in a family are expected to have same/similar profiles, and hence a sequence-profile can define a family of proteins.

With the template identified, the structure (Cartesian coordinates) can be downloaded from PDB, and modelling programs, e.g., MODELLER (see *Box* 2) can easily generate the possible structure of protein 'A' from the Cartesian coordinates of the template. The model can further be tweaked by mutating the amino acids (inserting or deleting residues) that differ between the target and template sequences.

This approach, based on the sequence-sequence alignment of related proteins is known as 'homology modelling' and is perhaps the easiest protein modelling method. As the name suggests, this approach refers to modelling the 3D structure of a protein based on the known experimental structure of a homologous protein as the template [4].

While this is the basic approach to protein modelling, various intricacies arise when there are major differences between the target, and the aligned sequences returned from PDB search or when it is not possible to construct sequence-profiles because there are not enough known sequences that are related to the target [5].

While similar sequences are seen to have similar structures and functions – the concept we exploit in homology modelling – Nature can throw up paradoxes. It has been observed that

Box 2. Homology Modelling using MODELLER

One of the earliest computational tools developed for homology modelling is MODELLER. Developed in the early 90s by Prof. Andrej Sali, while he was a doctorate student at Prof. Sir Tom Blundell's lab in the UK, MODELLER is written in standard FORTRAN 90. Decades later, MODELLER remains one of the most fundamental tools used by structural biologists aiming to predict the 3D structure of a protein.

MODELLER generates the possible 3D structures of proteins and their assemblies by satisfying spatial restraints. The program computes the possible model based on the alignment of the query sequence with its homologs whose structures are known. More generally, the program accepts restraints on the spatial structure of the amino acid sequence(s) and ligands to be modelled as input and generate a 3D structure that satisfies these restraints as well as possible. The restraints, in turn, are extracted from related protein structures available in PDB and their alignment with the target sequence. A complete 3D model is generated by MODELLER through the optimization of a molecular probability density function. MODELLER can also perform multiple comparisons of protein sequences and/or structures, clustering of proteins, and searching of sequence databases. To learn more about MODELLER see [8].

convergent evolution can drive completely unrelated proteins, with differing amino acid sequences to adopt the same fold [6]. Also, despite the vast number of proteins, there is a limited set of tertiary structural motifs (around 1300) to which most proteins belong. This scenario calls for a different approach called the 'protein threading' or 'fold recognition'.Protein threading is basically for target proteins which show only fold-level homology with other proteins and is based on the structure of similar and known folds, in place of homologous templates in PDB.

The idea here is to align the sequence of query 'A' to known protein structures. This is unlike homology modelling where the sequences are aligned first, and the protein is modelled based on the structure of the best-aligned sequences. This is a kind of reverse modelling wherein each amino acid in the target sequence is threaded (placed/aligned) through the structure of the template folds by matching it against a library of 3D profiles. An objective function scores the sequence-structure compatibility between the sequence of amino acids and their corresponding positions in a core template. Further, it deduces the possible spatial arrangement of amino acids and evaluates their interaction preferences, preferences for solvent accessibility, and residue contact potentials [7]. Hence, while the template in an alignment is treated as a sequence (with a structure) in homology modelling, the template is treated as a structure in protein threading, and both sequence and structure information, and statistical knowledge of the relationship between the structural folds extracted from the alignment are used for prediction.

In a nutshell, presently available comparative modelling approaches comprise sequential steps as follows [5]:

i. Search for proteins with known 3D structures which are related to the target sequence
ii. Select the template sequence/structures
iii. Align the template sequences/structures with the target
iv. Generate a 3D model for the target sequence based on the alignment
v. Evaluate the model

2.2 Template-Independent or *Ab Initio* Modelling

Ab initio translates to 'from the beginning' (also known as *de novo* modelling). As the name suggests, the approach seeks to predict the structure of a protein from scratch, based on its amino acid sequence and physical properties rather than on previously solved protein structures. This is particularly applicable in case of newly discovered and unique proteins with no homologous with known 3D structure information [4].

Ab initio methods try to arrive at the native structure of a protein by simulating the biological process of protein folding. In the process, the program generates multiple conformations iteratively and estimates the corresponding changes in energy using contact potentials. The aim is to arrive at the confirmation which has the lowest free energy, leveraging on the principle that native state of a protein has the lowest free energy minima in the vast conformational space [4] [6].

