Accepted Article

Title: Deriving ligand orientation on weak protein-ligand complexes by DEEP-STD NMR in the absence of protein chemical shift assignment

Authors: Ridvan Nepravishta, Samuel Walpole, Louise Tailford, Nathalie Juge, and Jesus Angulo

This manuscript has been accepted after peer review and appears as an Accepted Article online prior to editing, proofing, and formal publication of the final Version of Record (VoR). This work is currently citable by using the Digital Object Identifier (DOI) given below. The VoR will be published online in Early View as soon as possible and may be different to this Accepted Article as a result of editing. Readers should obtain the VoR from the journal website shown below when it is published to ensure accuracy of information. The authors are responsible for the content of this Accepted Article.

COMMUNICATION

Deriving ligand orientation on weak protein-ligand complexes by DEEP-STD NMR in the absence of protein chemical shift assignment

Ridvan Nepravishta,[†][a] Samuel Walpole, [†][a] Louise Tailford,[b] Nathalie Juge[b] and Jesus Angulo*[a]

DEEP-STD NMR is a recently developed powerful approach for structure and pharmacophore elucidation of weak protein-ligand interactions, reporting key information on the orientation of the ligand and the architecture of the binding pocket.[1] The method relies on selective saturation of protein residues in the binding site and the generation of a *differential epitope map* by observing the ligand, which depicts the nature of the protein residues contacting the ligand in the bound state. Selective saturation requires knowledge of the chemical shift assignment of the protein residues, which can be obtained either experimentally by NMR or predicted from 3D structures. Here we propose a simple experimental procedure to expand the DEEP-STD NMR methodology to protein-ligand cases where the spectral assignment of the protein is not available. This is achieved by experimentally identifying the chemical shifts of the residues present in binding hot-spots on the surface of the receptor protein using 2D NMR experiments combined with addition of a paramagnetic probe.

The 3D structure of a small bioactive molecule in complex with its receptor gives important atomic information essential for understanding biological effects triggered by biomolecular recognition processes, as well as for the discovery and design of new drugs. There are several techniques used to achieve this aim, with X-ray crystallography, NMR spectroscopy and more recently cryo-EM being the most relevant approaches.

Recently, we have developed the *DiffErential EPitope* mapping saturation transfer difference (DEEP-STD) NMR methodology for weak protein-ligand interactions,[1] as an extension of the general STD NMR method.[2] The DEEP-STD NMR technique allows the orientation of the ligand to be derived, through differential selective saturation of different sets of key protein residues in the binding site. Namely, two STD NMR experiments are carried out, each one saturating different sets of protein residues, and the difference between the resulting spectra is quantified and mapped onto the ligand structure (differential epitope map). In order to accurately perform the DEEP-STD NMR experiment, it is of paramount importance to know beforehand the chemical shifts of the residues present in the binding site, in order to identify which set of residues to target, i.e. choosing the irradiation frequencies to ensure that the

selective saturation is applied on residues that are present in the binding site. Experimental chemical shifts can be obtained by NMR, or derived from a 3D structure obtained by X-ray diffraction or by homology modelling.[3] The experimental DEEP-STD factors can be further combined with molecular docking and STD intensity predictions by CORCEMA-ST[4] in order to select the docking model that best fits the experimental data.

In the absence of chemical shift assignment of the receptor protein, we hereby propose a general approach to experimentally identify the chemical shifts of those binding pocket resonances, which relies on the identification of ligand binding hot-spots on the surface of the protein (a ligand-binding hot-spot is a site on the surface of the protein that has a high probability for interaction with a ligand[5]) using 2D NMR spectroscopy. This approach is compatible with the STD NMR technique, inexpensive, and relatively fast which should allow broad applicability.

