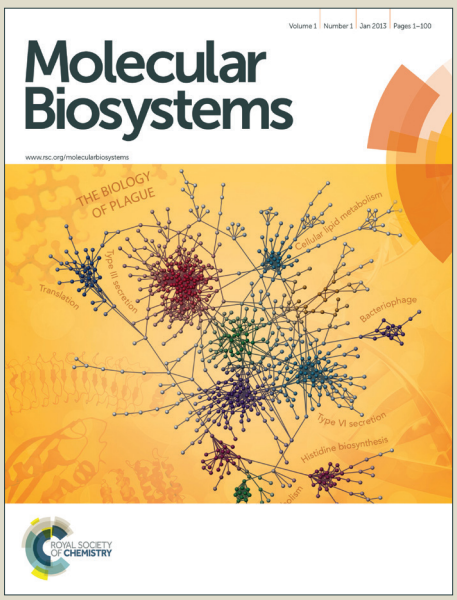
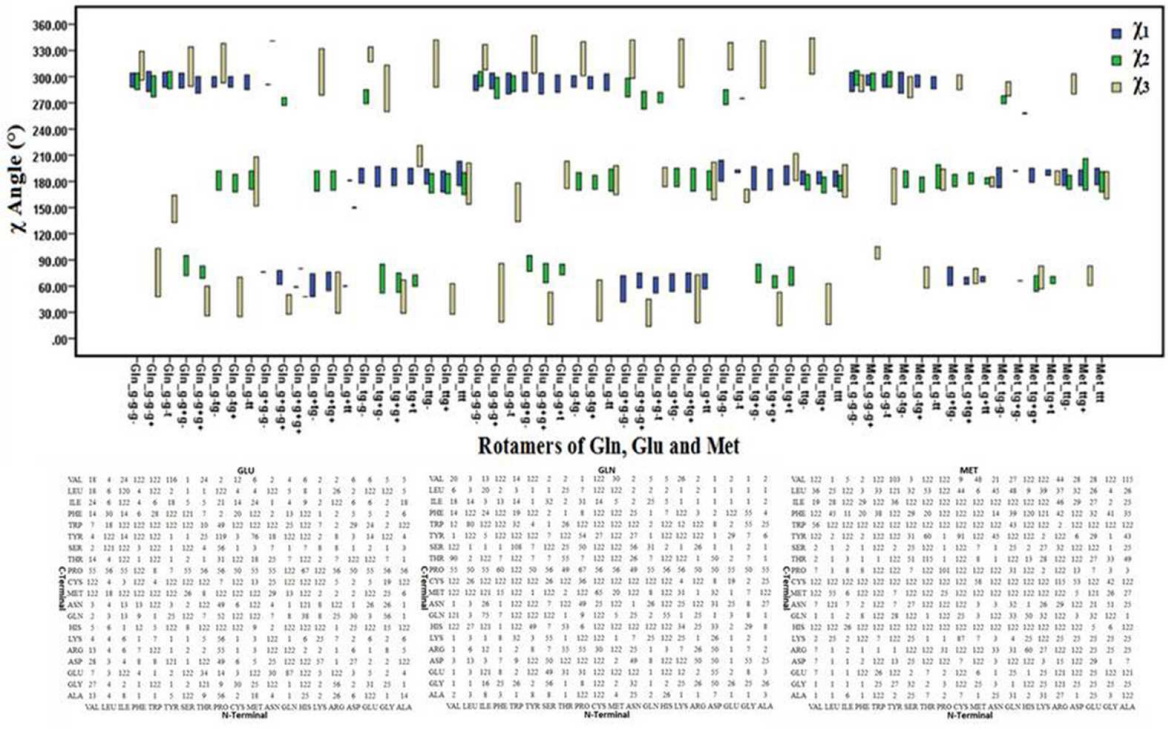
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems



We established a Sequence Dependent Rotamer Library(SDRL) to help side-chain modeling, better understanding of amino acid side-chain conformational selection and seeking neighbor dependency of this conformational selection.

Molecular BioSystems

RSC Publishing

PAPER

SDRL: A Sequence Dependent Protein Side-chain Rotamer Library†

Cite this: DOI: 10.1039/x0xx00000x

Mohammad Taghizadeh,[a] Bahram Goliaei*[a] and Armin Madadkar-Sobhani[a]

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Since the introduction of the first protein side-chain rotamer library (RL) almost half a century ago, RLs have been component of many software and algorithms in structural bioinformatics. Based on the dependence of the side-chain dihedral angles on the local backbone, three types of RLs has been identified: backbone independent, secondary structure dependent and backbone dependent. In all the previous efforts, the effect of sequence specificity on the side-chain conformational preferences was neglected. In an effort for developing a new class of RLs, we considered that the central residue's side-chain conformations of each triplet in the protein backbone, depend on the sequence of the triplet therefore, we developed a sequence dependent rotamer library (SDRL). To accomplish this, 400 possible triplet sequences of 18 natural amino acids as the central residue, which corresponds to 7,200 triplet sequences in total, were considered. Seeking the set of 11,546 selected PDB entries for the 7,200 triplet sequences resulted in 2,364,541 of occurred instances for 18 amino acids. Our results show that Leu and Val receive minimum impact from adjacent residues in choosing side-chain conformations and Cys, Ile, Trp, His, Asp, Met, Glu, Gln, Arg and Lys on the other hand, select their side-chain conformations mostly based on the adjacent residues on the backbone. The rest of residue types were moderately dependent on their adjacent residues. By using the new library, side-chain repacking algorithms can find preferred conformations of each residue more easily than other backbone independent RLs.

Introduction

In a nutshell, a side-chain rotamer (rotamer for short) is a single side-chain conformation represented as a set of values for each dihedral angle degree of freedom, known as χ angles in proteins, and collection of rotamers for each residue type plus their relative frequencies is called a (side-chain) rotamer library (RL)[1]. Since the introduction of the first RL almost half a century ago by Chandrasekaran and Ramachandran[1, 2], they have been component of many software and algorithms in structural bioinformatics. However, modelling side-chain conformations[3-11], protein-protein docking[12-17], crystal structure refinement[10, 18, 19], modelling site-directed mutations[20], small ligand docking with flexible receptors [21, 22] and side-chain conformational analysis[5] are among the RLs applications.

