I came across this article while tidying up my documents folder. The original request was for a blog-style piece, written in the third person, about an early piece of my work that turned out to have had a much wider impact on protein structure prediction, CASP and AlphaFold, which was very much in the news at the time. Nothing came of the proposed collection, but I rather enjoyed writing the article, so I have tidied it up and decided to share it. It is not meant to argue that I invented everything that followed. The intended spirit is much closer to James Burke's old BBC series Connections: a look at how one idea led, sometimes rather indirectly, to the next. In my head, at least, it is still being read in his voice.

The Quiet Line from NK-lysin to AlphaFold

David T. Jones's work in protein structure prediction is usually discussed as a series of separate methods and results. Put together, however, there is a fairly clear line from the first successful blind de novo fold prediction in CASP, through the use of evolutionary variation to predict contacts and distances, to the first AlphaFold system.

Most accounts of modern protein structure prediction are built around a few obvious landmarks. Rosetta made fragment assembly into a practical approach to de novo modelling. Covariation methods showed that alignments of related sequences contain information about which residues are in contact. Deep learning turned those contacts into more useful geometrical constraints, and AlphaFold brought the pieces together at a scale and level of accuracy that changed the field.

That account is broadly right. The problem is that it tends to split the work into a number of independent stories, and in doing so it misses an important line running through the work of David T. Jones.

In 1996, at the second Critical Assessment of Techniques for Protein Structure Prediction (CASP2), Jones submitted a model of the 78-residue protein NK-lysin. There was no suitable structural template. Even so, the calculation recovered the overall topology of the protein. It was not an atomically accurate model, but the C-alpha root-mean-square deviation was 6.2 A over the complete chain and below 5 A over the first 66 residues [1].

The independent CASP2 assessment found only one satisfactory full-coordinate ab initio prediction, pig NK-lysin [2]. It is therefore generally regarded as the first successful blind de novo prediction of a protein fold in CASP.

What is interesting in retrospect is not just that the prediction worked. The calculation already contained several ideas that would become central to structure prediction over the following two decades.

At a glance: CASP2 NK-lysin (1996) -> PSICOV (2012) -> MetaPSICOV and DMPfold -> AlphaFold 1 (CASP13, 2018).

A new fold, predicted blind

The word blind matters here. Before CASP, methods were normally developed and tested on proteins whose experimental structures were already available. Even with the best intentions, it was difficult to rule out conscious or unconscious tuning to known answers. CASP dealt with this by asking groups to predict structures that had been solved experimentally but had not yet been released.

CASP2 took place in December 1996, with more than 70 groups submitting predictions in categories that included homology modelling, fold recognition and ab initio prediction [3]. For NK-lysin, there was no known fold that Jones could simply identify and use as a template. Subsequent structural comparison found no significant match between the experimental NK-lysin structure and the library of known domain folds used in the assessment [1].

Put simply, this was not a case of finding the right answer in a structural library. The calculation had to construct a topology that was not recognisably present in the database.

Jones was careful not to overstate the result. He described it, to the best of his knowledge, as the first correct blind ab initio fold prediction produced by a computational 'heavy numbers' approach. He also pointed out that NK-lysin had a relatively simple alpha-helical topology, and that one success did not establish that the method would work generally [1].

That caution was entirely reasonable. Nevertheless, the result crossed an important boundary. Under genuinely blind conditions, a computer had searched a large space of possible conformations and selected an approximately correct topology for a previously unknown fold.

A recognisably Rosetta-like idea

The method was based on fragment assembly.

Rather than varying every backbone torsion angle independently, the algorithm took structural fragments from experimentally determined proteins and assembled them into candidate chains. Jones concentrated on recognisable supersecondary motifs, including alpha-hairpins, alpha-corners, beta-hairpins and beta-alpha-beta units. The fragments were generally between 9 and 31 residues long [1].

Candidate structures were changed by replacing sections of the chain with fragments from the library. A statistical energy function evaluated residue-pair interactions, solvation, steric overlap, hydrogen bonding and compactness. Simulated annealing, with the Metropolis criterion, was then used to decide whether each proposed move should be accepted [1].

In simple terms, the recipe was structural fragments, a knowledge-based energy function and stochastic search, with the aim of producing a complete fold.

That is also, in broad terms, the central idea behind early Rosetta.

The foundational Rosetta work assembled fragments whose local sequences resembled the target, again using simulated annealing and statistical scoring. In its CASP3 implementation, Rosetta inserted 3- and 9-residue fragments, generated around 1,200 structures for each target and selected models from favourable and well-populated regions of conformational space [4].

The timings are worth spelling out. Jones's NK-lysin prediction was made at CASP2 in 1996. Rosetta made its influential CASP debut at CASP3 in 1998, and the original Rosetta fragment-assembly paper appeared in 1997. The two approaches were therefore almost certainly developed independently over overlapping periods, and it would be wrong to suggest that one group simply copied the other.

The historical point is narrower, but still important: Jones demonstrated the fragment-assembly idea successfully in a blind CASP experiment before Rosetta appeared in CASP.