For large proteins, the slow molecular tumbling in solution is characterized by an overall correlation time expected to be in the range of 10^{-8} seconds. However, the internal correlation time of the surface residues of globular proteins may be significantly shorter. As a result, residues in the core of the protein follow a slow-motion regime due to their low flexibility, causing the signals to become broadened beyond detection. Conversely, the greater flexibility of surface residues causes them to follow a fast motion regime leading to cross-peaks in 2D NMR experiments that will be narrower and hence detectable. Therefore, these spectra are more likely to display signals from residues exposed on the surface or in very flexible regions of the protein.[6]

Using 2D ^{1}H-^{1}H TOCSY experiments, hot-spot mapping can be readily achieved by adding paramagnetic probes to the protein sample such as 4-hydroxy-2,2,6,6-tetramethylpiperidin-1-oxyl (TEMPOL).[7] A decrease in intensity of specific protein TOCSY cross peaks, compared with the spectra recorded without the paramagnetic probe, allows those residues interacting with the probe to be easily identified, since they are affected by the paramagnetic relaxation enhancement (PRE) effect.[8,9] Previous PRE studies with TEMPOL have demonstrated the greater accessibility of this paramagnetic probe to protein specific binding sites rather than surface regions.[7,10,11] As a result of these experiments, the identified hot-spot resonances will then be considered as input frequencies for the DEEP-STD NMR experiments.

The identification of binding hot-spots on the surface of proteins has previously been achieved by NMR spectroscopy using paramagnetic probes along with classical molecular dynamics as well as Mixed Molecular Dynamics using 5-50% probe-water mixtures.[7,12,13] For the development of our protocol, we combined this approach with DEEP-STD NMR using the structurally-characterized catalytic domain (belonging to glycoside hydrolase family 33, GH33) of the intramolecular *trans*-sialidase (IT-sialidase) from the human gut symbiont *Ruminococcus gnavus*, *Rg*NanH-GH33 in complex with 2,7-

[a] Dr Ridvan Nepravishta, Mr Samuel Walpole, Dr Jesus Angulo
School of Pharmacy, University of East Anglia
Norwich Research Park, NR47TJ, Norwich, UK
E-mail: j.angulo@uea.ac.uk
† These authors contributed equally to this work
[b] Dr Louise E. Tailford, Dr Nathalie Juge
The Gut Microbes & Health Institute Strategic Program, Quadram Institute of Bioscience
NR47UA Norwich Research Park, Norwich, UK
Supporting information for this article is given via a link at the end of the document.

anhydro-Neu5Ac (PDB code 4X4A) as a benchmark.[14] The *Rg*NanH-GH33 enzyme is a 489 residue domain which can be considered out of the typical range for swift assignment and structure determination by NMR spectroscopy and was previously used for the development of the DEEP-STD NMR approach.[1]

We first performed 2D homonuclear ^{1}H-^{1}H TOCSY experiments on *Rg*NanH-GH33. The protein was exchanged in 10 mM Tris-d$_{11}$ D$_2$O buffer pH 7.8 and 100 mM NaCl and used at a concentration of 1.2 mM. First, a 2D ^{1}H-^{1}H TOCSY reference spectrum of the protein was acquired, which was followed by two spectra in the presence of 2 mM and 12 mM of the paramagnetic probe TEMPOL. The spectra (Figure 1 a, b, c) obtained in the absence and in the presence of increasing concentrations of TEMPOL showed that the probe selectively interacts with some residues of the protein, as only some resonances in the spectra were significantly affected, as seen by a decreased intensity. The chemical shifts most affected by the presence of TEMPOL were 0.6, 0.74, 1.06, 1.15, 1.26, 6.6, 6.74, 7.04, 7.57, 8.56 ppm (Supporting Information Figure S1). These resonances, although lacking a specific assignment, are typical of aliphatic and aromatic amino acids, while we can exclude the presence of the NH resonances in the spectra due to the fact that the protein was solvated in a D$_2$O buffer. The identified resonances from the TEMPOL-attenuated TOCSY spectra of *Rg*NanH-GH33 were indeed in very good agreement with the predicted chemical shifts of key aliphatic and aromatic residues in the binding pocket of the enzyme (Ile-258, Ile-338, Val-502, Thr-557, Tyr-525, Tyr-677, and Trp-698).[1]

