

*A conversation with  
Gerard Kleywegt, Cambridge*



## A Guardian of Structural Integrity

Gerard Kleywegt, new head of EMBL's Protein Data Bank in Cambridge, talks about pitfalls and errors in determining 3D structures of proteins – and what has to be done to minimise them.

**G**erard Kleywegt obtained his PhD in 1991 from the University of Utrecht. His thesis was about automating the interpretation of 2D and 3D protein NMR spectra. Afterwards, he worked for a software company, Biosym, for half a year, “To earn some money because I was broke after my PhD. It’s so expensive in the Netherlands – you have to pay for 300 copies of your thesis and invite lots of people to a reception, a dinner and a party.” He then moved to Sweden, joining Alwyn Jones’ protein crystallography laboratory in Uppsala. “Initially I thought I was going there for a year to learn a little crystallography, just enough to understand how it works, do some computational biology, then go back to the Netherlands and work in some pharmaceutical company until I’m 65 – that was the theory.” He stayed there for 17 years, solving protein crystal structures and also doing protein crystallographic methods development and structural bioinformatics research. “The reasons were: I loved Sweden, I really liked crystallography, the lab was great and Alwyn wanted to keep me so he offered me a job. In addition, there was a cute, blonde reason.” He was coordinator (1994-2003) and director (2004-2009) of the Swedish Structural Biology Network (SBNNet), held a research fellowship from the Royal Swedish Academy of Sciences (2002-2006) and became an associate professor at Uppsala Uni-

versity in 2008 and a full professor in 2009. He has recently been appointed as head of the Protein Data Bank (PDB) in Europe at the EMBL outpost in Cambridge, UK.

*Lab Times: You’re in charge of the Protein Data Bank in Europe (PDBe)?*

**Kleywegt:** Yes, I started at the PDBe on June 1<sup>st</sup> and formally took over on 1<sup>st</sup> July.

*How does the PDB work? With three sites worldwide, who does what?*

**Kleywegt:** In fact, there are now four sites that together constitute the worldwide Protein Data Bank (wwPDB). One in Wisconsin (USA) has recently been added (the BioMagResBank, that deals specifically with nuclear magnetic resonance data), then there’s a site in Japan (Protein Data Bank Japan, PDBj), another in New Jersey (Research Collaboratory for Structural Bioinformatics, RCSB), and finally us here, at the EBI (European Bioinformatics Institute) in Cambridge, England.

*Does each site have a speciality?*

**Kleywegt:** No, the fourth site was added specifically for NMR issues but the other sites deal with deposition of any structures into the data bank. The site in New Jersey is the overall keeper of the archive

that constitutes the Protein Data Bank, but we all take depositions from scientists with the idea that, in principle, depositions from Europe and Africa should come to us, the Americas go to the US, Asia to Japan.

*How many protein structural entries are there in the PDB?*

**Kleywegt:** Almost 60,000.

*What is your estimate of the error rate in the databank?*

**Kleywegt:** That depends on what you call an error. The problem with structure determination is that you’re using experimental data and there’s always going to be experimental error. Okay, you have error bars on anything that you do, but we like to minimise the number of serious mistakes that people make, not just questions of misplacing an atom by a few hundredths of an Ångström, but putting loops or sidechains in the wrong way, or misidentifying a ligand.

*Have there been retractions from the PDB?*

**Kleywegt:** There was one just yesterday from *Nature Structural & Molecular Biology*, announced in the July 2009 issue on page 7. They have retracted a structure that was originally published in 2000, together with

*“There was this ‘great penta retraction’ as someone sarcastically called it.”*

the paper. Next week, this will be formally removed from the databank.

*On 22 December, five papers on structural studies of ATP-binding cassette (ABC) transporters – three in Science, one in PNAS, and one in the Journal of Molecular Biology – were simultaneously retracted in a letter to Science. At the time, Alwyn Jones and you wrote to Science, “We have much sympathy for your readers but very little for the magazine. This is not the first time incorrect structures have been published in Science, and it will not be the last time!” Is this still the most serious case to date?*

**Kleywegt:** Yes, that was a bit of a shock on Christmas morning! There was this “great penta retraction” as someone sarcastically called it in a later correspondence to *Science*. There were five ABC transporters that all turned out to be seriously incorrect.