There were, of course, substantial technical differences. Jones used relatively large supersecondary motifs, so a single move could introduce a sizeable piece of tertiary organisation. Rosetta generally used smaller fragments, allowing the global topology to emerge from many local substitutions. The smaller fragments turned out to be more flexible and scalable, and the Rosetta programme subsequently received sustained development in sampling, all-atom refinement and protein design.

Even with those differences, the conceptual similarity is hard to miss. The first successful blind de novo fold prediction in CASP used a strategy very close to the one that later became associated with Rosetta.

Folding an evolutionary ensemble

In retrospect, the more interesting part of the NK-lysin calculation may not have been the fragment assembly at all.

Jones did not score each candidate structure using only the pig NK-lysin sequence. He used an alignment of six related sequences. For each proposed backbone, most of the energy terms were calculated separately for every sequence and then averaged. Steric compatibility was handled more strictly, by taking the worst value across the family [1].

He described this as folding an 'evolutionary ensemble' of polypeptide chains in parallel.

The immediate reason was statistical. Averaging the pairwise energy function over several homologues reduced the noise associated with any one sequence and improved convergence [1]. The more interesting implication, however, is that a candidate fold had to accommodate the pattern of amino-acid variation across an entire protein family.

Suppose two positions are close together in the real structure. One homologue might have lysine and glutamate at those positions, another arginine and aspartate, and another two neutral polar residues. The identities change, but the substitutions preserve a chemically reasonable interaction.

A correct fold brings those positions together and can accommodate the observed combinations. An incorrect fold may place together positions whose substitutions are repeatedly incompatible. Averaging over the homologues therefore strengthens the structural signal shared by the family and reduces the influence of oddities in any one sequence.

This was not covariance analysis in the modern mathematical sense. It did not calculate correlations between alignment columns, fit a Potts model or try to separate direct from indirect coupling. The signal included ordinary conservation, conservative substitution and the reduction of sequence-specific noise, as well as genuinely compensatory changes.

It is therefore better described as an implicit evolutionary constraint than as explicit coevolutionary contact prediction.

Even with that qualification, the underlying idea was important. Natural sequence variation was being treated as evidence about three-dimensional structure. Evolution had sampled alternative amino-acid combinations, and the correct backbone was expected to be the one most compatible with the resulting family.

From evolutionary compatibility to direct coupling

More than a decade later, Jones returned to the same biological problem in a much more explicit form.

Traditional covariance measures could identify pairs of alignment positions that changed together, but they could not reliably distinguish direct interactions from indirect ones. If residue A interacts with B, and B

interacts with C, then A and C may look correlated even though they are not in physical contact. Shared phylogenetic history introduces further correlations that have nothing to do with structure.

In 2012, Jones, Daniel Buchan, Domenico Cozzetto and Massimiliano Pontil introduced PSICOV. The method used sparse inverse-covariance estimation to identify likely direct statistical couplings between positions in a multiple-sequence alignment. On their benchmark, PSICOV substantially improved the precision of long-range contact prediction compared with mutual-information and Bayesian-network approaches [5].

The link to the NK-lysin work is not that the algorithms were the same, because they were not. The link is the underlying biological proposition: a protein family contains information about the contacts needed to preserve its common fold.

A protein family contains information about the contacts needed to preserve its common fold.

In the 1996 calculation, that information was extracted by testing whether complete family sequences were compatible with a proposed backbone. In PSICOV, it was extracted directly from the statistical relationships between alignment columns.

The direction of inference had therefore been reversed. In the NK-lysin work, a candidate structure was tested against a sequence family. In PSICOV, the sequence family was used to predict structural contacts.

That reversal mattered. Once contacts could be predicted before a structure was built, they could be used as long-range restraints to guide the folding calculation.

From contacts to learned geometry

PSICOV was followed by MetaPSICOV, which combined several coevolutionary methods with more conventional sequence-derived features in a two-stage neural-network predictor. The method adjusted the balance between classical and covariation features according to the depth of the alignment, allowing useful predictions to be made for a wider range of protein families [6].

DeepMetaPSICOV then replaced much of this hand-built feature engineering with deep residual networks. In CASP13, the Jones group combined DeepMetaPSICOV with improved alignment generation and metagenomic sequence data to predict contacts for difficult targets [7].

The obvious next step was to move beyond binary contacts. Knowing that two residues are within, say, 8 A is useful, but a probability distribution over possible distances contains much more geometrical information.

DMPfold, developed by Joe Greener, Shaun Kandathil and Jones, predicted interatomic distance bounds, backbone hydrogen bonds and torsion angles. These predictions were converted into 3-D models and then fed back into the network, allowing mutually inconsistent constraints to be revised iteratively. The method extended useful de novo modelling to proteins with relatively shallow alignments and produced confident models for previously uncharacterised protein families [8].

The progression is quite straightforward when set out in this way: compatibility with a sequence family, direct evolutionary couplings, predicted contacts, predicted distances and finally a complete 3-D structure.

At each stage, the evolutionary information became more explicit and was converted into a more complete description of protein geometry.

Jones and AlphaFold 1

This line of work eventually converged with AlphaFold.