3. Protein Modelling – Applications

With massive improvements in technology has come the era of data-deluge. The enormous amount of genomic and proteomic data generated offers tremendous scope for furthering the understanding of life and diseases. While the structure of a protein can guide the design of new experiments, for example, site-directed mutagenesis, structure-based drug discovery, and rational drug design, structural information of proteins becomes unavoidable when it comes to furthering studies on biological pathways and networks, dynamics, and interactions. In this context, protein modelling has emerged as a powerful technique with the potential to help answer many crucial questions on life.

GS: To start with, could you please tell us a little about your background and how you got interested in using computational approaches to solve biological problems.

Prof. Andrej Sali

AS: Right, so how early in my life do you want me to start (laughs)? I was interested in science since my primary school, maybe by 3rd or 4th grade. I don't know why, but I liked to read textbooks. In particular, physics and chemistry were attractive to me as subjects. I also had some chemistry kits that my father got me when I was in 4th or 5th grade. This was to do some small experiments – to make different colors or blow up something (laughs), that sort of thing, nothing special. Then in 7th grade or so, I got a calculator – HP 41C – I think it was called, which was programmable. It had a very small memory, I think a 150-byte memory card, but you could program it. So, you could solve simple mathematical problems like quadratic equations, if you programmed it. I loved that, and I loved the combination of programming, physics, and chemistry. Then in my early high school, I went to the local institute – the Jozef Stefan Institute in Ljubljana,

the capital of Slovenia. Jozef Stefan was a physicist who along with Boltzmann discovered the Stefan–Boltzmann law that quantifies radiation of black objects as a function of temperature. There is a biochemistry department at the institute, and I had framed a project proposal to them. I wanted to measure the degree of water pollution by measuring the ability of water to conduct current after some reactions that depended on the degree of oxygen in the water, which in turn depended on the level of water pollution. I forget the details. It had absolutely nothing to do with what they were doing in the department (laughs), and of course, they laughed at me (laughs). But my first major stroke of luck was that Professor Vito Turk, the Head of the Department, proposed that I work on something else that was of interest to them. Hence, I was assigned a mentor who was a PhD student in the department, and I spent quite a lot of time in that lab doing some biochemistry – actually experimental biochemistry and not computational. Computers in the meantime were getting better and better. At the end of my high school studies, I managed to buy an Apple II+ microcomputer. We are talking about the early 80s, 1981 or so. So, I actually purchased an Apple II+ computer privately, with a lot of help from my parents and one of my aunts (laughs).

Having a personal computer was amazing. You could program with a high-level programing language, and I started writing more complex programs – for example, fitting enzyme kinetics models to data of different kinds. modelling, broadly speaking, is still what I am doing. Then, when it was time to go to the university – University of Ljubljana, Slovenia – I went on to do chemistry. This was because Prof. Turk (I mentioned earlier) wanted me to do chemistry and not physics. I am still not sure if I did the right thing (laughs). But yes, I went on to do chemistry. Moreover, University of Ljubljana was a very small university, and there was no specialization. So, I learned some generic chemistry – organic and inorganic. But I also continued to work at Jozef Stefan Institute, and I started publishing papers. I did more and more computational work and less and less experimental work, and I convinced Prof. Turk to buy several additional computers. It was at this point that the second major stroke of luck occurred. There was a big international biochemistry conference in Slovenia. Prof. Turk was one of the organizers. He introduced me to a very senior scientist from England who was the plenary speaker at the conference. His name was Tom Blundell. So, one thing led to another, and I visited Tom's lab in London, England, first for two weeks at a time and then two months at a time on a British scholarship as an undergraduate student. By the time I graduated from the university in 3 years, I was all set up to go for my PhD in London. It was possible because Tom supported me, fortunately.

With Tom, I did a little bit of X-ray crystallography in the beginning, but then I mostly worked on what is called structural bioinformatics now. This is an approach to study proteins inspired more by statistics than physics. So, I designed algorithms and programs, wrote papers, and had

a great time with Tom who is an excellent mentor. After I received my PhD, I decided to do a postdoc, but not in a totally new area that would require me to start from scratch, without benefiting from what I had done in my PhD years. I also did not want to do more of the same thing, but extend to something different that still benefited from my past experience. I felt that I should keep looking at proteins, which I was interested in, but now from a different perspective. Because I was taking a statistics-based view of protein sequence-structure relationship during my PhD studies, I now wanted to gain a more physical, statistical mechanical perspective on proteins.

