CHEMCOS
THE CHEMICAL SOCIETY, IIT DELHI
November 2008
Issue III

Searching for needles in haystacks: determining conformations of flexible, bioactive molecules in solution

Ashutosh S. Jogalekar,

Department of Chemistry, Emory University, Atlanta, USA.

The basis of all life is molecular motion. The complex organic and inorganic structures that serve as fuels, building blocks, signaling agents and genetic messengers in biological systems are dynamic assemblies that constantly interact with each other and with the fluid environment in which they are immersed. If we want to understand the behaviour of these molecules and extract useful information from that behaviour that can be converted into practical knowledge such as medical therapies, we need to understand both their structure and dynamics.

A hundred years of development in physics, chemistry and instrumentation have provided us with exquisite tools to study both structure and dynamics of molecules. In the early days of chemistry, structure determination dependend on arduous chemical detective work in which a molecule of unknown structure would be broken down into fragments using chemical methods. It would then take an impressive amount of detective work to mentally reassemble these fragments and deduce their sum total. In the twentieth century, a remarkable set of tools developed initially by physicists converged on chemistry and biology to provide a wealth of structural information. Out of all these methods, two stand out for their lasting importance. The first one, x–ray crystallography, developed by Max von Laue, the Bragg duo of father and son and others provides detailed atomic-level resolution of molecules in the solid state. This technique was first used for small organic molecules, after which a series of post-world war II breakthroughs enabled scientists to apply it to proteins. This development was extremely important and that importance continues to this day. However, molecules in the solid-state are static and we already mentioned their dynamic nature. While x–ray crystal structures are usually a fair representation of molecular structures in living systems, a technique that determined their structure in solution would be of inestimable importance. Fortunately, such a technique was developed in the 1950s- NMR spectroscopy.

Ask any organic chemist what's the most important technique available to him for the study of structure and dynamics and he or she will almost certainly name NMR spectroscopy, a technique that distinguishes atoms in different environments based on their magnetic properties. In its early days, NMR spectroscopy made huge contributions to the routine structure determination of organic molecules. Today no organic synthesis or natural products study proceeds without NMR spectroscopy. Stereochemistry is powerfully determined by NMR for example. In the last two decades NMR has also been used to determine the structure of proteins, developments that led to the receipt of the Nobel Prize to Kurt Wüthrich in 2002. Complemented with x–ray diffraction, NMR provides a set of tools that has allowed unprecedented insight into the behaviour of molecules.

What distinguishes NMR spectroscopy from many other techniques is that while it is very useful for determining configuration of chiral centers in molecules, it is also very useful for determining conformation. A molecule can exist in several conformations in solution. The relative proportions of these conformations, usually referred to as a Boltzmann population distribution, are determined by their relative free energies. Determination of configuration has always been important for organic chemists, but determination of conformation has special importance for insights into the behavior of flexible molecules, especially those that are used as drugs. Let us see why that is so.

All drugs with few exceptions are small organic molecules of molecular weight less than 500 Daltons that can have several rotatable bonds. Recall that about a single rotatable bond there are three dominant conformations; two gauche and one anti as shown in Figure I below.

CHEMCOS
THE CHEMICAL SOCIETY, IIT DELHI
November 2008
Issue III

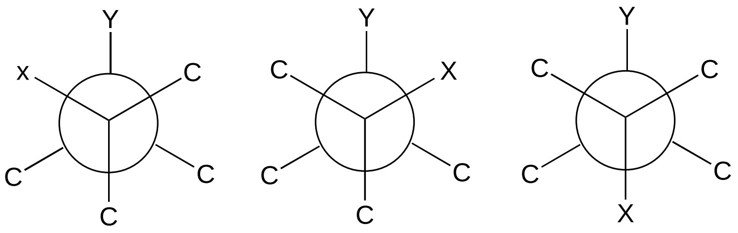

Figure I: The three dominant conformations of two substituents X and Y around a single bond: from left to right, gauche−, gauche+ and anti

Thus, since one bond has three possible conformations, a molecule with say ten rotatable bonds will have a maximum of 3^{10} or about 60,000 conformations, a huge number. Many of these conformations may be forbidden by high-energy steric clashes between atoms, but still there will be a considerable number in solution.