*What happened?*

**Kleywegt:** The structures all came from the same lab. If they had all been determined in different labs this probably wouldn't have happened because everyone would have had to have made the same mistakes in different labs – most unlikely. In fact, the problem came to light thanks to the fact that a related structure was solved in Switzerland and published in *Nature*. It showed that the structures from the other lab were wrong. It seems that the original lab had solved one of these structures incorrectly using very low-resolution experimental data – where you can't see much detail. Once they got the first one wrong, they probably used that as a starting point for constructing the other models with data collected later on. It became a self-fulfilling prophecy. If you start from something which you determined previously, then you're going to end up with something that looks like what you determined previously. The four later structures were probably based on the original one, which was tragically incorrect.

*Have they proposed a corrected version?*

**Kleywegt:** They retracted all their papers, and the structures that were deposited in the PDB were also removed. Later they published a paper in *PNAS* where they had re-done the structure determination and came up with the (hopefully) correct models, which were deposited in the PDB in the year following the retraction.

*Can we go into the technical side of how protein structures are determined? You state that the main sources of error are more subjective, coming from the experimenters rather than their hardware. Do you think the field suffers from a mixture of “inexperience, incompetence, and over-confidence”?*

**Kleywegt:** If you'd called me two months ago (before I became head of the PDB) I would have agreed immediately! However, there are two things to note: first of all, crystallography is an experimental science and the data that you can collect from your crystals can either show an awful lot of detail or very little, depending on how flexible your molecules are and how ordered the crystals are that you have obtained. Sometimes you cannot collect any better data than, say, 4 Å resolution which is very low – it means you see helices as cylinders of density and you don't see individual atoms or even sidechains. On the other hand, if you're lucky, you have a rock-solid protein that crystallises and permits very high resolution. This means that you may even see the hydrogen atoms. In the case of high resolution, it's almost impos-

**“Your crystals can either show an awful lot of detail or very little.”**

sible to make serious mistakes; you can tell from the data immediately that this is a leucine, this is a tryptophan etc. But at low resolution, that's where the subjectivity comes in, because there you need to have experience. You're presented with a map which has very little detail. It's almost anyone's guess how you should interpret that. If you give it to different crystallographers, they're bound to come up with different models. This depends on how much experience they have and how conservative they are. In my previous career, working in Uppsala with Alwyn Jones, what we tried to do was to explain to the crystallographers that they should be conservative, that they should start with a model with few assumptions. Only if your data is good enough can you start making more assumptions and adding more detail to your model. This is something you have to teach every new crystallographer. Otherwise, they make the same mistakes that have been made many times before. To some extent it's an educational issue that crystallographers have to be trained in being somewhat conservative and not over-interpreting the data when the data doesn't allow you to see the detail.

*Concerning the resolution question, you often write about the 2 Å resolution*

limit. With better than 2 Ångström resolution, you can be reasonably sure you have good data but at less than 2 Ångström, a lot of error can creep in. However, this limitation is not actually due to the equipment, it's due to the quality of the crystal.

**Kleywegt:** Yes, it's a matter of how well your molecules are ordered inside your crystal. When you have a lot of disorder, it means that the molecules may be slightly differently oriented, or the domains are slightly differently oriented with respect to one another or, for example, that there is movement of a whole loop, giving it different conformations. The crystal is like an amplifier – if all the molecules are exactly identical and they are all perfectly oriented then you get a very good signal and very high resolution. But if there is disorder, where molecules are slightly different throughout the crystal, then it sort of blurs the data and you get lower resolution. You don't see the features that you'd like to be able to see.



Sculpture of the BlyS protein structure in Cold Spring Harbor.

In addition, you've identified problems associated with the actual conditions under which the crystals are made: pH, problems of water molecules incorporated in the structure etc. These can have quite dramatic effects on how the protein crystal forms.