So, I applied to Martin Karplus, an expert in that kind of studies at Harvard University and was very lucky again to succeed in joining his research group on the Jane Coffin Childs Memorial Fellowship – which I would say was a lottery. Fellowships are always a lottery. I am not sure I would have been able to join him if I didn't have the fellowship and therefore a major step in my trajectory depended on a lottery. And I say that because I think young people should know that sometimes there is a lot of chance involved, and in order to eventually succeed, you have to keep trying. So if you fail first, you have to try and try again. Anyway, I joined Martin's lab, and I spent three years there working on statistical mechanics of protein folding using computer simulations.

One day Martin said, "Would you be interested in being an Assistant Professor at Rockefeller University in New York City." He said it was a big city and not everyone loved it and all that. And I said, "Oh sure! Why not, that is great." So I visited the University, and soon after that in 1995, I joined there as an Assistant Professor. I continued to work on merging the statistical and physical aspects of protein structure as a group leader. I spent eight years at Rockefeller, building a research team of students and postdocs, writing grant proposals and papers, collaborating, and attending conferences; one of those conferences was here in Bangalore in 1999 (smiles). That was when Madhu [1] and I met. Madhu was a student at that time, just like I was when I met Tom Blundell. Soon after that, he joined our group as a postdoc. Following that, Madhu managed to come back as an Assistant Professor at IISER Pune (laughs). So, there is this cycle of scientists renewing themselves, so to speak. After eight years at Rockefeller, there was an opportunity to move to University of California, San Francisco. The QB3 Institute at UCSF was very much reflecting our goals and the type of science I wanted to do more. There was a huge mass of structural and computational biologists at the institute, and there were a lot of students and postdocs interested in these areas, more so than at Rockefeller, which is a smaller institute and focused more on cell biology and not so much on computational biology. So there was an opportunity to collaborate locally more richly. The group moved to UCSF, and we continued to model proteins, trying to understand their functions, evolutionary processes

[1] Prof. M S Madhusudhan is Associate Professor of bioinformatics and structural biology at IISER, Pune.

that modulate these functions, etc., but hopefully in a different way, maybe in some respect in better ways. As a result, after a number of years, I find myself here with you asking me the question about how I got here. I probably gave you a slightly longer answer than you wanted (laughs).

GS: Life sciences in the past few decades have undergone a major transformation, and there has been a move towards more computational biology and less of experimental biology. What is your take on this trend?

AS: I don't think any of the big problems in biology will be solved using computational approaches alone. I think that most of the big questions have been and will continue to be solved by a combination of theoretical, including computational, and experimental approaches. So, broadly speaking, in science, you collect data doing some experiments and then use this data to guess or sometimes compute a model of the system to which the data apply. Then with that model in hand, you do more experiments. So you need to do both experimenting and modelling in order to solve any significant question.

GS: What are some of these most important biological problems which might be solved using integrated computational approaches?

AS: I think, maybe slightly myopically, that one big question in biology is the one we are interested in. I have to think that way because I guess everyone works on the problem that one considers the most important, right (laughs). We aim to describe in great detail – quantitatively and predictively – how the cell looks like, how it works, and how we can modulate its functions. Modelling a prokaryotic or a simple eukaryotic cell to understand how it functions is a daunting task in itself, and we need to understand human cells right. That is even more complex, and there is a great variety of them; I think there is something like 250 types of human cells. The cell is a very complicated entity compared to individual molecules or their complexes, which is what we can model relatively well at this point, provided we have sufficient information. We want to be able to describe how the cell is structured in terms of its atoms, in terms of its molecules, networks, complexes, organelles, pathways, compartments, and mass and energy flows. We want to develop a spatial model, and we want a model that changes in time because life occurs in time. We also want a model that would allow us to turn which ever knob we want and change the behavior of cells, modulate the cells using small molecules and macromolecules. A major purpose of our ability to modulate cellular functions is to cure diseases in a very deliberate, planned, rational, efficient, and precise fashion. The model would increase the chances of success in the discovery of new drugs compared to the current trial and error method. This is what we want. We don't have such a model right now; it is not even clear in detail what this model would look like. But I believe that the experimental technologies for

collecting information about different aspects of the cells have progressed much. Computers are also getting faster, with more memory. Likewise, computational tools are becoming better, and algorithms are improving. So it is time for modelers like us to engage on the modelling of the cell. It is time to get going on constructing a model of the cell for real, and there are clusters of scientists who are beginning to address this head-on in a serious fashion. It may take multiple granting cycles (1 granting cycle in the US means five years) before we have useful models, say 20 or more years from now. Obviously, there will be isolated successes in some aspects and partial successes if we are willing to reduce our demands enough. But I think it is a long-term project.