Most drugs are flexible molecules that bind to proteins and modulate their activity. Many of the best selling drugs on the market today such as the ubiquitous aspirin and the cholesterol–lowering atorvastatin, for example, bind to proteins or enzymes and regulate their activity. When a flexible molecule binds to a protein, it binds in a single conformation. However, the protein has to pay an energy penalty for converting the conformations of the molecule in solution to the bound conformation. This penalty is largely entropic since the molecule loses its degrees of freedom when it binds to the protein.

Determining this protein-bound conformation is very important for drug design and is a much–pursued goal in pharmaceutical science. For example, knowing what the bound conformation is, organic chemists can make modifications to the structure to 'lock' the molecule into that conformation and thus avoid paying the entropic penalty, forcing the molecule to be already constrained into the favourable conformation before it binds. If one could have an x–ray crystal structure of every drug bound to every important protein, the problem would be resolved. But unfortunately protein crystallisation is still an art and it has been very difficult to determine x–ray structures for some of the most important proteins bound to their corresponding ligands. Now it has been seen several times that the conformation that actually binds to the protein- termed the 'bioactive' conformation- is usually identical to or close to one among the several in solution. Thus, it is clearly valuable to have a list of the conformations of a druglike flexible molecule in solution. Note that there doesn't have to be a correlation between the population of the conformation in solution and the bound conformation. For example, the bound conformation can be present to the extent of only 2–5% in solution; as long it's a part of the conformational equilibrium, the protein can extract it out of solution.

The problem then boils down to determining the population of flexible molecules in solution. This is admittedly quite difficult. To see why this is so, it is important to understand that the relative populations of equilibrating conformations in solution are governed by their relative free energies in solution. There is a relation between free energy and equilibrium constant that is all-important in this situation; $\Delta G= -RT \ln K$, where K is the equilibrium constant, ΔG is the free energy difference, R is the gas constant and T is the temperature. A look at the equation tells us that since K is exponentially dependent on ΔG, a small difference in ΔG makes a huge difference in K. For example consider the example of a molecule that can exist in only two conformations in solution. If the free energy difference between them is only 1.8 kcal/mol (compare this to the strength of a typical carbon-carbon bond; 83 kcal/mol), the lower energy conformer will be about 97% while the higher energy one will be just 3%! Hike up the energy difference to 3 kcal/mol and the higher-energy one will be virtually non-existent, with the other one about 99.96%. If there are several conformations, one can clearly imagine an ensemble of 10–15 conformations, if anything an underestimate in case of highly flexible molecules, in which even the "dominant" conformation would be about 20%, and several other conformations could range from 2-5%, with the energy differences between all these being tiny. So clearly the energy window that includes the several conformations of a molecule in solution is very small. Since the energy provided by the environment at room temperature is much more than this, a flexible molecule will have conformations that will very rapidly interconvert at room temperature.

Going down to low temperatures could possibly 'freeze' some of these populations out, but one would need to go down to an impossibly low temperature to take care of such small energy barriers. Can NMR parse this small energy window and extract the individual conformations?

Unfortunately not, and we will shortly see why. NMR provides two very important variables that relate to conformation. One is the proton-proton coupling constant, 3JH-H. The coupling constant relates to conformation through the valuable Karplus equation; plug the coupling constants into the equation and one derives the dihedral angles about bonds and therefore the conformation. Another important variable is the distances between protons. One can acquire them through a powerful technique with the colourful name NOESY (Nuclear Overhauser Effect SpectroscopY). Together, the dihedral angles and the interproton distances define the conformations of the molecule.

But there is a fundamental problem here. As noted above, conformations at room temperature rapidly interconvert between each other. The time scale on which NMR spectroscopy operates is on the order of tens of milliseconds, while conformational interconversion is much faster. Trying to gauge the conformations using NMR is like trying to make out individual blades of a fan through your eye when it is moving very fast, or trying to photograph a horse race with a camera with a low shutter speed. In fact the former is a good analogy: what do you see when you look at a rapidly moving fan? You see a disk instead of individual blades because the rotation speed of the blades is much more than the resolution time of your eye. Thus, the disk you see is an average of all the motions. Similarly, when we use NMR spectroscopy to determine coupling constants and interproton distances, we get average values for these. Now, is the disk that you see for a rapidly rotating fan a real structure? No. It is non-existent since it is an average structure. Similarly, any structure that one might assign to the average coupling constants and interproton distances from NMR for a flexible molecule will be an average and therefore non-existent, often called a 'virtual' structure. As a simple example, consider two protons that exist in two conformations as shown below in Figure II. In one they are apart and in another one they are close. They rapidly interconvert between the two orientations. Clearly the average is the intermediate structure which does not exist in reality.