To further validate our approach and exclude false positives (i.e., binding hot-spots outside the binding site), we carried out molecular dynamics (MD) simulations, an approach successfully used in the past to identify ligand binding pockets for the development of small molecule inhibitors.[15] Here, MD simulations were used to confirm the accessibility of TEMPOL to the specific binding pocket of *Rg*NanH-GH33. To efficiently explore the configurational space of the *Rg*NanH-GH33-TEMPOL system, three different MD approaches were considered: (i) long MD (1.0 µs) with low concentration of TEMPOL (10 mM) starting from a random configuration of the system, (ii) 16 independent replicas of short MD (0.8 µs total 10 mM of TEMPOL), and (iii) 16 independent short replica MD simulations with a high concentration (50% w/w) of TEMPOL in water known as the MixMD approach.[12]

In each case, we first analyzed the backbone RMSD of *Rg*NanH-GH33 for each trajectory and showed that the presence of TEMPOL did not affect the structure of the protein, even with MixMD, in which the solvent contained 50% w/w TEMPOL, as the average backbone RMSD was only approximately 1 Å (see Supporting Information). In the case of the long MD and the 16 short replicas, in which there were relatively few molecules of TEMPOL in the simulation, the interaction between TEMPOL and *Rg*NanH-GH33 was analysed by computing the contacts between TEMPOL and each residue in *Rg*NanH-GH33 over the course of each trajectory, in order to construct a fractional occupancy map for each residue. In the case of MixMD, in which there were 381 molecules of TEMPOL in the bounding box, the occupancy was measured by using a 0.5 Å grid to create bins for each TEMPOL molecule in each

frame of the trajectory. The resulting 3D histograms were then visualized by means of the isomesh feature in PyMOL,[16] using
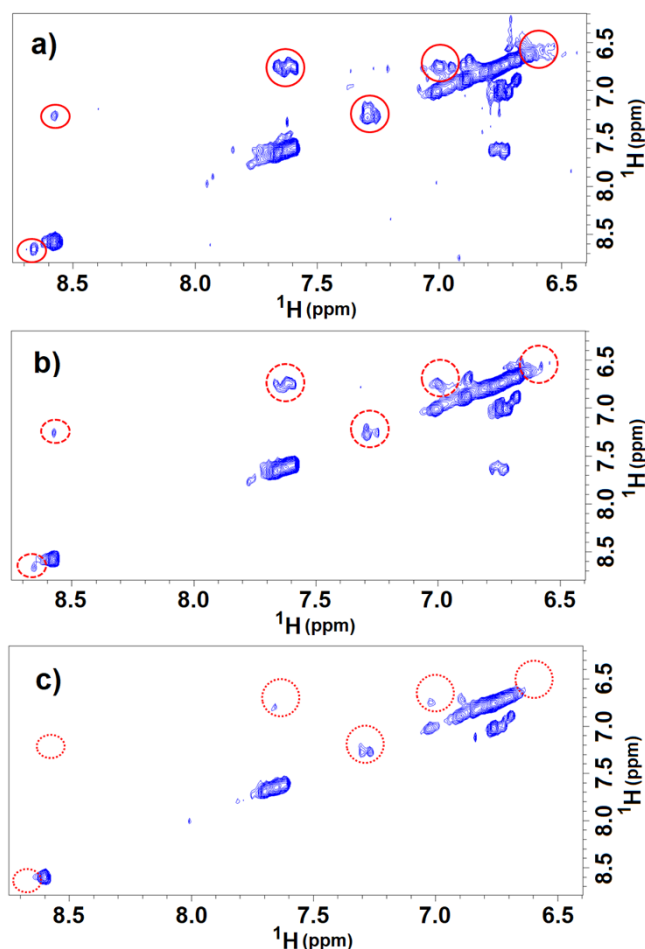

Figure 1. Expansion of the aromatic spectral region of the TOCSY spectra of *Rg*NanH-GH33 1.2 mM in the absence **a)**, in the presence of 2 mM **b)**, and 12 mM **c)**, of the paramagnetic probe TEMPOL. Red circles highlight some resonances affected by the presence of TEMPOL (see also Supporting Information Figure S1).

a structure averaged over the whole trajectory for RgNanH-GH33 (Figure 2).