GS: So biological structures, pathways, and even networks are being modelled, and you have developed an Integrative Modeling Platform (IMP) (see *Box* 3) to model eukaryotic ribosomes, ryanodine receptor channels, proteasomes, chaperonins, etc. You also mentioned about modelling entire organisms. Can you tell us about these efforts?

AS: As of now, I don't think there is any single method that can describe most complexes, such as those you listed, let alone the cell at the required level of detail. Maybe there will be one such method in the future, but as of now, models of biological systems are best built based on information from different experimental methods, at different scales of resolution. All of these data sets have to be integrated computationally into a model of a system, to understand how it works.

GS: Could you briefly introduce our readers to the concept of modelling 3D structure of proteins?

AS: Yeah (smiles). Structures of proteins can be modelled at different levels of precision. Most commonly, we define the Cartesian coordinates of the constituent atoms, i.e., the position of each atom of the protein in space. You can do that based on some information about the struc-

Box 3. Integrative Modelling Platform

Today, the focus has shifted from modelling individual, isolated proteins to taking a holistic look at biological systems using integrated models of protein complexes and pathways; so much so that the study has emerged as a new discipline called systems biology. The crux is to generate models based on multiple sources of data (various biochemical and biophysical experiments), rather than relying on a single source. The resulting models are likely to be more accurate and efficient. In this context, Sali lab has developed an Integrative Modelling Platform package (IMP) [9]. Such integrated approaches are and will continue to impart crucial information on the organization, function, and evolution of complex biological systems and give us an upper hand on how to modulate them to our advantage.

ture. Different experiments produce different kinds of information. Some of them, like NMR spectroscopy, produce distances between atoms while electron microscopy determines an approximation of electron density in space, and X-ray crystallography gives diffraction patterns of many copies of the protein in a crystal. In all of these cases, you next compute a model of the structure satisfying the experimental information. All methods try to find those protein models whose computed properties match the measured properties as much as possible, and if possible within the experimental uncertainties. So protein modelling, based on either theoretical or experimental information, always follows a common scheme that relies on finding the positions of individual atoms of a protein that reproduce what we know about it. Therefore, it is reasonable to take information from multiple sources and come up with the most accurate, precise, and complete model satisfying all the available information, not just one type of it. This is what we try to do. In addition, we are not interested in studying only the 3D structures of single proteins but also of complexes of proteins and then the processes in which these complexes are involved. This could be just assembling and disassembling of the complexes or their transport from one location to another, or how they are regulated by changes in the structure and dynamics. So, we are interested in the spatio-temporal dynamics of proteins and their complexes. We are interested in building the models of the various biomolecular processes by considering all the information from different experiments, statistical analyses, and physical theories.

GS: So the accuracy of a model depends on the quality of the data?

AS: Yes. In fact, it depends on two things, so as to say. It depends on the quality of data, and it depends on the quality of the modelling process. And one can, of course, discuss each aspect at great length, but this is the short answer to this question.

GS: What are the general processes involved, requirements, and challenges of such an integrative structure modelling approach?

AS: I will give you a semi-technical answer. So, there are three aspects of any modelling approach, including integrative modelling of macromolecular structures. You first have to decide how you are representing the modelled system. For example, you have to decide whether you are using individual atoms or want to make it coarser by combining atoms to get coarser objects. These objects may be bigger spheres instead of small spheres, and maybe they are not even spheres; maybe they are cylinders or ellipsoids. There is an uncertainty in general as to what is a good representation of the modelled system. This is the first aspect and in the difficult cases, perhaps the main challenge.

The scoring function is the second aspect of computational modelling. That is, once you gather the information for modelling, it needs to be converted into a scoring function whose only

purpose is to rank alternate models based on the input information. There is no other way around it. You have to have a ready way to rank alternate models. Exactly how you construct the scoring function, how you convert the raw data into a scoring function to rank the models, is the requirement and challenge. This process should consider the uncertainties in the data and our understanding of the relationship between the data and the structure, and also the computational burden.

The third aspect is generating or sampling the alternate models. We now have a scoring function to rank the models, but we also need a program to suggest alternate models so that the scoring function can rank them. For complex biological systems with many degrees of freedom, we have a lot of possible models, many of which are bad and maybe only one of them is a good model. Maybe a few are good models, but you need to generate them somehow so that the scoring function can recognize them as good models and rank them accordingly. This is computationally intense and requires robust algorithms. Additional developments benefiting from the increased computational power of modern computers are needed.

Once you have all these three aspects of modelling covered, the only part left is to apply them and analyze the results. The analysis includes validating the models in different ways, visualizing them, annotating, drawing conclusions, archiving, disseminating, and publishing. Certainly, there are challenges in these aspects as well.