**Kleywegt:** Whether or not you get this well-ordered crystal can sometimes be modulated. For example, if you crystallise a protein together with a ligand, you often lock the protein into a particular conformation. You may be able to get much higher resolution for a protein-ligand complex than for the protein by itself. That's not uncommon. People do that. They do a lot of experiments to try and get the initial crystals, to try and

optimise them, to get as high a resolution as possible. They will attempt to vary ionic strength, pH, additives and hopefully get crystals that diffract to high resolution.

*Bombarding your crystal with X-rays gives you an electron diffraction model but then you have to reconstruct the structure using the computer and mathematical models. Does this require a lot of subjective interpretation?*

**Kleywegt:** What we get from the experiment are the 'diffraction amplitudes' but, in fact, we're missing half of the information. We don't have the so-called 'phases'. If you have an optical microscope, you can see a perfect image of whatever you put under it, but we only get half the information in X-ray experiments. The other half we have to recover using all sorts of clever, partly chemical, partly mathematical, tricks. In the end, if we assume that this works then what we get is the 'electron density map'. This tells us where the electrons are concentrated in three-dimensional space, but there are no labels to say that this density belongs to a carbon atom, or to an oxygen, or whatever. This is where the interpretation of the density is difficult, where the crystallographer comes in.

Nowadays, software can often do it all by itself. It can interpret the electron density because we know the amino acid sequence if it's a protein. For example, a very big blob of density could be a tryptophan or a tyrosine and then maybe you see another big blob a bit further down and this could also be one of these amino acids. But if you look in the sequence and see there's only one place where you have two of these big residues close together, then you can get a handle on it and say, 'okay, so maybe this blob of electrons belongs to tyrosine-178', and then you can start building your model and assigning atom positions to that electron density. But the density itself just says there are electrons concentrated here and there. It is the crystallographer's task to interpret this using the protein sequence to come up with an atomic model that says, 'the carbon that you see here is from alanine-52, lysine-xyz etc.' for all the atoms in the protein sequence.

This is why the interpretation is tricky, especially when it is done by hand. That's when the experience of the crystallographer comes in, because this is obviously the most difficult step. Nowadays, at high resolution, software can do a lot of this for the

easy parts, a lot of different programmes can do it automatically. But you're always going to be left with parts that are difficult, which you have to do by hand, because the electron density is very poor in certain places, for all sorts of reasons. Again you have to use your judgement and your experience in order to interpret the experimental electron density.

*How do other structural analysis methods, like NMR, compare?*

**Kleywegt:** NMR is completely different. You don't use a crystal, you use a solution of your protein and you use a magnetic field. The kind of information you get from NMR is also different. In crystallography, what you can get is information about how the electrons are distributed in 3-D space; in NMR the most common type of information is about the distances between pairs of atoms in your protein. You have to determine which atom is responsible for which signal in the NMR spectra. Then you can say, for example, that 'this particular hydrogen in my tyrosine-56 side-chain lies within 6 Ångström of the amide proton in alanine-12.' You get lots of information that allows you

to construct a model based on many calculations that explains, as well as possible, all these observations about the distances of pairs of atoms in the protein.

**"You're always going to be left with parts that are difficult, which you have to do by hand."**

*Are there still size limits with NMR?*

**Kleywegt:** It's better nowadays – typically in the 20 to 30 kilodalton range – but bigger structures can sometimes be solved. But if the protein gets much bigger, you get problems with NMR, whereas with X-ray crystallography we can do anything from a simple salt to a virus or a ribosome. These have all been solved by X-ray crystallography. You can look at even bigger systems with techniques such as single particle cryo-electron microscopy and, increasingly, tomography.

*What sort of extra problems can arise when passing from individual proteins to multimolecular systems, for example, receptor-ligand complexes? You've written a couple of reviews with researchers from AstraZeneca that discuss such problems ("Application and limitations of X-ray crystallographic data in structure-based ligand and drug design." Davis A. et al., Angewandte Chemie vol. 42: 2718-36; "Limitations and lessons in the use of X-ray structural information in drug design." Davis A et al., Drug Discovery Today*

vol. 13: 831-41), since they concern drug design and the huge investments that are made by the pharmaceutical industry in order to find new drugs using structure-based design and combinatorial chemistry.

