Summaries of Presentations at CCPN Conference
April 2001
David Snyder

CCPN Web-Site: http://www.bio.cam.ac.uk/nmr/ccp/
Presentations from conference should be available soon

General Summary

Even though there were actually a number of established scientists in attendance, the meeting seemed to be geared toward graduate students and post-docs. Many of the presentations focused on giving advice about and descriptions of the process of structural determination. There was little actual discussion of either algorithms or biology, but many of the opinions were quite thought provoking and the descriptions of structural determination are useful in determining acceptable performance for automation methods.

Paul Driscoll (Introduction)

This presentation described the CCPN and its support. The Collaborative Computing Project for NMR is supported by the BBSRC and commercial subscribers. The funding is for a platform for software development and the encouragement of a common data standard. The CCPN will also hold meetings and workshops, produce www reports, etc., in order to define and encourage the best practice in NMR. In particular, the development of AZARA and ANSIG will be a focus.

Dr. Driscoll also mentioned a particular paper of interest: “Does NMR Mean ‘Not for Molecular Replacement'?” The authors encouragingly answer “no” and suggest that what is important is that the whole ensemble of NMR structures be used and not just the energy minimized mean structure. This, of course, begs the question of “how many structures is enough” - a question that was not really addressed in the conference, but maybe the authors of the aforementioned paper have a suggestion?

Peter Domaille (“Automating NOE Assignments [...]”)

This speaker addressed the issue of the automation of NOE assignments. Dr. Domaille feels that “‘automated' methods still require [too] much user interaction and there is still a steep learning curve” and also that NOE assignment is the bottleneck step in structure determination. The problem is that chemical shifts are not unique in identifying interacting residues (I am beginning to wonder how much this problem has been addressed in the context of hashing - after all, we essentially treat chemical shifts as hashes of atoms (albeit we do not know the hash function); these issues may also be addressed in coding theory where the atoms are encoded non-uniquely as chemical shifts). The speaker described some of the methods (based mainly upon the ANSIG and AZARA packages) he uses as well as some of their weaknesses. He suggests that, in particular, peak-picking requires care and experience. Additional suggestions made were to collect the best possible data with attention to detail, get as many (even redundant assignments) as possible, not to underestimate tolerances and to pay close attention to peak-centering and systematic processing errors. It would seem from this that proper registration of peak lists is important, although this issue was not addressed. Dr. Domaille also suggested that one should carefully screen contact maps for “orphans”, a concept I do not fully understand, although, I am sure we can detect these via clustering.

As far as NOE assignment is concerned, an important concept is that of ambiguous distance restraints (Reference: Michael Nilges, *J. Mol. Biol.*, 1995). The speaker commented that, when using an ARIA type rejection strategy, it is important to keep spectra separate for recalibration and to reduce the “level” very slowly. A final suggestion was to look at each residue separately in the Ramachandran plot - this, I notice, is now a feature in ProCheck.

Alexandre Bonvin ("How far can we get?" without NOE data)

According to this speaker, two bottlenecks in NMR structural determination are experimental data acquisition and analysis of NOE data/structure calculation - in the latter case, manual analysis is always needed. To speed NMR structure determination, we can use cryogenic probes, per-deuterated samples and forget about NOEs, thus circumventing (at least for a time) the second bottleneck.

It turns out that with available data of chemical shifts, H-bonds and residual dipolar couplings, one can get a bundle of structures, some, though certainly not all, of which are reasonable. The problem is picking out the correct structures: most methods one can imagine that could accomplish this task simply do not work. Empirically (not theoretically) derived potential functions work sometimes, but what works best is clustering. The clustering method used by the speaker's group is *ad hoc* (I asked about more sophisticated methods: they have not really explored these). The general idea is that correct structures are all similar whereas others are more uniformly distributed in structure space.

The results are not really good, but should be good enough for screening and as initial folds to start NOE analysis. My concern, however, is that such structures still seem like Y. Ito's first DinI structure in that they only may capture the gist of the correct structure - the question is how can such structures be perturbed by refinement to correctness?

The conclusion reached by the speaker is that correct folds can be generated at stage of backbone assignments, although this method doesn't work so well with all alpha-helical proteins. Similar analysis will need to be done with per-deuterated data to see that method is still applicable In the future, this kind of NMR analysis should be able to complement *ab-initio* fold prediction - results are of similar quality but former is good with beta sheets whereas the latter is better with alpha helices.

Flemming Poulsen ("Obsticles in high throughput structure determination [...]"