GS: What are the advantages of this kind of a computational approach over directly moving into experiments?

AS: I would guess that almost no scientist, especially these days, is just an experimentalist. Every scientist does some modelling; they make some predictive interpretation of the data. No one just measures the data looking through the microscope, prepares a table of numbers, and publishes a paper about it. That I think is very rare in today's world. There is always an interpretation of the data – this is what they mean, this is a scheme implied by the data, here is a proposal based on the data, here is a hypothesis, etc. In my view, as soon as you do this, you are modelling. You may perhaps be doing that in your head. It may not always be done using computers, but you are necessarily always modelling your data to generate a picture of how things are or how things work. Modelling can guide and thus reduce the experimental effort and provide a sense of direction to experimentalists.

GS: MODELLER developed by you is one of the fundamental tools used in bioinformatics and structural biology. The recent version MODELLER 9.19, was released on 25 July 2017. What are the updates on this?

AS: This is a sensitive one (laughs). The first version of MODELLER was created when I was

a PhD student with Tom Blundell, at the end of my first year there in 1988. That was a long, very long time ago. But the basic idea of how to solve the problem of homology modelling was already implemented then. The tool has changed over the years. We have improved it in incremental ways and fixed some bugs. Other advances helped as well, but not because of anything special we did. Instead, the database of known structures, which MODELLER relies on, has grown due to the efforts of others. Also, as we discussed, sampling is one of the key aspects of modelling, and when you have faster computers, as we now do, you can generate a larger number of alternate models, and you often do better with the same modelling method.

So MODELLER program was more or less smoothly improving on this basis, through some of my efforts and later based on the efforts of our research group at Rockefeller University. However, in 2007, it occurred to me that we should start another modelling program for integrative structure modelling of biological macromolecules, and I submitted a proposal to NIH. NIH liked the proposal, but they didn't agree to fund both the new proposal and the MODELLER grants. They gave me a choice, and I took the grant for the new program – for integrative structure modelling. As a result, since 2007, there has not been any significant developments in MODELLER, just because we don't have the funding anymore.

That said, I also have to admit that it could be that we have almost hit the glass ceiling in comparative modelling. This is corresponding to the accuracy of comparative models you get by simply mimicking the structure of the template used to construct the model while relaxing the model a little bit so that the chemistry of the model is what is expected, and it fits the physical potentials and the molecular mechanics force field. Now, that level of comparative modelling was already reached in the mid-90s, and it has been very difficult for anyone to go beyond it. So, maybe even if we had the money to continue working on MODELLER beyond 2007, we would still be hitting this glass ceiling. Maybe NIH did a favor by saving me a lot of time.

GS: Your lab has developed a hybrid approach to model biological pathways. How has this been accomplished?

AS: You could define integrative structure modelling as a structural modelling of either a simple protein or a complex of proteins (see *Figure* 2) that takes information of different types, from different methods, different physical theories, and different statistical analyses, so that you construct a model consistent with all of that information, and not just one type of information at a time. That is why it is called integrative modelling. And because you are using all information, if you don't misuse it, you generally get a more accurate, precise, and complete model compared to using only a subset of information. That is the motivation behind integrative approaches. They can also be applied to other kinds of modelling, including the modelling

of biological pathways, which we began to do recently.