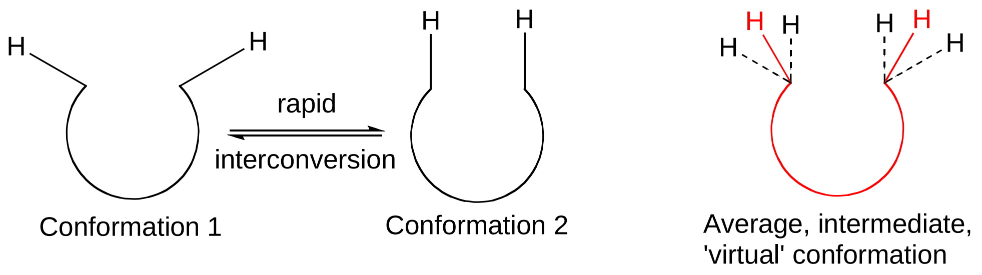

Figure II: A pair of protons existing in two dominant, rapidly interconverting conformations. The intermediate conformation is an average and is virtual.

This is a very important point. NMR is extremely valuable for getting information for rapidly interconverting conformations in solution. But this information is an average for all those conformations, and assigning a single structure to this average data is inherently flawed. In this sense it is a misnomer to refer to the singular 'conformation' for a flexible molecule in solution; one should always talk about 'conformations'. In addition, the detection limit of NMR precludes it from determining conformations that are less than about 2% in solution; the data is overwhelmed by other dominant conformations. What if a 2% conformat-ion is the one binding to the protein? The problem is equivalent to finding a needle in a haystack. Clearly NMR would be useless in finding it. Or would it?

Fortunately there is a way out. Computational advances in the last three decades have provided several theoretical methods for investigating conformations of organic molecules. One of the most important methods is called molecular mechanics, which uses collections of equations and experimentally derived parameters called force fields to investigate the potential energy surface of a molecule. A typical force field will have equations for bond stretching, angle bending, dihedral rotations, electrostatic interactions and Van der

Waals interactions. The equations are based on classical mechanics and treat the atoms and bonds as balls as springs. Because this picture is approximate, experimental parameters are used to augment the equations and make them accurately reproduce geometries and energies of complex organic molecules. Using force fields, one can do a conformational search of a flexible molecule. This is nothing more than exploring the conformational potential energy surface and finding all possible low-energy conformations of a molecule. In practice it is quite challenging to find all possible conformations, but most often using a good energy window to capture many conformations and using a large number of search steps, one can do reasonably well. After a conformational search then, one ends up with a large number of conformations, usually numbering in the thousands.

The key point is that one can now combine this set of conformations with the average NMR data above using a novel method called NAMFIS to finally get a glimpse of the coveted conformational ensemble in solution. NAMFIS stands for NMR Analysis of Molecular Flexibility In Solution. It was developed by researchers at the Istituto di Ricerche di Biologia Molecolare (IRBM) at Pomezia in Italy and transformed into a workable method by researchers at Emory University in Atlanta in the United States [1]. NAMFIS takes as its input two sets of variables; one is the set of average coupling constants and average distances from NMR, and the other is the set of conformations computed through the conformational search. NAMFIS can calculate the distances and coupling constants for all the theoretical structures. What it then does is to vary the mole fractions of the theoretical conformations, calculate the resulting coupling constants and distances, and compare these weighted parameters to the corresponding experimental data. In effect, by varying the mole fractions, NAMFIS simulates different conformational populations in solution and then compares the resulting calculated parameters to the average experimental parameters. The fit of the two is expressed as Sum of Squares Differences (SSD), a standard measure of fits between sets of data. Clearly the 'best-fit' solution would be the one for which this number is the smallest. The accompanying mole fractions would then represent the ideal ensemble of conformations in solution. Note that there several combinations of conformations that could fit the data more or less equally well. NAMFIS chooses the one that does the best job. In any case, NAMFIS derives a whole family of conformations that together satisfy the average NMR data, a much realistic proposition than assigning all the data to a single conformation.