This talk may be summarized by the following statement: "high throughput means not just rapid structure determination but also decent samples to start with". After thus summarizing his opinion, Dr. Poulsen then listed some criteria for target selection and pointed out that in structural genomics, the concept of an "interesting" target has a different meaning.

An important obstacle in structure determination is protein degradation: there was some discussion of this: one possibility is that the protein samples have not been sufficiently purified from proteolytic enzymes and that some enzymes were easier to remove/less "dangerous" than others.

Some interesting data from this talk were that the calculation of sheet/helical regions with WASS agrees with RDC measurements and that it took a student three days to peak-pick six COSY type spectra.

Mike Williamson (Requirements for structure determination)

The speaker began by asking how good does a structure need to be. The answer due to G. Wagner is "good enough to provide biological information", but non-specialists will be using structure, so should put our best foot forward. Dr. Williamson then pointed out the importance of "holonomic constraints" (i.e. constraints from what we know of molecular structure in general) in structural determination, especially in the case of NMR.

In determining protein structures, we want the best precision (generally taken to be RMSD) possible as long as an accurate structure is obtained, but accuracy is hard to measure, An important question is what bundle is trying to depict: the speaker thinks it represents the best guesses at the mean solution structure. The RMSD then is the best guess at the deviation in the mean of the structure.

After giving this "philosophical" background, the speaker proceeded to discuss the "problems with NMR". These problems include, in particular, motion which results in time-averaging to incorrect structures (although the speaker underestimated the degree to which this also creates problems in crystallography - in general, many speakers seemed to be unreasonably easy on crystallographic structural determination relative to NMR) and spin diffusion which can lead to strong than expected NOEs, especially when motion is also occurring. One particular concern of the speaker is that NMR structures cannot be directly validated.

Dr. Williamson then concluded with a list of many recommendations regarding structural determination which I list below with some marginal commentary:

Use lots of restraints (at least 15 per residue)
Use many different kinds of restraints
NOEs in particular are important

(Speaker suggested that CCPN should work on software to pick/assign NOE data)

Stereospecific assignment of Val/Leu is easy and worth it (contrast with opinion of Fosner (sp?) et al, JBNMR, 9:245-258, 1997)

NOEs:

Should calibrate NOEs with distances that are similar to those you wish to calculate - in particular, be cautious in calibration, but strong/medium/weak classification looses too much information

When should ambiguous NOEs be considered? Toward the end?
Remove trivial NOEs from list (others suggest otherwise)

Hydrogen bonds - showed example of how H-exchange data is misleading, but in combination with amide proton temperature coefficients can be useful

Chemical shifts

C13 chemical shifts (+/- H1 shifts) are great restraints so long as they are not inaccurate
H1 chemical shifts should not be used in refinement

Refinement:

Clear criteria for selection of structures: use fair sampling (or all) of low energy structures at each stage
All procedures should somewhere be clearly documented and justified
Relaxation matrix analysis: maybe should not do this as we do not know overall motion/tumbling of protein
Use H-bonds only for final refinement

Analysis:

Compare structure to input restraints and know protein geometry - also compare structure to restraints not used
Pay particular attention to “riskier violations” - i.e. NOEs that are, in some sense non-redundant

DAS running commentary:
I may have algorithm to ID these

Make all details available

Ramachandran plot is important (Doreleijers et. al., *JMB* 1998, 281, 149ff)
List table of statistics

DAS running commentary:
This last recommendation was surprisingly controversial

Chemical shifts (including H1 shifts) might be useful for cross-validation, these already can be used to ID wrong assignments

Brian Smith ("The HP-1 Chromo-Shadow Domain Homodimer")

This speaker discussed his experiences with the determination of the protein mentioned in the title. Some key lessons from this experience is that it is important to evaluate the protein, both by methods such as analytical ultra-centrifugation and by such NMR techniques as T_2 measurements before getting too involved in structural determination. Such measurements allow one to know what to expect, e.g. in terms of the dimerization state of the protein, in the structural determination process. Additional ideas that come out of this talk are the need to have an evaluation of problem areas in alignment of peaks and that one should also re-filter, re-center the results of automated peak-picking. I would say that the people at the conference seemed to be enamored with restricted peak-picking. One interesting question remaining in my mind from this talk is: why does, without direct adjustment otherwise, energy minimization not do much to improve Ramachandran plot?

