Response to the Reviewers

We would like to thank the reviewers and the associate editor for their constructive comments and criticisms. Reviewers 1 and 3 were generally satisfied with our revised manuscript and had only minor comments. Reviewer 2 raised a number of additional issues. Most importantly, Reviewer 2 pointed out a highly relevant prior study (Singh and Ahmad 2009). We have revised the manuscript accordingly. Below follows a point-by-point response to the comments.

Comments by the Associate Editor
The authors need to mention in Introduction two other popular methods for RSA prediction - SABLE (Adamczak et al., 2004) and NetSurfP (Petersen et al., 2005). But they can ignore the specific case of RSA application for beta-turns prediction suggested by Reviewer 3, as it is too specific for broad readership. Also, Introduction has to relate the presented work to the previously published study pointed by Reviewer 2 in Singh and Ahmad, 2009.

We now refer to all these works in the introduction and/or discussion where appropriate.

Per Reviewer 3 comment, there should be a remark made that some alternative approaches of computing RSA consider the Ala-X-Ala tripeptide for maximal solvent exposure, supported by references Ahmad et al., 2003; Nguyen and Rajapakse, 2005.

We agree and we have added this topic to the discussion.

Reviewer 2 also raises the question about the removal of redundancy in the dataset, as many protein structures may have structural flexibility and, despite the identical or conserved sequence, adopt different conformations changing solvent exposure of the same residues. This concern goes in line with the comments by Reviewer 1 asking the authors to expand discussion on the structural variability (hence RSA) in protein families, and how their refined normalization can better estimate RSA of residues on the surface. Per Reviewer 2, the authors have to provide a specific example to demonstrate benefits of adjusted RSA.

We believe that removing redundancy is crucial to perform a statistically valid analysis; see our response to Comment 1 of Reviewer 2. Further, we have no reason to believe that using a larger, more redundant data set would yield substantively different results; see our response to Comment 3 of Reviewer 2. We believe that the main novelty of our study is the theoretical modeling of tripeptides in all allowed conformations. This allows us to state confidently what the maximum allowed solvent accessibilities (ASAs) are for each amino acid. Thus, our work can resolve the concern that Singh and Ahmad (2009) state in their concluding sentence: “There may still be some difference between the highest *possible* ASA and highest *observed* ASA due to an insufficiently covered space of ASA distribution in the PDB […].”

Comments by Reviewer 1

Comment 1: The authors made a real effort to address the technical deficiencies pointed out before (in particular filtering out potential artifacts in PDB structures and improved assessment based on CORE/ALLOWED regions of RP). Therefore, the resulting estimates of maximum ASA values can be now regarded as established using a sound protocols and should be published. However, in order to avoid confusing the readers, I believe that the authors should expand on their comments on the variability of RSA in protein families. The problem is not really related to mutations, but it has a much more fundamental dimension. Namely, even with the same amino acid residue (or even within a conformational ensemble in solution for one particular protein), an exposed position may have dramatically different surface exposed areas and thus RSAs (defined with either of standard or the newly proposed maximum values as denominators) in different conformations. On the other hand, fully buried positions are much more likely to have similar RSAs (i.e. to remain fully buried) in multiple conformations and family members. Renormalizing by slightly larger in general) maximum values is not going to change much, and if anything actually would likely exacerbate the problem (as opposed to what the paper claims). While this is within minor changes, I hope the authors will use this additional opportunity to improve the manuscript.

Response: We would like to thank the reviewer for his/her evaluation of our results as “established using a sound protocol” and worthy of publication. We also agree with the reviewer's suggestion to explain in more detail how RSA may change due to thermodynamic fluctuations. We have added the following text to the relevant paragraph in the Discussion:

> At the same time, we have to realize that RSA can show variability even in the absence of mutation, in particular for exposed residues. A residue in a surface loop will undergo thermodynamic fluctuations, and its solvent exposure state will vary over time as neighboring residues move closer in or further out. By contrast, a residue in the core will likely remain solvent-occluded at all times. To obtain a reliable RSA value for a surface residue, one would thus ideally calculate an average over a thermodynamic ensemble of structures.

Comments by Reviewer 2

Comment 1: The choice of PDB data sets is questionable as I do not see why redundancy should be removed at 20% sequence identity.

Response: We respectfully disagree. It is crucial to remove redundancy, because the PDB does not represent an unbiased sampling of protein structures. Some structures are sampled over and over again while others are sampled rarely. For example, there are over 100 structures of the influenza protein haemagglutinin. At the sequence level, the vast majority of these structures are exactly identical, or differ at most by one or two substitutions. At the structural level, these structures are also nearly identical, showing only very minor distortions in the backbone (Meyer et al., Phil. Trans. R. Soc. B. 368:20120334, 2013). It would be unreasonable to include all of these structures in an analysis of the kind we carry out and treat them as separate, independent data points. They are not. The statistical terminology for this kind of situation is “pseudoreplication,” and it is important to minimize pseudoreplication for statistical validity. Indeed, it is conventional practice in studies of PDB structural variation to work with curated data sets where redundancy has been removed. One could argue the exact cutoff in sequence identity that should be used, but there is no getting around the point that some cutoff is needed.