There are two statistical analysis methods in general for deriving RLs: conformational clustering and bins selection[1]. In the conformational clustering method, side-chain conformations are clustered based on the χ_s or χ_s plus φ/ψ angles in backbone independent or dependent RLs, respectively. On the other hand, the idea behind bins selection method is dividing dihedral angle space based on physical-chemical propensities (e.g. rotation about sp^3–sp^3 or sp^3–sp^2 bonds) into several bins and determining an average conformation in each bin.

Conventionally, based on the dependence of the dihedral angles on the local backbone, three types of RLs has been identified: backbone independent, secondary structure dependent and backbone dependent[1]. Lovell–Richardsons[23], Dunbrack et al[24] and McGregor et al.[25] are typical examples of backbone independent, backbone dependent and secondary structure dependent RLs, respectively. In addition, there is a variant of the backbone-dependent rotamer libraries, recognized as position–specific, which uses a fragment of odd numbered (e.g. five or seven) amino acids with similar backbone φ and ψ angles, whose central residue's side-chain conformation is examined[26]. Dunbrack and Cohen have used an backbone dependent RL for ecomplish side-chain repacking related to homology[27].

Besides the above-mentioned general trends, in recent years there have been some contributions in the field to incorporate insight gained from the first–principles approaches into the existing knowledge of the experimentally determined structures in order to enhance the side-chain prediction accuracy. The used of huge number of rotameric states obtained from experimental data in conjunction with an *ab initio* potential energy function or statistical methods[7, 28] and inferring sidechain rotamer preferences from

molecular dynamics simulations in water[29] are some of the efforts in this regard. Based on the Anfinsen's dogma, at least for small globular proteins, the native structure is determined by the protein's amino acid sequence alone[30]. This sequence-structure relationship was investigated for fragments of two consecutive amino acids or doublets[31-33], triplets[34], and even longer sequences[11, 17], which the later case commonly considered as the structural alphabets for proteins. Nevertheless, all the previous studies focused on the effects of sequence specificity on the backbone rather than side-chains. In an effort to better understand the sequence specificity of side-chain conformational preferences, here we present a new type of backbone independent RL based on the preceding and succeeding adjacent residues which we call it sequence dependent rotamer library (SDRL). The development of RLs based on the sequence information is a practical approach because in almost all the applications of RLs, it is readily available. In addition, because the bulk of the rotamers for each amino acid already has been divided into 400 triplets, side-chain modelling algorithms will be faced with a smaller assignable subset of rotamers which result in reduced search space. This is very essential for some of the problems related to RLs in protein science, particularly those for which even finding a approximately reasonable solution is NP-complete[35].

To accomplish this, 400 possible triplet sequences of 18 natural amino acids as the central residue were considered, which corresponds to 7,200 triplet sequences in total. Seeking these triplet sequences in the set of more than 11000 selected PDB entries resulted in more than two million occurred instances. Our results show that by using SDRL, side-chain modelling algorithms could find preferred conformations of each residue more easily than other backbone independent RLs.

In addition, some analyses which illustrate patterns of side-chain dihedral angle preferences of each residue type based on the immediate adjacent residues will be presented. Using this information, all residue types can be divided into three classes of high, moderate and low dependency on their immediate adjacent residues.

Methods

Experimental data

The PISCES server[36] was used to cull PDB structures with ≤ 50% sequence similarity, resolution ≤ 2.0 Å, R–factor ≤ 0.25, sequence length between 40 and 10,000 and excluding non X-ray structures and those with only alpha carbons. The set of 11,546 PDB structures (SMs-File 2) was chosen from the RCSB database[37] using this methodology. Two PDB structures: 1WTE, 1M0K which have redundant chains with the exact same coordinates were eliminated. Residues with poor electron densities (i.e. damaged residues in PDB structures) and also the first and last residues of each chain were also eliminated.

χ Angles extraction and bin selection

For extracting the χ angles from the PDB Files, Dangle software from the Richardson laboratory was used[38]. Bin selection and method of calculating dihedral angles and choosing zero reference for χ angles was adopted from Lovell–Richardsons and colleagues[23], i.e. three 120° bins centered on each staggered conformation (g^+ = 60°, t = 180°, and g^- = –60° for gauche positive, trans and gauche negative, respectively) considered for χ angles. However, rather than different bins for the terminal *sp3–sp2* χ angles present in some of the amino acids (i.e. Phe, Tyr, Trp, His, Gln, Glu and Arg), we used the same three 120° bins for all the χ angles (see Results and Discussion sections χ_{1+2}, χ_{1+2+3} and $\chi_{1+2+3+4}$ amino acids for the rationale).

For each amino acid with only χ_1 angle (χ_1 amino acids; Ser, Cys, Thr, Val and Pro) as a central residue of a triplet fragment, we considered a set of 1,200 rotamers (20×20 = 400 possible triplet sequences, multiplied by three 120° bins, hence 1,200). For amino acids having χ_1 and χ_2 angles (χ_{1+2} amino acids; i.e. Ile, Leu, Asp, Asn, His, Phe, Tyr and Trp), a set of 400×3×3 = 3600 rotamers were considered. Using the same methodology, 10,800 and 32,400 rotamers were considered for the χ_{1+2+3} (Met, Glu and Gln) and $\chi_{1+2+3+4}$ (Lys and Arg) amino acids, respectively.

Relative frequencies and bin preference orders

Relative frequencies (RFs) defined for each rotamer according to the following formula:

$$\mathbf{RF_{itc}} = \frac{\mathbf{P_{itc}}}{\mathbf{P_{it}}}$$

Which P_{itc} is all the observed cases for the c^{th} conformation of the t^{th} triplet of the i^{th} amino acid and P_{it} is all the observed conformations of the t^{th} triplet of the i^{th} amino acid.