Annalisa Pastore (EF Hand Proteins)

This talk concerned, in large part, what the speaker termed “the role of bioinformatics”. She defined this role as that of predicting, defining and suggesting what is expected. The speaker began by listing the steps of structural determination and identifying what she felt was the bottleneck phase: (1) Sequence Analysis, (2) Sample Preparation (Rate Limiting Step), (3) Data detection, (4) Structure Calculations (Tedious but not bottleneck), (5) Structure analysis (Fun step, not a bottleneck phase).

As in illustrative problem, Dr. Pastore posed the question of where to cut a protein. The presence of the same building block in different architectures is taken as an indication that the module is also an independent folding domain. This assumption (which is, for each case, a *hypothesis to be proven*) is correct about 50% of the time. Examples of non-domains include the KH module, the WW module and the Dep module. Knowledge of which building blocks are actually domains is critical in producing structures - this knowledge, presumably may eventually be obtained by bio-informatic methods (e.g. “Domain Parsing”), but now much trial and error is required - evaluation is by making constructs and testing, e.g., whether folded.

Frank Delaglio (Talos and Pales)

This speaker discussed both Talos and Pales. The former is a (cross validated) procedure for predicting backbone angles. The latter predicts alignment tensors from structure.

Talos has no prediction about 35% of the time and the prediction it does make is wrong about 3% of the time. Talos also has a certain uncertainty in the angles it can predict accurately. An interesting idea that Dr. Delaglio gave is that the database used in Talos has also been used to build surfaces in phi/psi space to make chemical shift prediction.

The speaker then proceeded to talk about dipolar couplings and Pales. He described a strategy for using homology as homology implies compatible chemical shifts and dipolar couplings. His idea is to use the converse of the above statement to generate an initial fold. The strategy seems to be based on linear regression. I think a Bayesian perspective would be useful for all of this speaker's ideas.

Andy Pickford (Extracellular Matrix Proteins)

This speaker primarily talked about the software he has found useful in his work. He discussed NMRView which uses NMRStar format and is easily adaptable and extendible with tcl as well as X-Plore and CNS, both of which can handle ambiguously assigned NOEs. Dr. Pickford also defined the concept of “ambiguity level” which is the square root of the number of ambiguous combinations at some stage of NOE assignment. Finally, the speaker mentioned a new procedure for dealing with pro-choral groups: SOPHIE.

Helen Mott (Structure of Protein Complexes)

This speaker gave an overview of some particular challenges associated with NMR based structural work on protein complexes. According to the speaker, important considerations when working with complexes include the need for multiple, concentrated samples (which, especially for a complex) can be hard to obtain, the lack of sensitivity of the experiments which are useful to analyze complexes and specific problems with small peptides. The experiments involved in understanding the structure of protein complexes include all of the standard experiments for resonance assignment and also various NOE experiments. The NOE experiments may be classified as “mixed”, “intermolecular” and “doubly rejected”. Some recommendations were made as well. For example, it is better to have C13 editing on the larger component and assigning part(s) of the complex separately before attempting the whole complex. The importance of ambiguous restraint analysis (cf. Nilges paper) was also mentioned.

Marc-Andre Delsuc (Gifa, Residue Typing and Fold-recognition)

This speaker discussed the protein assignment module in Gifa and various methods of residue typing and fold recognition. The protein assignment module in Gifa consists of a macro that allows Gifa to help with assignment, a database for spins and for spin systems and a topology file for residues. The typical project involves peak picking with the internal picker, assignment of TOCSY peaks with graphical, assignment of additional NOEs, calculation of distances and building assigned strip plots. Extract features include the ability to superimpose spectra, no limitations on file sizes and relaxation analysis.

One approach to assignments is the Rescue algorithm of Pons, et al. (1999), which is a neural net based approach to calculating assignments. This approach computes reliabilities for its assignments which correlate well to the quality of the final structure. Dr. Delsuc also discussed the Fire algorithm which extracts geometric information without any assignment step. This method correlates strips associated with peaks in HSQC to build a similarity matrix. The result can be used to explain a structure, but one wishes to be able to use this algorithm to predict structures: the problem is peak overlap (although maybe we can solve this problem with a good multiple registration algorithm?) - also, since we do not know the order of the peaks, it is hard to build a structure.

The final focus of this talk was on the Fireman algorithm which “rescue”s the difficulties of Fire in order to use the latter as a fold production tool. The principle idea of this program is to order peaks so that the resulting similarity matrix as produced by Fire, the “fold matrix”, is like a contact map. Essentially this algorithm builds a rough assignment. Fireman depends on the ability to use various ideas of a how a good ordering looks to optimize the ordering of the peaks. In this case, a heuristic approach seems to work best. It turns out that errors are more prevalent in secondary structure regions.