**Kleywegt:** The problem here is that most crystallographers understand proteins and amino acids quite well, but when you add ligands, you're adding an extra level of complication. You have to understand the chemistry of the ligand. First of all, you have to be sure that you know exactly what you put into your crystallisation solution. Sometimes there are reactions going on that you're not aware of and sometimes you're not looking at what you thought you were looking at. For example, you might have an oxidation product, or it may be broken down. Second, building a model for a small molecule can be just as difficult, or more difficult than building one for a protein because usually crystallographers are less well-versed in chemistry than they are in dealing with amino acids and nucleic acids. Another problem is that all the software mostly deals with proteins and nucleic acids, so as soon as you introduce something that these programmes don't know about, you have to specify in advance what you think this molecule is going to look like. That requires quite a bit of chemical knowledge – for example, you have to specify that certain groups are going to be planar because they're aromatic, and if you lack experience and a chemistry background, you can make all sorts of interesting mistakes!

There's the case in the *Drug Discovery Today* review, where we cite a structure published in *Molecular Cell* in 1999. This was a peroxisome proliferator-activated receptor (PPAR – nuclear receptors for fatty acid ligands) that had been modelled with a bunch of water molecules that they identified in the active site. But this structure was a conundrum because it didn't explain any of the biology (because the conformation of the pure protein was the same as the activated/ligand-bound state).



**Gerard Klewegt:** Sometimes playing advocatus diaboli when it comes to protein structures.

Then, a number of years later (in 2006), another lab realised that the original researchers hadn't looked very closely at their data, because the electron density features that they had in the ligand-binding site were not for water molecules. There was in fact a fatty acid bound by the protein! Just the fact that they hadn't added the fatty acid themselves to the crystallisation solution, meant that they had never even considered the possibility that the protein might have retained something during the purification. It was something they hadn't even thought about! (In fact, a mixture of fatty acids had been acquired by the protein from the bacterial expression system.) So, they completely missed the existence of a very interesting fatty acid ligand bound to the active site. All of a sudden the whole structure made sense. All the biology made sense in the light of the reinterpreted structure.

*Have there been any major errors with clinical consequences?*

**Kleywegt:** No, I don't think it's quite that dramatic because if you make mistakes you usually end up with something that is inactive. I don't know of any such cases.

*Is there a very heavy investment in drug discovery?*

**Kleywegt:** Yes, the structural work is usually at the very early stages of drug development, where they try to optimise the binding of small molecules to a receptor. But then there are so many checks along the way before you get to the clinic, it's unlikely that a mistake in the crystallography would actually go all the way through to the market.

*But it can influence the choice of research direction on a given receptor or class of molecules?*

**Kleywegt:** It could potentially mean that a lot of money and effort is wasted if you draw the wrong conclusion. There was an example in the *Drug Discovery Today*

paper from AstraZeneca's own laboratory, where they had misinterpreted something and it may have cost them money because they were drawing the wrong conclusions from the model. AstraZeneca were looking for inhibitors of inducible nitric acid oxide synthase (iNOS) as a potential treatment for inflammatory disease. Their strategy required a ligand that could interact with a critical aspartic acid residue in iNOS. Their design strategy eventually resulted in the synthesis of a compound that had much

lower activity than expected. The crystallographer analysed the crystal of iNOS with the compound. The low resolution crystallography confirmed the predicted interaction at the aspartic acid residue. However, shortly before publishing their work, a new crystallographer arrived who reanalysed the data and concluded that no valid analysis could possibly be made from this data. Unable to obtain better structural information, the manuscript was not submitted and the project was abandoned.