For each residue type, RF values in g^+, t and g^- bins may be different among 400 triplets and by ordering them from highest to lowest value, a pattern of bin preference order (BPO) can be defined for the sake of classification. For example in case of χ_1 amino acids, there are seven BPOs as follows:

1. g+ RF >g– RF >t RF
2. g– RF >g+ RF >t RF
3. t RF >g+ RF >g– RF
4. g+ RF >t RF >g– RF
5. g– RF >t RF >g+ RF
6. t RF >g– RF >g+ RF
7. g+ RF = g– RF = t RF or g+ RF >g– RF = t RF or g+ RF = g– RF >t RF

However, only for the χ_1 amino acids all the combinations of RFs were considered. For amino acids with two or more χ angles, all possible combinations of RFs will make a big number (e.g. 9! for the χ_{1+2} and 27! for the χ_{1+2+3} amino acids, respectively). Therefore, we considered top four RFs (in average of all the triplets) for the χ_{1+2} and top five in case of the χ_{1+2+3} and $\chi_{1+2+3+4}$ amino acids, respectively. By considering one additional BPO in each category representing all sort of conformational frequencies with at least one equality, we included 25, 122 and 122 BPOs for the χ_{1+2}, χ_{1+2+3} and $\chi_{1+2+3+4}$ amino acids, respectively (SMs-File 3).

Comparison with other RLs

The Lovell–Richardsons[10] and Dynameomics[29] RLs were chosen for the comparison. Lovell–Richardsons rotamer library introduced in 2000 has a much stricter criteria for including side-chains in the dataset and as a result it was the first backbone independent RL with a higher accuracy. The Dynameomics rotamer library is the first RL based on the side-chain rotamer preferences and dynamics in solution, consequently charged surface residues are better represented.

In order to compare the SDRL with the two above mentioned RLs, we calculated number of rotamers plus the percent of their covered side-chain conformations. For the χ_1 amino acids which nearly all the observed conformations are involved in sufficiently populated rotamers, the difference between the three RLs was very small (except for Cys), therefore, the number of rotamers and percent of included real side-chains were not calculated. On the other hand, for the χ_{1+2} amino acids we considered triplet rotamers with RF more than 3%. However, for Ile, Phe and Asn we used 1%, 4% and 3.5% as threshold, respectively. We used two thresholds (2.5% and 3.5%) for Asn to reach to a compatible number with each of the two selected RLs. For Met, Glu and Gln (i.e. χ_{1+2+3} amino acids), we used 3%, 2.8% and 2.7% as thresholds respectively, and finally for the $\chi_{1+2+3+4}$ amino acids, rotamers with RFs of at least 1% were included in the calculation. To calculate this ratio for the SDRL, the average of the parameter for all the triplets of each amino acid with above thresholds was used. Nevertheless, since for choosing a side-chain conformation of a residue on the protein backbone, the preceding and succeeding adjacent residues are known, the average parameter for SDRL is totally compatible with the form of parameter for the other two RLs.

Results and Discussion

The complete backbone independent sequence-dependent side-chain rotamer library is provided in the Supplementary Materials (SMs)-File 1. In developing the resulting library, contrary to the previous efforts[1, 23, 28, 39-44], we considered that the central residue's side-chain conformations of each triplet and/or their frequencies, depend on the sequence of the triplets. To accomplish this, 400 possible triplet sequences of 18 natural amino acids as the central residue were considered, which corresponds to 7,200 triplet sequences in total. Gly and Ala were excluded from the list of 20 natural amino acids, because Gly does not have any χ angle and low electron density of hydrogen atoms prevents χ_1 angle of Ala to be traceable in crystallography. Seeking the set of 11,546 selected PDB entries (SMs- File 2) for the 7,200 triplet sequences resulted in 2,364,541 of occurred instances for 18 amino acids. In the following sections, we present the results based on the number of χ angles in the amino acids.

χ_1 amino acids

The χ_1 amino acids are Ser, Cys, Thr, Val, Pro and Ala, of which Ala does not have a traceable χ angle because of missing hydrogen atoms and χ angle of Pro happens only in two 120° bins rather than three because of a cyclic side-chain.

Table 1 General properties of rotamer seeking for the χ_1 amino acids

Amino acid	Triplet case No.		Low frequency triplet rotamer	No. of rotamers With at least 10 cases[a]
	Before filter	After Filter		
SER	166985	166553	8/1200 (0.66%)	1192
CYS	36557	36536	229/1200 (19.08%)	971
THR	154158	153941	64/1200 (5.33%)	1136
VAL	205354	205135	53/1200 (4.42%)	1147
PRO	132007	131937	5/800 (0.62%)	795

[a]Number of rotamers calculated based on 400 forms of triplets for each of three possible 120° bins minus low frequency (less than 10 cases) ones. It reaches to 1200 rotamers for each of the χ1 amino acid in fully form (with three bins for each one of triplets).

General properties of rotamers for the χ_1 amino acids have been summarized in Table 1. For the χ_1 amino acids, 695,061 triplets were encountered in the dataset, of which 694,102 were considered and 959 were omitted because of problems in their side-chain structures. Rotamers with less than 10 occurrences have also been labelled as low frequency (LF) in the supplementary materials. For Cys, which happens less frequently than the other χ_1 amino acids low frequency rotamers are about 19% of all possible rotamers (i.e. 400 triplets × 3 bins = 1200). For the rest, they are less than 6%. As there is low correlation (r = 0.52) between the number of considered conformations and high frequency rotamers (columns 3 and 5 in Table 1), these low frequency rotamers can be considered as conformations that rarely happen in protein structures.

In this category of amino acids average deviation in RFs of same bins of same amino acids within different triplets is 41% and moreover in average there is 12.68° of variations in χ_1 angle averages and in maximum point it reaches to 21.2° for trans conformation of Thr (SMs-File 4, Table A).

The pattern of RF values for three considered bins in descending order can be different among 400 triplets of each residue type. We refer to this pattern as BPO which reveals shuffling of the order among triplets of the same residue type (see Methods). Each 400 possible triplets of each χ_1 amino acids prefers one of these BPOs. For Ser and Cys all 7 BPOs among 400 triplets were observed, but for Val, Thr and Pro only 3 forms were observed (Fig. A of SMs-File 4).