The talk concluded with a discussion of future directions. The principle need to be addressed is “molecular replacement”: if you have a homologue, you have some idea what a contact map should look like, so you should be able to use a modification of Fireman for molecular replacement.

Yutaka Ito (“Problems in iterative assignment and structure determination [...]”)

This speaker discussed his experience with the determination of the structure of DinI. Dr. Ito began with a discussion of the biology of DinI which led to the questions to be addressed by knowing something about the structure of DinI. A particular question of interest was whether DinI inhibited the ssDNA binding activity of RecA. From transferred NOESY experiments, it appeared that DinI had no effect on RecA DNA binding - a result which disagrees with the results of Voloshin, et al, *Genes Dev.*, 15:415-427. The problem with this experiment is that the ssDNA used in the TRNOE may not have been sufficiently long from the experiment to detect an effect of RecA on DNA binding.

The actual structure calculation of DinI was performed twice at RIKEN (as well as concurrently at NIH). The first trial consisted of complete side-chain assignment and some initial NOE assignments that were completed by an iterative NOE assignment procedure. The result was a structure that, when compared with the NIH structure, is wrong. Problems that may have contributed to the calculation of a wrong structure include poor separation of side-chain methyl resonances, low digital resolution in indirect dimension and, perhaps most importantly, the incorrect setting of the C13 offset and C13 CPD which can lead to difficulty in assignment of aromatic related NOEs. The second trial, which resulted in a reasonable structure involved the use of Aria and floating chirality, but did not use iterative assignment. Instead, the Azara connect command was used.

At this point, the focus shifted from answering questions about the biology of DinI to the question of distinguishing wrong structures from correct ones. For this purpose, measures of the violation of constraints are not too useful as in iterative assignment the violations are removed as noise. The calculation of H-alpha chemical shifts seems possibly useful in theory, but in this case, the first ensemble would not seem wrong on the basis of its H-alpha shifts. Also, there are no obvious localizations of shift differences to problematic regions of structure. The discussion of iterated assignment concluded with Dr. Ito's view of the risk of automated analysis: at an initial stage, relative amount of unambiguous restraints are limited. Therefore, strong noise peaks can affect structures.

Michael Sattler ("Practical Aspects of NMR Structural Determination")

This talk centered on subject areas relating to NMR structure determination: sample preparation, time summaries, structure calculations with ARIA/CNS and cross-correlated relaxation experiments.

According to Dr. Sattler, the important aspects of sample preparation are the definition of domains, considerations relating to tags, purification, evaluation of the sample (by 1-D NMR or HSQC) and further sample characterization. Relating to tags and purification, everyone seems to like TEV-protease based systems for tag cleavage as, evidently, the protease activity is removed with relative ease. Recommendations for further sample characterization included a quick H1-N15 NOESY to verify that there are not too many flexible residues, analytical ultra-centrifugation, gel filtration and CD measurements. Also noted at this point in the talk was the fact that the existence of mutation induced chemical shift perturbations do not necessarily imply that the mutation has significantly modified the protein structure. Additionally, the importance of water suppression with isotope filtered NOE data and the need to take BSP-shifts into account with off-resonance pulses were mentioned. The latter concern may be addressed by either amplitude modulation or better an additional pulse.

The time summaries given were that backbone assignment took two weeks as did side-chain assignment. The remaining time to get the structure should be around 4-8 weeks. I forget the details of this time summary, however.

For the structure calculation with ARIA/CNS, the speaker particularly mentioned the use of molecular dynamics and simulated annealing, but as a guide to making manual assignments. Also possible with this package is a reduced relaxation matrix approach as well as distance recalibration and peak exclusion. The latter two functionalities are "dangerous" as they involve fudge factors. Dr. Sattler used a mixed strategy for stereospecific assignments.

Cross-correlated relaxation experiments (Griesinger, et. al, 1997; Kay, et. al., 1997) are useful to measure the psi (?) dihedral angle (The results may be compared with Talos.). The question is whether such data are all that useful in structure determination. Also, one needs to have the correlation times to use this approach. The speaker recommended that, if psi angles are to be used in structural refinement, they should be treated like J-coupling data and that force fields should be modified as necessary.

The speaker also made a general comment on residual dipolar coupling data: these make most sense to use when there is more than one "domain", although they can also be used to "tweak" even well-defined structures.