Firstly, the long MD simulation containing a low concentration of TEMPOL did map several binding hot-spots including the area of the binding site, but the outcome was dependent on the starting coordinates of the system. To overcome this issue, the same experiment was repeated with 16 different independent short replicas of 50 ns each, following a previous protocol.[7] In this case, although mapping of the binding hot-spots was clear, the extension of sampling of the surface was not complete (Supporting Information Figure S5). Using the MixMD approach, high concentrations of the probe enabled most of the protein surface to be mapped in a short time. In order to avoid biasing the system by the starting coordinates, 12 independent trajectories were run starting from different initial random configurations of the system. Figure 2 displays the average structure of the protein together with the occupancy grid,

showing that the area of the known binding site of the 2,7-anhydro-Neu5Ac ligand is the major site for the interaction with the paramagnetic probe.
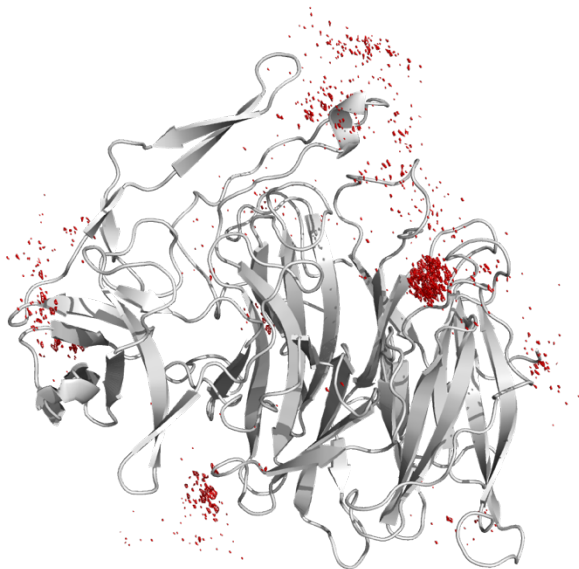
Figure 2. Distribution of TEMPOL (red mesh) around RgNanH-GH33 (grey cartoon) as determined by MixMD. The distribution reveals a single large hot-spot at the known 2,7-anhydro-Neu5Ac binding site, revealing that TEMPOL can act as a probe to selectively target residues of the binding site.

This result clearly excludes the presence of false positives and, more importantly, confirms that TEMPOL is selective for the binding site. Together with the TEMPOL/*Rg*NanH-GH33 interaction TOCSY experiment, these MD data build a solid argument for the use of TEMPOL-based TOCSY experiments to identify specific chemical shifts from residues in the binding pocket to perform the DEEP-STD NMR experiments.

We then carried out the DEEP-STD NMR study using the frequencies identified by the TEMPOL approach. *Rg*NanH-GH33 (50 uM) in the presence of 2,7-anhydro-Neu5Ac (1 mM) in 10 mM Tris-d_{11} D_2O buffer pH 7.8 and 100 mM NaCl at 298 K was saturated with a train of Gaussian pulses of 50 ms each for 0.75 seconds, centered on the chemical shifts of the binding hot-spots 0.6, 0.74, 1.06, 1.15, 1.26, 6.6, 6.74, 7.04, 7.57, 8.56 ppm.