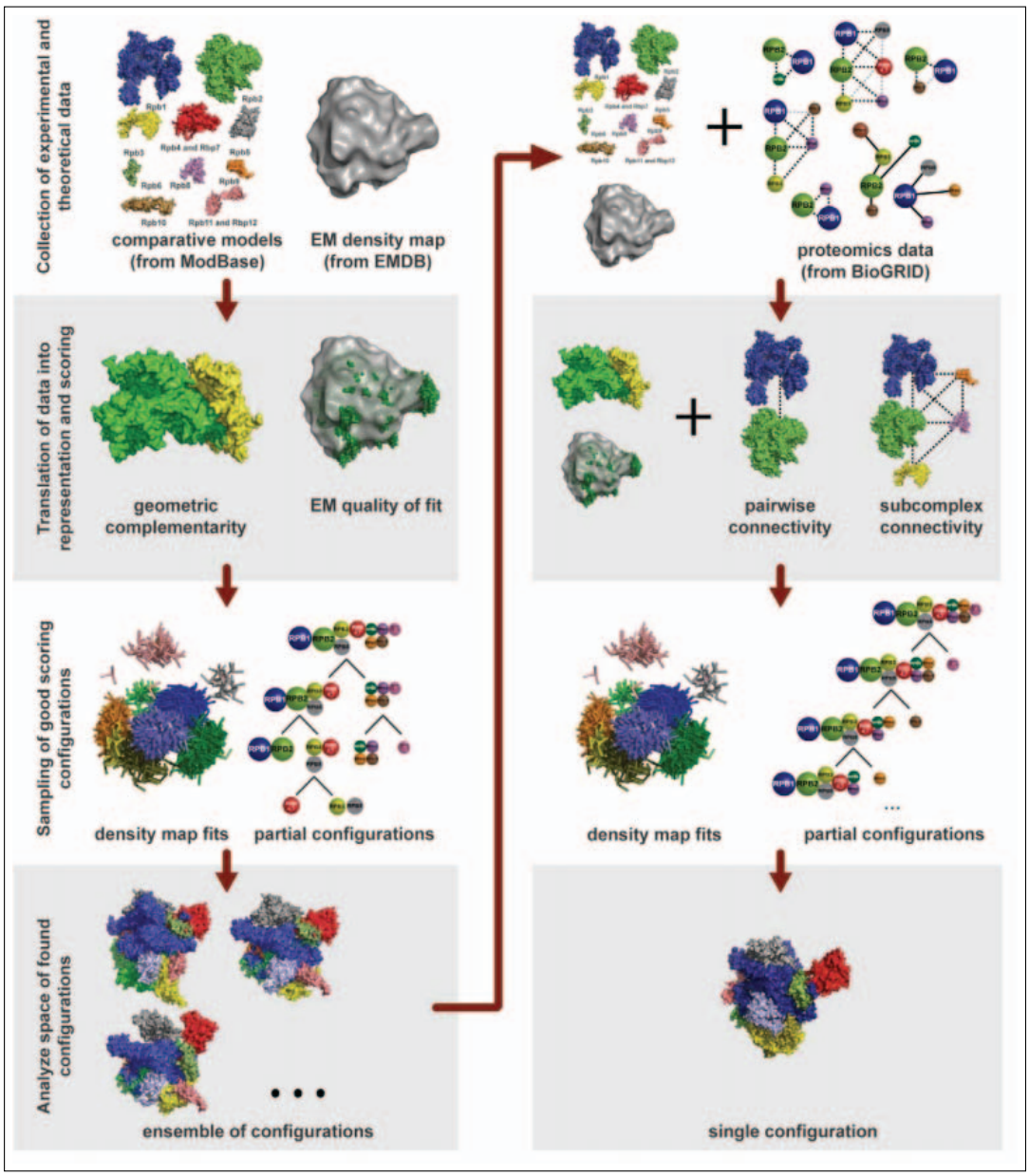

Figure 2. Integrative structure modelling of Human RNA Polymerase II (Source: A Sali *et al*., *PloS Biology*, Vol. 10, Issue 1, 2012) [10].

GS: There has been an explosion of biological data in the post-genomic era. Do you think that the available computational tools suffice to match the data being churned out?

AS: Oh no! Absolutely not! There is always a time lag. I think it is a feature of biological research that most people involved in biology are experimentalists. Maybe you can even say that most of the big discoveries are attributed to new experiments as opposed to new models, with a few exceptions such as the double helix of DNA. Most of the granting agencies and most of the papers generally insist on experiments first. So, biology is generating large amounts of data, but unfortunately, there is a lag in developing the tools required for analyzing this data. This is because of the sociology of doing science (smiles). Moreover, how can you develop new computational tools unless you have the data first? So naturally, there is going to be some time lag between the data and the tools. But unfortunately, I think that the time lag is often longer than it has to be, and maybe we should work ahead of time finding and developing computational support a little more proactively. A very good example is the Human Genome Sequencing Project. All those involved in the genome sequencing project – NIH, Francis Collins, Craig Venter – they were all doing great, but then initially they didn't have the databases, they didn't have the annotation software, the sequence comparison tools, etc., to interpret the genomes they successfully sequenced. It took quite a while to develop sufficient infrastructure on the computational side. But in the end, it happened. And all is well that ends well.

GS: What is the scope for aspiring students in this area?

AS: The sky is the limit (smiles). We kind of come back to what we discussed previously. Absolutely everything that matters is getting better. There is more data, more variety of data, there are faster computers with a lot more memory, and the world is more connected and collaborative. Now, maybe it is fair to say, sometimes there are very long cycles in science. For example, in biology, Lamarck systematized many organisms, describing their differences and similarities, but there was no understanding of how evolution worked. However, enough data was accumulated from around the world, including by Darwin himself, that made it possible for him to come up with the theory of evolution. This is a common occurrence in all fields. Maybe I am just kidding myself, but, today, we have so much new data and so many new technologies in biology that there are good conditions for young researchers to postulate new models and theories on how life has evolved, how it functions, and how it could be modulated and designed. The future is bright.