We also would like to emphasize that we have no reason to believe that using a larger data set would yield substantially different results. Already, on the empirical side, regardless of whether we restrict our results to the central 80% of data, the central 97% of data, or all data, we get virtually the same normalization scales. Further, our empirical results are largely in agreement with the results of Singh and Ahmad (2009), who used a much larger data set.

As far as the theoretical modeling is concerned, we now provide (in a public online repository, see Methods) backbone-dependent maximum SA values for **all** backbone configurations, allowed or disallowed. Thus, if at a time in the future somebody wanted to find maximum SA values for backbone configurations we excluded in our analysis, they could do so trivially from our raw data.

Comment 2: In this manuscript authors aim to define upper bound of ASA values, so that their relative measures fall within a bound range. Authors do not provide enough justification for setting up upper bounds, except that they yield more than 100% RSA in some cases. I do not see a problem if a real structure has more than 100% RSA, as long as I understand what is the meaning of this value. Nonetheless, I understand the value of having an upper bound of ASA value, for example in computing differences in relative values of non-identical residues.

Response: The reviewer touches exactly upon the main point that we argue, in his/her last sentence, “I understand the value of having an upper bound of ASA value, for example in computing differences in relative values of non-identical residues." This is exactly our position. The issue is not so much that the normalizations are somewhat off, but that they are off by different amounts for different amino acids.

Further, on a more philosophical note, we find it unsatisfactory to work with normalization values that don't normalize RSAs to between 0 and 1. The main reason why we embarked on this project in the first place was that in numerous studies in our lab, we were encountering RSA values >1 and we didn't know how to interpret this finding. The present paper explains this finding and provides a new set of normalization values that don't generate this issue.

Comment 3: I am also surprised that the authors have completely neglected a very elaborate study on computing highest observed ASA values from empirical considerations in Singh and Ahmad (BMC Str Biol, 9(1) 25 (2009)). In that work, authors have analyzed observed ASA values in much more details than in this manuscript. For example, sequence context has been considered and the value to predictability of RSA has been evaluated. Furthermore, I do not agree why authors in the current work remove redundancy in the PDB data? There must be many local structures which would be missed if protein structures having more than 20% sequence ID are discarded. This will therefore fail in reproducing the real empirical range of ASA observed values.

Response: We would like to thank the reviewer for pointing out this paper to us. We were not aware of it. It is indeed very relevant to our study. Ours and their study are broadly in agreement. At the same time, the two studies complement each other, since the Singh and Ahmad (2009) study was entirely empirical while ours has a large theoretical modeling component. In fact, Singh and Ahmad identified only Highest Observed Accessibility (HOA), while we identify also the Highest Possible Accessibility (HPA). We further demonstrate that HPA is relatively close but distinct from HOA.

Singh and Ahmad analyzed solvent accessibility on a considerably larger empirical data set, and they showed that the largest solvent accessibilities appear in turns and loops, not in extended conformations. They suggested that normalization by extended conformations (as is current practice) is sub-optimal, and that normalization by HOA is preferred. Our empirical analysis, while carried out on a smaller data set, generally agrees with their findings. Further, our theoretical analysis provides HPA in addition to HOA, and it demonstrates that HOA is generally close but not identical to HPA.

Singh and Ahmad chose to look at a larger set of data points in order to address how amino-acid neighbors and backbone conformations influence ASA. We carried out a similar analysis on our data; however, by our assessment, these finer-scale normalizations, in particular normalization by amino-acid neighbor, add substantial complexity and overhead for little gain. This assessment is broadly consistent with the results of Singh and Ahmad. Note, however, that we now make all our raw data available, including HPA values as a function of backbone dihedral angles. Further, the conformations that Singh and Ahmad identified to yield the highest observed solvent accessibility are consistent with our results. Arguably, though, we provide a more precise definition of these conformations, since we list specific backbone angles corresponding to maximum SA values. Singh and Ahmad did not provide this information.

Finally, we would like to emphasize a fundamental difference in philosophy between Singh and Ahmad's approach and our empirical approach. They attempted to identify the absolute highest observed accessibility value. As such, they had to use as large a data set as possible. By contrast, we chose to approach the empirical analysis from a sampling perspective. We do not claim that the empirical maximum RSA values we report are the highest ever seen anywhere in the PDB. We claim that taking a reasonable sample based on a variety of PDB files can give rise to ASA values that far exceed the previous standard of the extended conformation. In addition, our sampled data set yields very similar results to the larger data set of Singh and Ahmad.