Since the absence of chemical shift assignment of the protein prevents the irradiation frequencies to specific protons in the binding pocket from being known, we propose here a novel approach: instead of using a single pair of frequencies for determining the differential epitope map of the ligand (DEEP-STD map)[1], an *averaging* approach should be followed. First, the DEEP-STD factors are calculated for each experiment resulting from all the possible pairs of aliphatic and aromatic frequencies experimentally identified before (Supporting Information Figure S3). In our case, this resulted in a total of 25 differential epitope maps. Finally, all the obtained DEEP-STD factors are averaged to obtain a unique DEEP-STD map. This approach produces a more accurate depiction of the orientation and the nature of the amino acids surrounding the ligand in the

binding pocket, particularly when no chemical shifts from the protein are available.

Figure 3a shows the resulting experimental average DEEP-STD map of 2,7-anhydro-Neu5Ac binding to *Rg*NanH-GH33.
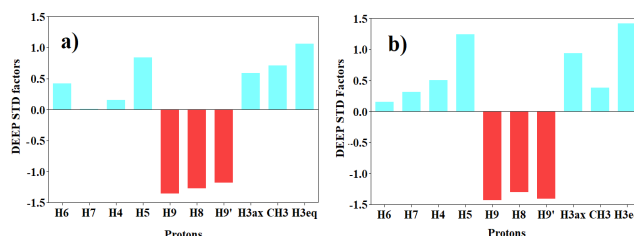
Figure 3. a) Experimental average DEEP-STD factors of the binding of 2,7-anhydro-Neu5Ac to *Rg*NanH-GH33, obtained from 25 differential epitope maps through selective saturation at pairs of frequencies from the two sets experimentally determined by the TOCSY+TEMPOL experimental approach; set 1 (0.6, 0.74, 1.06, 1.15, 1.26 ppm) and set 2 (6.6, 6.74, 7.04, 7.57, 8.56 ppm) (see Supporting Information Figure S3). b) Theoretical average DEEP-STD factors calculated with CORCEMA-ST using ranges of saturation frequencies comprehending the experimental values in the calculation.

The map highlights that CH$_3$, H3a, H3eq, H5 are oriented toward aliphatic residues protons, whereas H6, H7, H4 present little to no preferred orientation, and finally H9, H9' and H8 protons are oriented toward aromatic residues. This result is in excellent agreement with the crystal structure of the complex between 2,7-anhydro-sialic acid and *Rg*NanH-GH33.[1] To confirm that our average DEEP-STD map is a reliable representation of the architecture of the binding pocket, we further carried out a comparison to theoretical predictions of the average DEEP-STD map using the CORCEMA-ST approach (see Supporting information).[4] The calculated average DEEP-STD factors shown in Figure 3b using CORCEMA-ST are in excellent agreement with the experimentally obtained ones. This result further validates our approach, demonstrating that the TEMPOL-based TOCSY approach is a reliable and powerful approach for the identification of the suitable set of saturating frequencies to carry out DEEP-STD NMR studies in the absence of protein chemical shift assignment.

About the general applicability of this approach, although here we have applied it to an enzyme with a polar binding pocket favoring H-bond interactions with TEMPOL, it has been previously shown that the interaction between proteins and TEMPOL can involve weak van der Waals forces, hydrogen bonding and hydrophobic interactions. Several authors have described interactions of TEMPOL with proteins such as Ubiquitin, Lysozyme, Tendamistat, Sso7d, Cyclophyllin, BTPI [17-22] presenting different hydrophobicity/hydrofilicity profiles in their binding sites, which makes us confident on the general applicability of the new protocol for different types of protein targets. Yet, some expected limitations are that TEMPOL should bind the protein with low affinity, to allow an easy interpretation of the spectra in the absence and in the presence of the paramagnetic agent, and it should not induce changes in the conformation of the protein upon binding, which would lead to misinterpretation of resonances to consider for DEEP-STD

NMR or to conformational instability of the protein. It is also worth noting the competition of TEMPOL with water tightly bound to the protein, that in unfavorable cases might prevent the probe to approach the protein surface.[18]