David Jones was a co-author of the original AlphaFold Nature paper, now usually referred to as AlphaFold 1. The author-contribution statement records that he provided advice and guidance on protein structure prediction methodology and contributed to the manuscript [9].

AlphaFold 1 used neural networks to predict probability distributions over distances between residue pairs. These distributions were converted into protein-specific potentials of mean force, which were then minimised to generate complete structures. At CASP13, the method produced high-accuracy models for 24 of the 43 free-modelling domains, compared with 14 for the next best group [9].

The technical argument in the AlphaFold paper starts from evolutionary information. Covariation between homologous sequences can indicate which amino acids are likely to be in contact, and predicted distance distributions contain more useful information than binary contact maps [9].

It is important not to overstate the connection. Jones was not the principal architect of AlphaFold, and AlphaFold was not simply an implementation of his earlier methods. It was the work of a large team and involved major advances in neural-network architecture, optimisation, data processing and engineering.

At the same time, his involvement was not an arbitrary advisory connection. By then, Jones had spent more than twenty years working on the chain of ideas that linked evolutionary sequence information to protein geometry.

Why the contribution is easy to overlook

Scientific histories tend to organise themselves around successful platforms.

Rosetta became a large and sustained programme spanning structure prediction, molecular modelling and protein design. AlphaFold produced a dramatic improvement in accuracy and was then deployed at extraordinary scale. It is not surprising that these two programmes dominate retrospective accounts.

The NK-lysin result was rather different. It was one conspicuous success from a method that did not become the dominant software platform in the field. The overall topology was correct, but the detailed packing was poor. The C-terminal region was substantially wrong, and some models incorrectly introduced beta-structure. Jones described the result as resembling a native-like molten globule, with approximately correct global topology but inadequate side-chain packing [1].

His conclusion was correspondingly cautious. Although he described the prediction as a major step, he also wrote that it was difficult to envisage the method becoming a practical ab initio prediction tool in the near future [1].

That restraint probably helps to explain why the work has had a limited place in popular accounts. The paper did not claim that fragment assembly had solved de novo structure prediction. It did not present the use of multiple sequences as the start of a future coevolutionary revolution. It reported one successful blind experiment, explained what had gone wrong and made clear how much remained to be done.

Nevertheless, several of the ideas turned out to be durable.

The work showed that an approximately correct global topology could be recovered without a structural template. It demonstrated fragment assembly and statistical scoring in CASP before Rosetta's CASP debut. It used a multiple-sequence alignment to constrain tertiary structure by treating homologues as an evolutionary ensemble. Jones later helped turn amino-acid covariation into explicit contacts, contacts into learned distances and learned distances into complete structures. He then contributed methodological advice to AlphaFold 1.

A more complete history

It would be wrong to draw a perfectly straight line from NK-lysin to AlphaFold and assign the whole route to one person.

Fragment-based modelling had important precursors and was developed independently by the Baker group. Direct-coupling analysis involved contributions from statistical physicists, computational biologists and several competing groups. Deep learning for contact and distance prediction advanced rapidly in a number of laboratories. AlphaFold itself was very clearly a team achievement.

The opposite simplification is also misleading, however. Jones was not simply a peripheral adviser who happened to appear on the first AlphaFold paper.

A more complete account would recognise four substantial contributions: the first successful blind de novo fold prediction in CASP; an early and effective fragment-assembly strategy, demonstrated before Rosetta entered CASP; foundational work in extracting structural contacts and distances from evolutionary sequence variation; and direct participation in the methodology and publication of AlphaFold 1.

Perhaps the most revealing phrase remains Jones's description of the NK-lysin calculation as the parallel folding of an 'evolutionary ensemble' [1].

In 1996, this was presented mainly as a practical way to reduce noise in a statistical energy function. Seen from the present, it also sounds like an early statement of a principle that now sits near the centre of protein structure prediction: evolution has carried out an enormous natural mutational experiment, and the record of that experiment contains information about the geometry of the fold.

Jones's impact therefore lies not just in one early CASP result or one successful algorithm. It lies in repeatedly returning to the connection between sequence, evolution and structure, and making that connection progressively more explicit and more useful.

References

[1] Jones, D. T. (1997). Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins, Supplement 1, 185–191. Link

[2] Lesk, A. M. (1997). CASP2: report on ab initio predictions. Proteins, Supplement 1, 151–166. Link

[3] Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K., and Pedersen, J. T. (1997). Critical assessment of methods of protein structure prediction (CASP): Round II. Proteins, Supplement 1, 2–6. Link

[4] Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. (1999). Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins, Supplement 3, 171–176. Link

[5] Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics, 28, 184–190. Link

[6] Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics, 31, 999–1006. Link

[7] Kandathil, S. M., Greener, J. G., and Jones, D. T. (2019). Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins, 87, 1092–1099. Link

[8] Greener, J. G., Kandathil, S. M., and Jones, D. T. (2019). Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nature Communications, 10, 3977. Link

[9] Senior, A. W., Evans, R., Jumper, J., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577, 706–710. Link

Source note. The account of the NK-lysin method and its results is based mainly on Jones's original CASP2 paper, supplied with this article. The later publications are cited for the subsequent methodological developments.