NAMFIS is a powerful tool because it can tease out individual conformations from average NMR data; something that we have seen cannot be accomplished by NMR alone. Another significant advantage of NAMFIS is that it derives a relatively small number of structures; 10-20 in all the cases studied until now, compared to the theoretical thousands. It has been used to provide Boltzmann populations for many important molecules and shed light on their bioactive conformations. Let us now look at some past and current applications of NAMFIS.

1. NAMFIS analysis was used to probe the conformations of Taxol. This molecule is one of the most important molecules in cancer therapy and kills cancer cells by binding to the important protein tubulin and disrupting cell dvision. The structure of the bound conformation of Taxol in the tubulin binding pocket was elucidated using electron crystallography, a technique similar to x-ray diffraction that uses electrons. The bioactive conformation of Taxol turned out to be one that was present to the extent of only 2% in solution (Figure III) [2]. This fact underscores the value of NAMFIS in finding minor conformations in solution that would not be found by NMR. Similar NAMFIS analysis was also applied to another anticancer molecule binding to tubulin called epothilone.

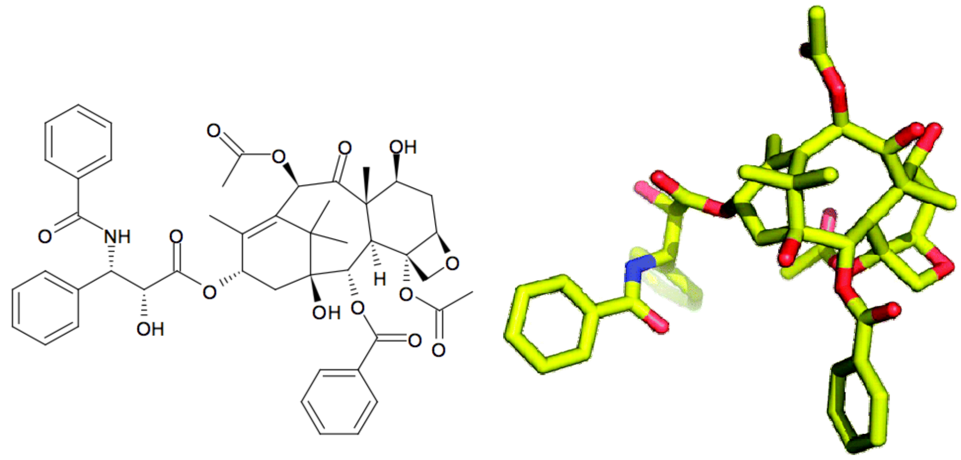

Figure III: 2D representation of Taxol (left) and the tubulin-bound conformation of Taxol called T-Taxol (right) which is a minor conformation in solution as found by NAMFIS. The figure to the right is from Reference [2] while that on the left is from Wikipedia (http://en.wikipedia.org/wiki/Taxol)

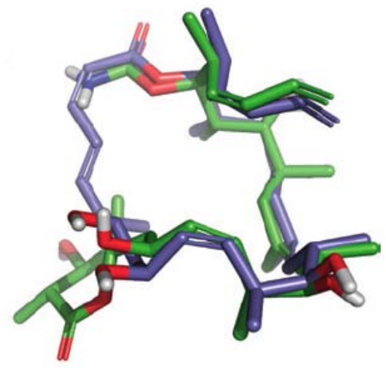

Figure IV: The proposed overlap of a dominant conformation of dictyostat-in (purple) and discodermolide (green). The figure is from Reference [4]

2. In another interesting application, NAMFIS was used to find out the family of solution conformations for a 5-residue peptide that was claimed to form an alpha helix in solution. Confidence in the existence of the helix came from the observation of characteristic signals in the NMR data and the fact that the two ends of the peptide were tethered together by a metal to apparently constrain the peptide in a helical conformation. NAMFIS analysis indicated however that the NMR data could be much better satisfied by a collection of conformations that did not include the alpha helix even as a minor conformation [3]. The main conclusion from this study was that average NMR data may give the illusion of a dominant conformation, and secondly that even constraining a molecule may not actually lock it in a given conformation.

3. In current studies, I am using NAMFIS to investigate the conformations of two other important tubulin-binding anti-cancer molecules, discodermolide and dictyostatin. These molecules show very similar behaviour and potency towards cancer cells, and are therefore thought to bind 'similarly' to the protein [4]. However, especially when it comes to molecules binding to the same protein, 'similarity' is in the mind of the beholder. There are many ways in which one can overlap molecules on top of each other. Which one of these overlaps indicates the greatest similarity? In case of discodermolide and dictyostatin, dominant conformations in solution corresponding to the average NMR data were derived and overlaid on each other (Figure IV). But our NAMFIS analysis indicates conformations that are different from previous ones [5]. Clearly the overlap cannot be then correct if the conformations are not present in solution in the first place.