In order to rule out the effect of adjacent residues on the central residue's preference, we used G–X–G triplet, in which adjacent residues have the smallest side-chains, as the reference. Last column in Table 2 shows the percent of BPO patterns among 400 triplets of χ_1 amino acids that are not the same as G–X–G's BPO. It represents the percent of triplets for the X residue types that their BPO patterns are mostly related to their adjacent residues on the backbone. Low percentages, e.g. 2.75% for Val, is an indication of no or little dependence on the adjacent residues. In other words, 97.25% of Val

triplets' BPO is same as G–V–G's BPO. On the other hand, high percentages like 80.25% in case of Cys, shows that Cys selects the three possible bins mostly under the influence of its adjacent residues on the backbone. Ser, Thr and Pro are between these two extremes. Due to the fact that Ala has the second smallest side-chain, BPOs of A–X–A were also included in Table 2. For Cys, Pro and Thr which G–X–G's BPO is not equal with A–X–A's, the non G–X–G percent is larger than 40%.

To answer the question about why some amino acids have dependency to their adjacent residues in selecting their side-chain conformations and some others not, generally it can be subjected to their physico-chemical properties. Among five χ_1 amino acids, Val which has nearly no dependency on its adjacent residues, has a symmetrical side-chain and two bulky methyl groups which force the central residue of most triplets to choose trans conformation no matter what its adjacent residues are. On the other hand, the rest of χ_1 amino acids have smaller and asymmetrical side-chains. Ser, Thr and Cys also have one polar group in their side-chains that can participate in various interactions with adjacent residues, which contributes to asymmetric properties of their side-chain structure. In case of Cys, in addition to great difference between G−C−G triplet with other ones, there are great variability of BPO of different triplets, which probably is a sign of the effect of triplet sequence on determining S−S bond orientation and choosing its side-chain conformation and vice versa. Despite the fact that Pro has only two of the three possible 120° bins, it shows considerable variation in its BPO within different triplets. Probably because of Pro's special side-chain structure, role of backbone in this variation for Pro is considerable.

As can be seen in Table A of SMs-File 4, wide range of variations are associated with χ_1 angles and these variations are not just limited to few triplets, but it is highly distributed in nearly all of them (Fig. 1 and Fig. B of SMs-File 4).

Fig. B of SMs-File 4 illustrates statistical distributions of χ_1 angle means for all 400 triplets of the five χ_1 amino acids with a frequency greater than 9. The normal curve has been fitted on these distibutions for the sake of comparisons and as can be seen, they have near normal distributions but skewness also exists in few of them. All the distributions show a large spread which indicates that a large number of triplets have means with considerable deviations from the grand mean.

Fig. 2, summarizes the correspondance between grand mean of each χ_1 amino acid in the SDRL and Lovell–Richardsons' RL. The average difference between the two RLs is 1.74° for Ser, Thr, Cys and Val and 2.15° for all 5 amino acids. However, in average there is more than 12° difference between 400 triplets of the same amino acid within the same bin. In case of Thr, the difference can reach to 21.2° for *trans* conformation.

Table 2 An analysis of BPO patterns for the five χ_1 amino acids; 1-7 represent seven types of patterns which has been introduced in Fig. A of SMs-File 4. G–X–G is the sign for the triplet with glycine in its N-terminal and C-terminal immediate adjacent sides and A–X–A has alanine in both sides of its central residue.

Amino acid	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)	7 (%)	G–X–G	A–X–A	Non G–X–G (%)
SER	67.5	4.75	4.25	19.5	0.75	0.25	3	**1**	**1**	32.25
CYS	0.5	20.75	0.25	0.25	67.5	6	4.75	**2**	**5**	80.25
THR	58.75	40.5	0	0	0	0	0.75	**1**	**2**	40.75
VAL	0	0	1.25	0	0	97.75	1	**6**	**6**	2.75
PRO	46.75	51.5	0	0	0	0	1.75	**1**	**2**	48.5

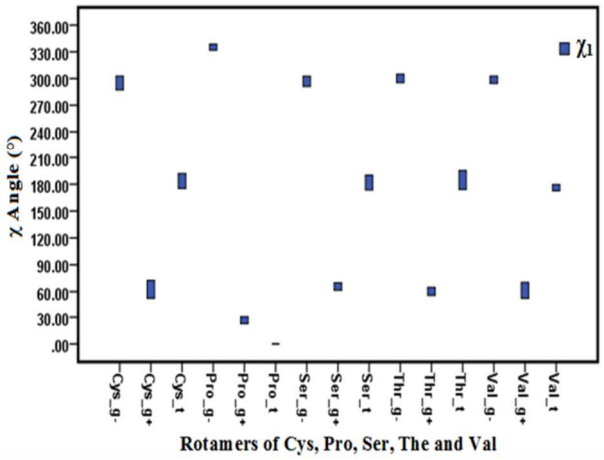


Fig. 1 Range of variation of χ angle averages within 400 triplets of χ_1 residue types.

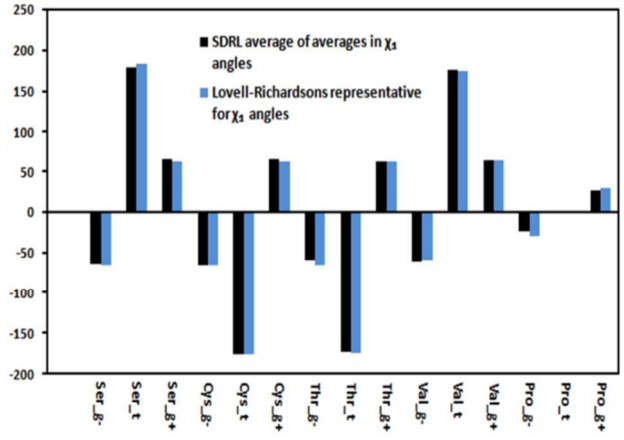


Fig. 2 Comparison average of averages in the χ_1 amino acids within the SDRL with Lovell–Richardsons RL χ_1 angle representatives.

χ_{1+2} amino acids

General properties of rotamers for the χ_{1+2} amino acids including Leu, Ile, Phe, Trp, Tyr, Asn, His and Asp have been summarized in Table 3. In total 1,041,647 triplets were encountered in this category, of which 1,038,567 were considered.