GS: Could you share one of your 'Eureka' moments?

AS: Yes, when I figured out that $E = mc^2$ (laughs). I am going to give you one, but then I am afraid it won't measure up to the Eureka moment of the kind you had in mind. Well,

one of my mini Eureka moments was when I was a PhD student and started my project by comparing protein structures. So when there are two proteins, they could be similar or quite different. Hence, it is often helpful to compare their amino acid sequences, residue by residue. That is, this residue in this protein is equivalent to that residue in the other protein, and so on. The output is an alignment of the residues comprising the two protein sequences. Alternately, one could also superpose the two compared structures. I realized that we could consider both types of representation at the same time, the sequences and the structures. Moreover, there was no reason why we couldn't add additional properties of sequences and structures, like their intramolecular interactions and local shapes of the polypeptide chains. So, you obtain a comparison by optimizing the similarity across all these different features and not just one at a time. You can see a little bit of the integrated approach already here. But that was not the Eureka. The Eureka moment came when I realized that if I have such a comparison in hand, I can then forget that I know one of the two structures available, and I can extrapolate structural restraints from the one I do know to the one I don't know; I can do that for many structural features, not just one. Then, I can get a structural model by satisfying the spatial restraints, whether they are residue distances, hydrogen bonds, or whatever. So that is how MODELLER was designed. Importantly, it immediately also led to integrative modelling. I didn't care exactly what those spatial restraints really were or how exactly they were obtained; all that mattered was whether or not they were informative. Even in 1989, towards the end of some of our papers, I realized that these restraints could, in principle, come from anywhere, and not just from related known structures as is the case for comparative modelling. Still, it actually took almost 15 years before we started working on integrative modelling in practice. So, that was probably one of the realizations about modelling that I am happy with. Maybe there were a few more, but I have to admit this is nothing compared to the discovery of the double helix of DNA. Can you imagine what kind of a 'click' that had to be, when Watson and Crick realized that such a simple model explains so many fundamental aspects of evolution and life.

SR: But don't you think Biology is past that stage when we can make any more of such fundamental discoveries?

AS: No! We don't know, right! We just don't know what we don't know. There must be more of such simple and powerful concepts out there, yet to be discovered. We need to motivate young people to engage in pursuing such discoveries. Sky is the limit.

GS: Could you share your experience working with your mentors Martin Karplus and Sir Tom Blundell, who are pioneers in their fields?

AS: Yeah. As I said, I was very lucky to work with all the three main mentors I had – as an

undergraduate student, as a PhD student, and as a postdoc. Unbelievable luck. My first mentor – Vito Turk, obviously, I would not have even started unless he got me going. He supported me immensely. About Tom, he is a very broad person. He can talk about a lot of different areas of science very well. He is, in fact, a broad-minded person and can talk to any type of person on any topic. He is also very down-to-earth, generous, and supportive of his students. He let me take MODELLER, which I had developed during my stay in his lab, and let me continue developing it on my own later. MODELLER was licensed to a company, and Tom didn't want any royalties. You know, normally he should get some. Instead, he said, "It is your thing. Go and do it." Tom has been extremely supportive, and I will always be very grateful to him. Martin Karplus, my postdoc mentor, is a premier scientist. He is on the top of the pyramid so to speak, and almost everyone in the field of protein structure has been touched by him, maybe as a student or as a postdoc, or as a student of a student. He got the Nobel Prize in 2012. Martin is an amazing teacher, teaching by example, and a great scientist with the ability to ask very good questions. In science, a lot depends on asking the right question first and then trying to answer it as opposed to just trying to answer a question. So, I think Martin excels in asking very good questions whose answers then have a lot of impact. All my mentors have retired, but they are all very much active, and I see them every so often. I am even collaborating with Tom now, which is especially pleasing.

GS: Biology today is no longer an isolated field. What are the essential skills that a student aspiring to enter the field of biology and in particular, computational biology must posses?