As seen from the case studies above, not only does NAMFIS allow us to represent the NMR data much better with a set of multiple conformations, but it also can be insightful in correcting or modifying previous conclusions. It's a fast method; most calculations take not more than several minutes. Also note that in many cases, the bioactive conformation deduced through other techniques was very similar to a minor conformation in solution found by NAMFIS, and not the dominant conformation in solution. As seen above, it would not be possible to find this conformation using NMR alone.

There is another valuable function that NAMFIS performs. One can ask whether the protein-bound bioactive conformation is truly present in solution. While there is no reason to necessarily assume this, in many past studies this has been so. But even if this is not the case, the conformations from NAMFIS provide a tractable set of structures for exploring the active site of the protein as indicated below.

There are many cases in which the bioactive conformation is not known because of problems in crystallizing the protein-ligand complex for example. To try to circumvent this problem, one very common computational technique used in drug design today is molecular docking. In this technique, a program places or 'docks' thousands of conformations for a bioactive ligand generated by a conformational search in the active site. It then twists and turns them and determines the interaction of each conformation with the atoms in the protein; this includes interactions such as hydrogen bonding, electrostatic interactions, and hydrophobic interactions. The program then calculates the sum total for the favourable and unfavourable interactions for each conformation and assigns a number called the 'score' which denotes how well the conformation binds. It then ranks the scores from best to worst, and the best scoring conformation is supposed to be the bioactive conformation of the ligand.

It would be extremely valuable if such fast theoretical methods could provide bioactive conformations without doing a structure determination experiment. In practice, both the conformations from a conformational search and the scores determined by docking them are notoriously fickle. Conformational searches for example can generate spurious high-energy structures and docking programs may sometimes score these high-energy structures favourably, leading to false positives. One can also have low-energy structures that can score unfavourably, leading to false negatives. In general there is always a problem assigning theoretically generated structures as 'low' or 'high' energy since this assignment depends on the particular method used, and different methods give different results. But NAMFIS significantly bypasses this problem by deriving a relatively small set of bonafide experimental low-energy structures that can serve as docking candidates. Since the structures satisfy the NMR data, they have to be low-energy by definition otherwise they would not exist in solution. And of course the complexity of the problem is reduced by a factor of perhaps a hundred since NAMFIS delivers a dozen or so structures compared to the thousands generated from a conformational search.

There will undoubtedly be future applications of NAMFIS in determining conformations of medicinally important molecules in solution and generating hypotheses for bioactive conformations. One drawback of NAMFIS deals not with the method itself but with the data acquisition for it. Doing a conformational search is easy and does not take much time. But the other part of the NAMFIS input, acquiring accurate NOESY data from NMR studies and extracting distances from it, is a non-trivial exercise within the domain of skillful NMR operators. Nonethless, the efforts spent in this data acquisition are well worth the wealth of conformational data that one can unearth for the purposes of understanding the behaviour of molecules in both solution and inside protein active sites. NAMFIS is a reminder of the power of both NMR spectroscopy and modern computational chemistry to shed light on molecular structures and their behaviour in living systems. We can be assured that such techniques will increasingly continue to aid us in fundamental understanding of molecules and their involvement in health and disease.

Ashutosh Jogalekar is a graduate student in the Department of Chemistry at Emory University, Atlanta, USA. His interests include conformational analysis, structure-based drug design and other aspects of molecular modeling.

References:

1. Cicero, D. O., Barbato. G and Bazzo. R. *JACS*, **1995**, *117*, 1027-1033
2. Kingston, D. G. I. *J. Org. Chem*. **2008**, *73*, 3975-3984
3. James P. Snyder, Ami S. Lakdawala and Michael J. Kelso, *JACS*, **2003**, *125*, 632-633
4. Ian Paterson, Nicola M. Gardner, Karine G. Poullenec and Amy E. Wright. *J. Nat. Prod.* **2008**, *71*, 364-369
5. Ashutosh S. Jogalekar, Damodaran Krishnan, Won-Hyuk Jung, Shi Zhong, Dennis P. Curran and James P. Snyder (manuscript in preparation)