In other libraries unusual bins has been used for terminal χ angles of Phe, Tyr, Trp, His and Asn. However, we have used usual bins for these χ angles since usual bin selection for them actually gave us populated rotamers and their SDs for these χs with usual bins were compatible with SDs in RLs which unusual bins have been used. Nevertheless, by adding more bins we could decrease amount of SDs but in expense of increasing number of rotamers which was already high.

Among χ_{1+2} amino acids, Trp and Asp have the most and least LF rotamers, respectively. Again, as there is a low correlation (r = 0.11) between the number of considered conformations and high frequency rotamers, these LF rotamers can be considered as conformations that rarely happen in protein structures. The range of variations within χ angle averages in this category is even more than the χ_1 amino acids (Fig. 3), for instance in g^-g^+Leu, there are 33.43° and 44.2° of variation for χ_1 and χ_2 mean angles, respectively. Range of variations in all aspects of rotamers in this category has been abstracted in Table B of SMs-File 4.

For the χ_{1+2} amino acids, 25 different BPOs were considered, but same numbers for different residue types do not have the same meaning (*see* Methods and SMs-File 2). There are extensive variations in BPO patterns for most of the amino acids in this category (Fig. C of SMs-File 4). Based on the information presented in Table 4, Ile, Trp, His and Asp are highly dependent on the adjacent residues on the backbone for selecting their side-chain conformations. Asn, Phe and Tyr are moderately dependent and Leu is the least dependent in that regard. Patterns of G–X–G and A–X–A in Table 4 are not similar for the amino acids with non G–X–G percent bigger than 55.0%.

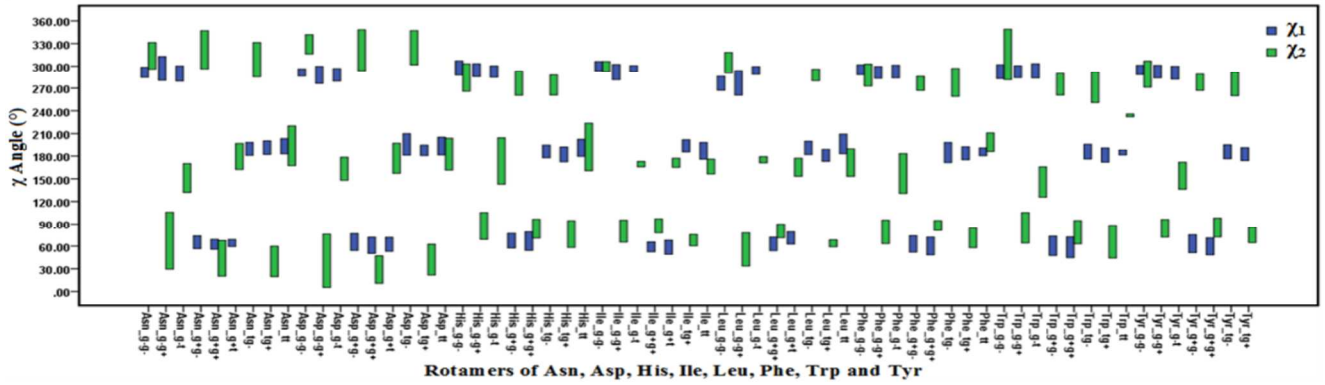


Fig. 3 Range of variation of χ angle averages within 400 triplets of χ_{1+2} residue types.

Side-chain of Leu has similarities with Val, nevertheless, it has two χ angles which means more degree of freedom for taking influence from its neighbors. Ile also has similarities with Val but its adjacent residues dependency in selecting its side-chain conformation is much greater than Val. As an explanation for this phenomenon, considering side-chain structure of Ile can reveal a fork with unequal branches, however, Val and Leu have forks with equal branches. It can be deduced that such an asymmetrical form of structure in Ile side-chain causes greater dependency on adjacent residues in selecting side-chain conformations. In comparison of Phe, Trp, Tyr and His, all of them have an asymmetrical side-chain structure, but Trp have more bulkiness and bigger structure in one side in comparison with Phe, Tyr and His. However, His side-chain has more polarity even a net charge beside its asymmetric side-chain structure. Therefore, these physico-chemical properties of Trp and His side-chains could be the reasons for higher dependency of their side-chain conformational selection on their adjacent residues in comparison with Phe and Tyr. This symmetric/asymmetric effect also is observable in Asp and Asn. Asp has a net charge in terminal part of its side-chain, which is not the case for Asn. To better understand the symmetrical/asymmetrical structures of χ_{1+2} amino acids, the effect of χ_1 dihedral angle of these eight amino acids in bending of side-chain structure must be considered.

χ_{1+2+3} amino acids

As can be seen in Table 5, 342,274 triplets were encountered for the χ_{1+2+3} amino acids (i.e. Met, Glu, Gln), of which 336,681 were considered. The percent of low frequency rotamers shows an increase in this category in comparison with the previous ones, which can be interpreted as the tendency of the χ_{1+2+3} amino acids to adopt fewer conformations amoung the all possible ones. It becomes evident from Fig. 4 that the range of variations of the χ_3 angle for this category is in general more than the χ_2 angle. The range of variations of the χ_3 angle grand mean is more than 40° in many triplet rotamers of Gln and Glu. Among three amino acids in this category, Met has the highest ratio between numbers of high frequency rotamers and numbers of encountered triplets (columns 5 and 3 of Table 5, respectively). It means that Met has capability to adopt more conformations in comparison with Gln and Glu. Table C of SMs-File 4 is an abstraction of variation in all aspects of χ_{1+2+3} rotamers.

We considered top five RFs in the definition of BPO patterns for this category and 122 different BPOs were defined (Fig. D of SMs-File 4). BPO number 122 consists of more than one pattern because

all the cases with equalisations of two or more RFs in a triplet are included in this BPO. As a result, the repeatation of BPO 122 reflects more equal RFs in an amino acid such as Met. In none of the χ_{1+2+3} amino acids, G–X–G's BPO was the dominant one. It can be an indication of that in larger amino acids the effect of adjacent residues in selecting side-chain conformation is more substantial than the effect of physico-chemical tendency of the side-chain itself.