In summary, we have developed a simple experimental procedure to expand the field of application of the DEEP-STD NMR methodology for deriving ligand orientation to protein-ligand cases where the spectral assignment of the protein is not available i.e. when (i) a full NMR assignment is not possible, (ii) the predicted chemical shifts from the structure are not in line with the experimental data (e.g. due to the dynamics of the protein, not accounted for in calculations using a static X-ray structure), or (iii) chemical shift assignments are lacking. Combining 2D TOCSY experiments in the absence/presence of a paramagnetic probe, with the determination of an average DEEP-STD map via saturation at all the experimentally determined frequencies, has been demonstrated to be a powerful approach to allow the type of protein residues most likely to interact with the ligand to be determined. The obtained information on the orientation of the ligand in the binding pocket of the protein opens several interesting applications of the DEEP-STD NMR methodology, for example in the hit-to-lead stage of drug discovery as in 3D-QSAR studies. Further, if combined with the K_D of the complex, the experimentally obtained averaged DEEP STD factors can be used as descriptors to evaluate success or failure of hit modifications during the hit-to-lead stage.

Acknowledgments

Keywords:

[1]   S. Monaco, L. E. Tailford, N. Juge, J. Angulo, Angew. Chem. Int. Ed. **2017**,129 (48), 15491–15495.
[2]   a) M. Mayer, B. Meyer, *Angew. Chem. Int. Ed.* **1999**, *38*, 1784-1788; b) M. Mayer, B. Meyer, *J. Am. Chem. Soc.* **2001**, *123*, 6108-6117; c) A. Viegas, J. Manso, Nobrega F.L., Cabrita E.J. *J.Chem.Ed.* **2011**, 88 (7), 990-994. d) J. Angulo, P. M. Nieto, *Eur. Biophys. J* **2011**, *40*, 1357-1369. e) A. Bhunia, S. Bhattacharjya, S. Chatterjee, *Drug Discovery Today* **2012**, 17, 505-513.
[3]   B. Han, Y. Liu, S. Ginzinger, D. Wishart. SHIFTX2: significantly improved protein chemical shift prediction. J. Biomol. NMR. **2011**, 50 (1), 43-57
[4]   V. Jayalakshmi, N. R. J. Krishna, J. Magn. Reson. **2004**, 168, 36−45.
[5]   B. S. Zerbe, D.R. Hall, S. Vajda, A. Whitty,   D. Kozakov, J. Chem. Inf. Model. **2012**, 27;52(8), 2236-44.
[6]   M. A. Keniry, J.A. Carver. Annual Reports On NMR Spectroscopy, *Vol. 48* (Ed. G. A. Webb), Academic Press, London, **2002**, pp. 32-63.
[7]   N. Niccolai, E. Morandi, S. Gardini, V. Costabile, R. Spadaccini, O. Crescenzi, D. Picone, O. Spiga, A. Bernini. Biochim. Biophys. Acta. **2017**, 1865(2), 201-207.
[8]   I. Solomon Phys. Rev. **1955,** 99, 559-566