*Can we discuss ways of cleaning up the Protein Data Bank and how to avoid repeating the mistakes of the past? This year, you wrote a very interesting paper on validation (Acta Crystallogr D Biol Crystallogr. Vol. 65:134-9). You've written it like a lesson in the philosophy of science, explaining from a very didactic viewpoint the "why, what and how of validation". Do you feel that all scientists, but especially crystallographers, would do well to understand the principles and practice of your validation model?*

**Kleywegt:** It was completely generic in principle, so it applies to all experimental science. I think that when you use experimental data, you should always validate your interpretation of that data. It doesn't matter if you study kinetics or crystallography. I tried to write it as partly generic, partly for crystallography, illustrating how we do this and this kind of study and how we should test our data and models. However, the principles are general to experimental science whenever you take data that you try to interpret in terms of a model.

*At the end of your validation paper, you speak of the creation of a validation taskforce. So is your aim to investigate how well everyone else has been validating their work?*

**Kleywegt:** This is something that the collaborating institutes in the wwPDB have initiated. Of course, I was a member and since I can't advise myself, I had to resign.

*"But this structure was a conundrum because it didn't explain any of the biology."*

I'm now on the other side of the fence. But the idea of this task force is to actually have people, who have been thinking about validation and quality controls, come up with a set of criteria that we should check when researchers deposit new structures into the data bank. We would like any problems to be caught as early as possible. Preferably, researchers should do this before they deposit their data with us. They should do it even before they start writing the paper, so that they are fairly sure that the model they are going to describe in their paper and deposit in the PDB doesn't contain any obvious problems and is not over-interpreting the data. This validation task force is going to produce a list of recommendations. I have a job opening for someone to implement these, for a crystallographer to produce a software package that we're going to use. It will be downloadable and usable by researchers in their own laboratories to do these validation checks, hopefully spotting errors before they end up in the literature and the PDB.

*You say people shouldn't be afraid of the validation task force because, "validation is your friend". You want people to take a positive attitude?*

**Kleywegt:** I've taught a lot of courses. When you talk to the students and post-docs who do the actual work they are usually eager to learn about these things because they don't want to look silly. They know that if they have to retract a paper and a structure that's not going to do their careers any good. They want to avoid making serious mistakes. They want to learn about better ways of producing their models and using ways to validate those models. It's not going to be a policing type of task force. It's just going to say that we, as a community, think that these are the kind of tests that you need to do on your structure when you deposit it. It will probably result in a one-page report that you can also send, for example, to the journal where you're going to publish the paper, to show the referees that the structure has a quality report and that it seems to be okay.

*Gert Vriend from the Centre for Molecular and Biomolecular Informatics in Nijmegen*

*en (Netherlands) wrote to Science about the ABC transporter penta retraction observing that in 1996, "Hooft reported one million anomalies in the PDB, and we recently detected ten times as many anomalies in a PDB that is ten times as large." What do think about his initiative to use his error-detecting software to systematically analyse structures in the PDB?*

**Kleywegt:** I've been using his validation programmes as well. He has a website called PDBReport (<http://swift.cmbi.ru.nl/gv/pdbreport/>), where every structure in



Photographed by Kleywegt in Beijing – as a kind of motto for protein structure research.

the PDB is being analysed. I've been using that for more than a decade. When I did crystallography I'd always run my structures through his software to find any sidechains that were misoriented or waters that shouldn't be there and other such things. It's extremely useful. He's also part

of the task force, so we'll probably be using some of his methods and software as part of this new validation package that we're going to produce.

*Will there be any more nasty surprises for the PDB?*

**Kleywegt:** Well, there was one yesterday! But I hope that once these validation tools are in place, we're going to minimise the occurrence of these errors in the future. What has already been published we can't do much about, but we can now try to do the validation early so that it doesn't even get deposited into the PDB if there's any indication that there might be something seriously wrong with a structure.

*Do all journals now agree that the coordinates for the structures have to be deposited in the PDB before publication?*

**Kleywegt:** Not all, but most of them. They require that the structure and the experimental data be deposited for the protein. If you only have the structure then, even if you think there's something fishy, you can never prove it. What you need is the experimental data to actually prove that there's something that could be improved or to show that there's something seriously wrong. Since February last year, it's mandatory to deposit experimental data in the PDB.

INTERVIEW: JEREMY GARWOOD