Based on the information presented in Table 6, side-chain conformational preferences for the χ_{1+2+3} amino acids are mostly dependent on the adjacent residues on the backbone and in general this dependency is more profound than previous categories. Difference between G–X–G's and A–X–A's BPO for all three amino acids also confirms this. Nevertheless, there are more BPOs than previous categories.

$\chi_{1+2+3+4}$ amino acids

As summarized in Table 7, 304,755 triplets were encountered for Lys and Arg, of which 295,191 were considered. Despite having equal number of χ angles, it can be deduced that Arg can adopt more conformations than Lys because the ratio between numbers of high frequency rotamers and numbers of encountered triplets (columns 5 and 3 of Table 7, respectively) is higher for Arg. The percent of low frequency rotamers also shows an increase in comparison with the previous categories of amino acids.

In general, the SDs of the $\chi_{1+2+3+4}$ amino acids are smaller than of the χ_{1+2+3} amino acids and for Arg and Lys, the SDs are similar. The range of variations of χ angle means for the $\chi_{1+2+3+4}$ amino acids within the majority of triplet rotamers is about 20°,but with more than 30° variations in few cases (Fig. E and Table D of SMs-File 4). For each amino acid rare conformations which have accepted frequency for small numbers of triplets reveal specificity in conformational selection. In other words some conformations (mostly in two recent categories of amino acids) happen with just specific or even very specific immediate adjacent residues. In Tables B, C and D of SMs-File 4, rows are sorted based on the numbers of low frequency conformations for each residue type and in last rows of each residue type side-chain conformations are more specific for few number of triplets.

For the $\chi_{1+2+3+4}$ amino acids, like the previous category, top five RFs were considered in the definition of BPO patterns and 122 different BPOs were defined, with number 122 for all the cases with equalisations. As can be seen in Fig. F of SMs-File 4, except for the number 122 which represents more than one type of patterns, it is hardly possible to find two equal BPOs number for Arg and Lys. This indicates that conformational preferences of Arg and Lys are almost completely dependent on the adjacent residues on the backbone.

In case of χ_{1+2+3} and $\chi_{1+2+3+4}$ amino acids, they have more χ angles and therefore, more degrees of freedom to form various side-chain conformations and more asymmetrical conformations, and this can explain their high dependency on adjacent residues in choosing their side-chain conformation.

BPO pattern in nearly all amino acids also could get some effects from nearby residues other than adjacent ones in the sequence but in this cases also the sequence of the triplet has a relative effect in choosing these nearby residues in 3D environment of the triplet.

Comparison of SDRL with other RLs

For comparing the SDRL with other backbone independent RLs, we considered number of rotamers plus their percent of covered real side-chain conformations. Table 8 summarizes results of comparing the SDRL with Lovell–Richardsons and Dynameomics. It becomes evident that despite all benefits that SDRL has in comparison with other backbone independent RLs this parameter for the SDRL is in average compatible or even better than Lovell–Richardsons and Dynameomics.

Table 3 General properties of rotamers for the χ1+2 amino acids.

Amino acid	Triplet case No.		Low frequency triplet rotamers	No. of rotamers With at least 10 cases
	Not filtered	Filtered		
LEU	264682	264051	2004/3600(55.67%)	1596
ILE	163133	162905	1837/3600 (51.03%)	1763
PHE	116160	116021	1650/3600 (45.83%)	1950
TRP	40818	40776	2292/3600 (63.67%)	1308
TYR	101053	100895	1692/3600 (47.00%)	1908
ASN	120040	119470	1193/3600 (33.14%)	2407
HIS	66818	66554	1582/3600 (43.94%)	2018
ASP	168943	167895	967/3600(26.86%)	2633

Table 4 An analysis of BPO patterns for the eight χ_{1+2} amino acids; Second column is included of number of patterns types with non-zero quantity in figure C of SMs-File 4 and numbers in 3th and 4th column are BPO patterns for G-X-G and A-X-A triplets.

	Patterns types	G-X-G	A-X-A	Non G-X-G (%)
LEU	5	1	1	25.7
ILE	6	3	1	81.2
PHE	10	1	7	55.2
TRP	17	6	7	92.5
TYR	10	1	7	55.5
ASN	9	1	1	54.7
HIS	16	3	2	76.5
ASP	9	5	1	91.5

Table 5 General properties of rotamers for the χ_{1+2+3} amino acids.

Amino acid	Triplet case No.		Low frequency triplet rotamers	No. of rotamers With at least 10 cases
	Not filtered	Filtered		
MET	48301	48168	9207/10800 (85.25%)	1593
GLU	189735	185605	6639/10800 (61.47%)	4161
GLN	104238	102908	7933/10800 (73.45%)	2867

Table 6 An analysis of BPO patterns for the χ_{1+2+3} amino acids; Second column is included of the number of patterns types with non-zero quantity in figure D of SMs-File 4.

	patterns types	G-X-G	A-X-A	Non G-X-G (%)
MET	52	**25**	**122**	91.7
GLU	42	**25**	**14**	96.7
GLN	43	**25**	**3**	95.0