AS: Let me tell you. I think there is no specific profile. You could come to computational biology from a lot of different directions and be successful. You might have been a physicist, a computer scientist, a biologist, or a chemist or you might have been a statistician. All these basic expertise may allow you to solve a computational biology problem. So the funnel is very broad. If you are an undergraduate student or even if you are looking for a postdoctoral position, you have to have some expertise in fundamental sciences or a combination of expertise. It is often helpful if you are collaborative; any computational scientist would benefit most if they have worked with experimentalists. You are more likely to become a group leader if you can show that you have collaborated in the past or you are interested in collaborations. You also will be more successful if you are seen as a person who knows what is going on inside computational biology tools and don't just use them as black boxes. So, I think for young people wanting to pursue computational biology, it is a good idea to spend a little time developing some tools or theories and then applying them to solve specific problems, involving specific proteins, complexes or pathways, which invariably relies on collaboration with experimental biologists.

GS: Your experience and vision as the Jubilee Professor of the Academy.

Box 4. Ajanta and Ellora – Ancient Wonders

The Ajanta and Ellora caves, located in Maharashtra, India, are one of the oldest surviving monastery-temple caves of the world. The caves marked as UNESCO World Heritage Site are famous for their intricate architecture, rock-cut sculptures, and wall-paintings. According to records, the first phase of the Ajanta caves was built during the Satavahana period (100 BCE–100 CE) and the second phase during the Vakataka, period (460–480 CE). Ajanta caves constitute ancient monasteries and worship halls of different Buddhist traditions and are believed to have served as a monsoon retreat for monks, and resting site for merchants and pilgrims. The Ellora caves contain Buddhist, Hindu, and Jain monuments, and artworks.

Figure B. Ajanta Caves (Credit: Archeological Survey of India)

AS: Oh! First of all, I am surprised and honored that I have been chosen for it. So, I have accepted it. Here I know Madhu, I know Raghavan, and they are responsible for this, and that is how I am here. I have very much enjoyed the trip so far. The idea is that I spent 2–3 weeks in India, going to different places and giving some research talks, as well as more down-to-earth undergraduate level talks. I have met students and faculty. I also saw a lot of places in India that are mind-boggling, like the caves of Ajanta and Ellora (see *Box* 4). That is a totally separate discussion, but it is striking. I am not aware of any such artifacts existing elsewhere, and I didn't think it was even possible for people to organize themselves over 20 generations, starting with a specific goal and working for 20 generations to achieve it. That was just great to see and very inspiring, even though it was not science.

You know, as a scientist, you can improve science – which is exactly what I want to do – in two ways. One is by contributing directly through your research, and the other one is by inspiring or educating young scientists and helping them in one way or another so that they contribute to science, and I get a little bit of credit (laughs) indirectly for that as well (laughs). And I

think this trip is rich in such opportunities. Slightly selfishly, I have also been on a lookout for potential collaborators, and I think there may be a few. I have not planned anything, but conversations happen, and then you realize...oh wow! We could do something together. This wouldn't have happened without being here.

4. Acknowledgment

We are thankful to Prof. Andrej Sali for sharing his insights, to the Indian Academy of Sciences for providing the opportunity for such an interaction, the reviewer for his/her constructive comments, and to Prof. Rajaram Nityananda for his constant support.

Suggested Reading

[1] A M Lesk, *Introduction to Protein Architecture*, OUP, Oxford, 2001.

[2] Ardala Breda, Napoleao Fonseca Valadares, Osmar Norberto de Souza, and Richard Charles, Chapter A06 Protein Structure, Modelling and Applications, Garratt, *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach*, National Center for Biotechnology Information (US); 2008.

[3] S K Burley, S C Almo, J B Bonanno, M Capel, M R Chance, T Gaasterland, D Lin, A Sali, F W Studier, S Swaminathan, Structural Genomics: Beyond the Human Genome Project, *Nat. Genet*., 23, pp.151–157,1999.

[4] Elmar Krieger, Sander B Nabuurs, and Gert Vriend, Chapter 25- Homology Modelling, *Structural Bioinformatics*, Edited by Philip E Bourne and Helge Weissig, ISBN 0-471-20199-5, Wiley-Liss, Inc., 2003.

[5] Andras Fiser, Template-Based Protein Structure Modelling, *Methods Mol Biol*., 673, pp.7394, 2010.

[6] Krzysztof Ginalski, Nick V. Grishin, Adam Godzik, and Leszek Rychlewski1, Practical Lessons from Protein Structure Prediction, *Nucleic Acids Res*., 33(6), pp.1874–1891, 2005.

[7] https://www.biostat.wisc.edu/bmi776/lectures/threading.pdf

[8] https://salilab.org/modeller/manual/

[9] https://integrativemodelling.org/

[10] A Sali *et.al*., Putting the Pieces Together: Integrative Modelling Platform Software for Structure Determination of Macromolecular Assemblies, *PloS Biology*, Volume 10, Issue 1, pp.1–5, 2012.