[9]   G. M. Clore, J. Iwahara. Chem. Rev. **2009**, 109(9), 4108-4139.
[10]   S. W. Fesik, G. Gemmecker, E. T. Olejniczak, A. M. Petros. J. Am. Chem. Soc. **1991**, 113, 7080-7082
[11]   A. M. Petros, L. Mueller, K. D. Kopple. Biochemistry **1990**, 29, 10041-10048
[12]   P. M. Ung, P. Ghanakota, S. E. Graham, K. W. Lexa, H. A. Carlson. Biopolymers. **2016**, 105(1):21-34.
[13]   K. W. Lexa, H. A. Carlson. J. Chem. Inf Model. **2013**, 25;53(2), 391-402
[14]   L. E. Tailford, C. D. Owen, J. Walshaw, E. H. Crost, J. Hardy-Goddard, G. Le Gall, W. M. de Vos, G. L. Taylor, N. Juge, Nat. Commun. **2015**, 6:7624, 1-12.
[15]   J. Seco, F.J. Luque, X. Barril. J Med Chem. **2009**,52(8):2363-71.
[16]   The PyMOL Molecular Graphics System, Version 1.7 Schrödinger, LLC.
[17]   N. Niccolai, O. Spiga, A. Bernini, M. Scarselli, A. Ciutti, I. Fiaschi, S. Chiellini, H. Molinari, P. A. Temussi. J. Mol. Biol. **2003**, 332(2), 437-47.
[18]   M. Scarselli, A. Bernini, C. Segoni, H. Molinari, G. Esposito, A. M. Lesk, F. Laschi, Temussi P, Niccolai N. J. Biomol. NMR. **1999**, 15 (2), 125-33.
[19]   N. Niccolai, R. Spadaccini, M. Scarselli, A. Bernini, O. Crescenzi, O. Spiga, A. Ciutti, D. Di Maro, L. Bracci, C. Dalvit, P. A. Temussi. Protein Sci. **2001**, 10(8), 1498-507.
[20]   A. Bernini, V. Venditti, O. Spiga, A. Ciutti, F. Prischi, R. Consonni, L. Zetta, I. Arosio, P. Fusi, A. Guagliardi, N. Niccolai. Biophys. Chem. **2008**, 137(2-3), 71-5.
[21]   G. Pintacuda, G. Otting. J. Am. Chem. Soc. **2002**, 23;124(3), 372-3.
[22]   S. W. Fesik, G. Gemmecker, E. T. Olejniczak, and A. M. Petros. J. Am. Chem. Soc., **1991**, 13, 7080-708.

For internal use, please do not delete. Submitted_Manuscript

Entry for the Table of Contents (Please choose one layout)

Layout 2:

COMMUNICATION
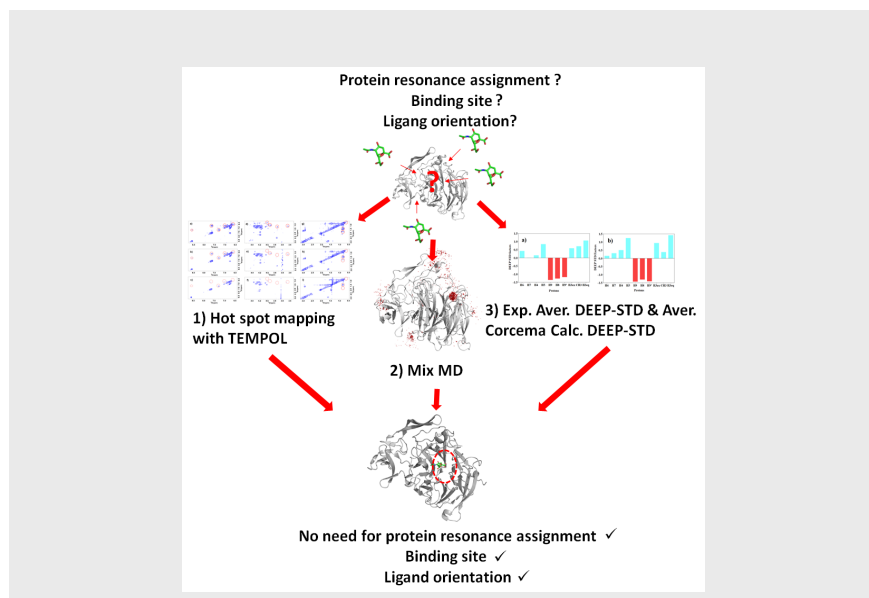
Ridvan Nepravishta, Samuel Walpole, Louise Tailford, Nathalie Juge and Jesus Angulo*

Page No. – Page No.

Deriving ligand orientation on weak protein-ligand complexes by DEEP-STD NMR in the absence of protein chemical shift assignment

Paramagnetic probes assist DEEP-STD NMR experiments to be able to provide information on ligand orientation in a protein-ligand complex even if the protein chemical shifts are unknown.