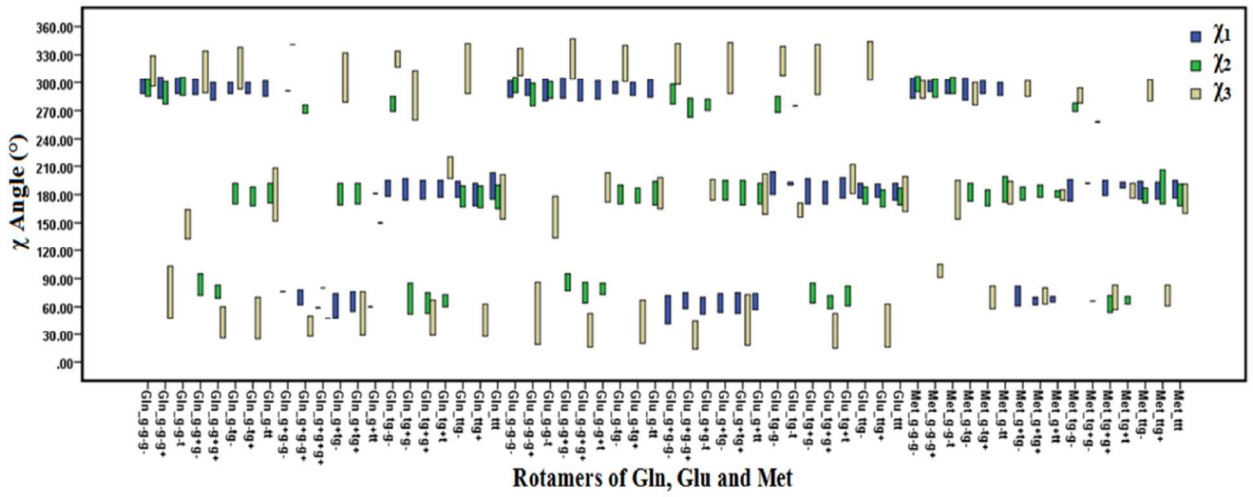


Fig. 4 Range of variation of χ angle averages within 400 triplets of χ_{1+2+3} residue types.

For the χ_{1+2} amino acids, three RLs are generally on par with each other. However, the SDRL in case of Leu, Ile, Trp, Asn and His is better and in case of Phe, Tyr and Asp is worse than Lovell–Richardsons. In comparison with Dynameomics, the SDRL in case of Leu, Phe, Trp, Tyr, Asn and His is better and in case of Ile and Asp is on par or a little worse. For the χ_{1+2+3} amino acids, the SDRL in case of Met is better and in case of Glu and Gln is worse than Lovell–Richardsons. In comparison with Dynameomics, the SDRL in case of Gln is a little better and for Glu is worse. They are almost on par for Met. Comparison of three RLs in case of the $\chi_{1+2+3+4}$

amino acids shows that the SDRL is much better than Lovell–Richardsons RL and on average is marginally better than Dynameomics.

It is kind of hard to justify but in Dynameomics very broad samples were used to calculate each rotamer and in the SDRL also there are many experimental conformations for each rotamer. As we have calculated in this paper, amount of neighbor dependency of different amino acids for side-chain conformational selection is very different and for some of them it is very high. Therefore, considering triplets (neighbor residues) in calculating rotamer parameters could act in favor of decreasing the number of rotamers.

It can be noticed that performance of amino acids is not consistent among three RLs in this comparison. While it is hard to provide an exact rational for all the observed cases, some plausible explanations are: different policies in bin selection for some amino acids such as Glu and Gln; diverse ideas for accomplishing RL; and difference in structural sampling methods (e.g. number of protein structures).

Table 7 General properties of rotamers for the $\chi_{1+2+3+4}$ amino acids.

Amino acid	Triplet case No.		Low frequency triplet rotamers	No. of rotamers With at least 10 cases
	Not filtered	Filtered		
LYS	159578	152831	28861/32400 (89.01%)	3539
ARG	145177	142360	28399/32400 (87.65%)	4001

Table 8 Number of rotamers and percent of their included real conformations of the three RLs for the χ_{1+2}, χ_{1+2+3} and $\chi_{1+2+3+4}$ amino acids.

	Dynameomics[a]	Lovell-Richardsons[a]	SDRL[b]
LEU	5 (98.51)	5 (93)	4.99 (98.67)
ILE	5 (98.65)	7 (99)	5.89 (99.22)
PHE	5 (93.05)	4 (98)	4.99 (94.42)
TRP	6 (89.18)	7 (94)	6.09 (97.62)
TYR	6 (96.94)	4 (98)	5.67 (97.11)
ASN	11 (97.87)	7 (94)	7.41 (98.19) and 6.82 (96.39)
HIS	11 (91.70)	8 (94)	7.28 (98.14)
ASP	6 (95.62)	5 (96)	6.87 (96.28)
MET	9 (90.23)	13 (86)	9.64 (90.64)
GLU	9 (90.64)	8 (91)	9.41 (83.89)
GLN	10 (85.52)	9 (88)	9.82 (86.52)
LYS	19 (91.93)	27 (81)	19.30 (91.66)
ARG	27 (90.92)	34 (82)	26.17 (91.77)
AVE.	**9.92 (93.13)**	**10.61 (91.85)**	**9.48 (93.93)**

[a] Number of rotamers; real case percent in parenthesis.
[b] Average number of rotamers; real case percent in parenthesis.

Conclusions

In this work as a backbone independent SDRL, our results for eighteen studied amino acids demonstrate that there are large variations in BPO patterns of most of the amino acids and also χ angle means and RF values for the same bin among 400 triplets of the same amino acid. This is a clear indication that adjacent residues on the protein backbone have substantial impact on the conformational preferences of the side-chains. As a result, by using the SDRL, side-chain repacking algorithms can find preferred conformations of each residue more easily than other backbone independent RLs. There is a possibility that the SDRL in a backbone dependent manner could predict side-chain conformation with a higher accuracy, which can be a subject of future studies.

The results were also analyzed in order to find some structural insights for individual amino acids. For instance, among the χ_1 amino acids, Val receives minimum impact from adjacent residues in choosing side-chain conformations. On the contrary, Cys mostly choose its side-chain conformations based on the adjacent residues. Among the χ_{1+2} amino acids, Leu behaves like Val, while Ile, Trp and Asp are the same as Cys. The χ_{1+2+3} and $\chi_{1+2+3+4}$ amino acids only select their side-chain conformations based on the adjacent residues on the backbone. Finally, these efforts with lightening up more structural roles for each amino acid could help protein engineering and design and also it could be helpful in solving problems like folding process.

Acknowledgements

This work was supported by research grant of Tehran university and we want to thanks Dr Sharifi-Zarchi and Dr Froughmand for some technical information

Notes and references

Institute of Biochemistry and Biophysics (IBB), Tehran university, Tehran, Iran.
Corresponding Author: Bahram Goliaei PhD, Laboratory of Biophysics and Molecular Biology, Institute of Biochemistry and Biophysics, University of Tehran, P. O. Box 13145-1384, Tehran, Iran. Tel: +98-21-66498672, Fax: +98-21-66956985, E-mail: goliaei@ibb.ut.ac.ir.