In response to the reviewer comment, we have made the following changes to the manuscript:

1. We now make all our data available. This includes maximum SA values as a function of backbone dihedral angles.

2. We mention the Singh and Ahmad paper in the introduction. We discuss the similarities and differences to our work in the discussion:

> Our results are broadly consistent with a recent paper by Singh and Ahmad [20]. These authors did an extensive empirical survey of ASA values in tripeptides from PDB structures. They found that the highest observed ASA values were found in loops and turns, not in the extended conformation used by Miller et al.. Their highest ASA values are generally consistent with ours. Further, Singh and Ahmad found that the highest observed ASA values were dependent on the neighboring residues around the focal residue. Finally, Singh and Ahmad showed that for RSA prediction from primary sequence, prediction accuracy could be improved by approximately 10% if ASA values were normalized by (neighbor-dependent) highest observed ASA values rather than by ASA values observed in the extended conformation [20]. Our work serves as a useful complement to their work, by (i) providing, through molecular modeling, highest possible ASA values rather than just highest observed ASA values, by (ii) providing highest observed and highest possible ASA values as a function of backbone dihedral angle, and by (iii) demonstrating that improved RSA normalization yields empirical hydrophobicity scales that are more similar to experimentally measured ones.

Comment 4: In addition to just the statistics of upper bound obtained empirically and otherwise, some practical value of these computations must be demonstrated.

Response: There are a number of immediate practical benefits. First, resolving the issue of having RSA values > 1 is a practical benefit to our lab. We use RSA frequently to predict evolutionary rates (see references cited in the introduction to this paper), and we find it unsatisfactory to have a “normalized” value that isn't properly normalized and that can take on values > 1. Through our modeling work, the Highest Possible ASA values are now known. Second, Singh and Ahmad (2009) demonstrated that improved RSA normalization results in improved RSA prediction (on the order of 10%) and may be useful in identifying DNA binding residues. Third, we demonstrate here that improved RSA normalization yields empirical hydrophobicity scales that are more similar to experimentally derived scales. This result may be of benefit to applications that make use of empirical hydrophobicity scales, such as protein-fold prediction.

More generally, however, we would like to respectfully disagree with the demand that “some practical value of these computations must be demonstrated.” Not every scientific study needs to provide an immediate, practical benefit of some sorts. The purpose of science is to generate knowledge. Further, the purpose of science is to increase the confidence in the knowledge we have, as well as its accuracy. We believe that our study falls primarily into the latter category. In particular, by systematically enumerating tripeptides in all allowed conformations, we overcome the potential issue of poorly-sampled side-chain conformations in the PDB, as identified by Singh and Ahmad as a potential caveat of the empirical approach.

We would also like to point out that according to the PLOS ONE Publication Criteria, PLOS ONE will publish works that present results of primary scientific research, carried out to a high technical standard and described in sufficient detail, and with conclusions that are supported by the data (http://www.plosone.org/static/publication). Nowhere in the document is “practical value” stated as a required condition. We believe that our paper meets the criteria required for publication by PLOS ONE.

Comments by Reviewer 3

Comment 1: In the introduction the authors mention computational methods to predict RSA from protein

primary and/or secondary structure. I do not see a mention of NetSurfP, which is a highly used and well cited web-server.
www.cbs.dtu.dk/services/NetSurfP/
A generic method for assignment of reliability scores applied to solvent accessibility predictions.
Bent Petersen, Thomas Nordahl Petersen, Pernille Andersen, Morten Nielsen and Claus Lundegaard1.
BMC Structural Biology 2009, 9:51 doi:10.1186/1472-6807-9-51.

Response: We thank the reviewer for pointing us to this article. It is now cited in the introduction.

Comment 2: It is mentioned that Gly-X-Gly tripeptides are conventionally used to evaluate the surface area. This is true but methods using Ala-X-Ala has also been used as described in Ahmad S, Gromiha MM, Sarai A: Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003, 50(4):629-635.

Response: We agree with the reviewer, and we now mention this point in the discussion:

> In our modeling approach, we calculated ASA values for Gly-X-Gly tripeptides. Other authors have considered normalizations based on Ala-X-Ala tripeptides [18, 33] or even neighbor-specific normalizations (i.e., a different normalization for each specific tripeptide [20]). We chose Gly-X-Gly tripeptides because we wanted to calculate the highest possible ASA values of tripeptides, and glycines will generally occlude less solvent than alanines. From a practical perspective, we prefer a simple normalization scheme, and hence highest possible ASA values are attractive to us. However, for certain applications, it may be the case that neighbor-specific or backbone-specific normalizations are preferable. Singh and Ahmad [20] provided neighbor-specific normalization values, but didn’t control for backbone angles. We have shown here that maximum ASA values depend substantially on backbone angles (e.g. Fig. 3), and we provide both highest observed and highest possible ASA values as a function of backbone angles (see “Data and code availability” in Methods). It is not known at this time whether neighbor dependent or backbone-dependent normalization is preferable, and the answer may depend on the specific application. In principle, one could also normalize by both neighboring amino acids and backbone dihedral angles. A modeling approach such as ours could be employed to calculate the highest possible ASA values for any tripeptide in any conformation. The computational resources required would be substantial, however, since we would have to model 400 times more tripeptides than we did for the present work.

Comment 3: The authors would maybe like to correct the spelling mistake in the last section "Hydrophobicity scales". At the 4th line hydrophobicity is misspelled as hydrophbicity.

Response: Thank you. It is fixed in the revision.