† Electronic Supplementary Information (ESI) available: Supplementary information included of Supplementary Materials (SMs)-File 1, SMs-File 2, SMs-File 3 and SMs-File 4. See DOI: 10.1039/b000000x/

1. R. L. Dunbrack, Jr., *Curr Opin Struct Biol*, 2002, **12**, 431-440.
2. R. Chandrasekaran and G. N. Ramachandran, *Int J Protein Res*, 1970, **2**, 223-233.
3. M. Lu, A. D. Dousis and J. Ma, *Protein Sci*, 2008, **17**, 1576-1585.
4. G. G. Krivov, M. V. Shapovalov and R. L. Dunbrack, Jr., *Proteins*, 2009.
5. R. Gautier, A. C. Camproux and P. Tuffery, *Nucleic Acids Res*, 2004, **32**, W508-511.
6. P. Francis-Lyon and P. Koehl, *Proteins*, 2014, **82**, 2000-2017.
7. M. V. Shapovalov and R. L. Dunbrack, Jr., *Structure*, 2011, **19**, 844-858.
8. K. Nagata, A. Randall and P. Baldi, *Proteins*, 2012, **80**, 142-153.
9. L. Quan, Q. Lu, H. Li, X. Xia and H. Wu, *BMC Bioinformatics*, 2014, **15 Suppl 12**, S5.
10. A. Bhowmick and T. Head-Gordon, *Structure*, 2014.
11. L. Heo, H. Park and C. Seok, *Nucleic Acids Res*, 2013, **41**, W384-388.
12. E. Mashiach, D. Schneidman-Duhovny, N. Andrusier, R. Nussinov and H. J. Wolfson, *Nucleic Acids Res*, 2008, **36**, W229-232.
13. Y. Gao, D. Douguet, A. Tovchigrechko and I. A. Vakser, *Proteins*, 2007, **69**, 845-851.
14. P. Carter, V. I. Lesk, S. A. Islam and M. J. Sternberg, *Proteins: Structure, Function, and Bioinformatics*, 2005, **60**, 281-288.
15. C. Wang, O. Schueler-Furman and D. Baker, *Protein Science*, 2005, **14**, 1328-1339.
16. O. Schueler-Furman, C. Wang and D. Baker, *Proteins: Structure, Function, and Bioinformatics*, 2005, **60**, 187-194.
17. T. Kirys, A. M. Ruvinsky, A. V. Tuzikov and I. A. Vakser, *Proteins*, 2012, **80**, 2089-2098.
18. J. J. Headd, R. M. Immormino, D. A. Keedy, P. Emsley, D. C. Richardson and J. S. Richardson, *J Struct Funct Genomics*, 2009, **10**, 83-93.
19. M. T. Stiebritz and Y. A. Muller, *Acta Crystallogr D Biol Crystallogr*, 2006, **62**, 648-658.
20. W. L. DeLano, 2002.
21. R. D. Taylor, P. J. Jewsbury and J. W. Essex, *J Comput Chem*, 2003, **24**, 1637-1656.
22. S. B. Nabuurs, M. Wagener and J. de Vlieg, *J Med Chem*, 2007, **50**, 6507-6518.
23. S. C. Lovell, J. M. Word, J. S. Richardson and D. C. Richardson, *Proteins*, 2000, **40**, 389-408.
24. R. L. Dunbrack, Jr. and F. E. Cohen, *Protein Sci*, 1997, **6**, 1661-1681.
25. M. J. McGregor, S. A. Islam and M. J. E. Sternberg, *Journal of Molecular Biology*, 1987, **198**, 295-310.
26. G. Chinea, G. Padron, R. W. Hooft, C. Sander and G. Vriend, *Proteins*, 1995, **23**, 415-421.
27. M. J. Bower, F. E. Cohen and R. L. Dunbrack, Jr., *J Mol Biol*, 1997, **267**, 1268-1282.
28. R. W. Peterson, P. L. Dutton and A. J. Wand, *Protein Sci*, 2004, **13**, 735-751.
29. A. D. Scouras and V. Daggett, *Protein Sci*, 2011, **20**, 341-352.
30. G. Wang and R. L. Dunbrack, *Bioinformatics*, 2003, **19**, 1589-1591.
31. B. Goliaei and Z. Minuchehr, *FEBS Lett*, 2003, **537**, 121-127.
32. Z. Minuchehr and B. Goliaei, *Protein Pept Lett*, 2005, **12**, 379-382.
33. N. A. Fonseca, R. Camacho and A. L. Magalhaes, *Proteins*, 2008, **70**, 188-196.
34. S. Anishetty, G. Pennathur and R. Anishetty, *BMC Struct Biol*, 2002, **2**, 9.
35. C. L. Kingsford, B. Chazelle and M. Singh, *Bioinformatics*, 2005, **21**, 1028-1036.
36. G. Wang and R. L. Dunbrack, *Nucleic Acids Research*, 2005, **33**, W94-W98.
37. P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlić, M. Quesada, G. B. Quinn and J. D. Westbrook, *Nucleic Acids Research*, 2011, **39**, D392-D401.
38. D. C. Richardson, *Dangle Software*, (2007), **Richardson Lab**, Duke University-USA.
39. R. L. Dunbrack, Jr. and M. Karplus, *J Mol Biol*, 1993, **230**, 543-574.

40. S. J. Shandler, M. V. Shapovalov, R. L. Dunbrack, Jr. and W. F. DeGrado, *J Am Chem Soc*, 2010, **132**, 7312-7320.
41. A. D. Scouras and V. Daggett, *Protein Science*, 2011, **20**, 341-352.
42. R. E. Smith, S. C. Lovell, D. F. Burke, R. W. Montalvao and T. L. Blundell, *Bioinformatics*, 2007, **23**, 1099-1105.
43. P. Tuffery, C. Etchebest, S. Hazout and R. Lavery, *J Biomol Struct Dyn*, 1991, **8**, 1267-1289.
44. J. W. Ponder and F. M. Richards, *J Mol Biol*, 1987, **193**, 775-791.

