PROTEINS: Structure, Function, and Genetics 20:301–311 (1994)

RESEARCH ARTICLES

Intrinsic Secondary Structure Propensities of the Amino Acids, Using Statistical ϕ–ψ Matrices: Comparison With Experimental Scales

Victor Muñoz and Luis Serrano
European Molecular Biology Laboratory, Structures and Biocomputing, Heidelberg, Germany

ABSTRACT **Today there are several different experimental scales for the intrinsic α-helix as well as β-strand propensities of the 20 amino acids obtained from the thermodynamic analysis of various model systems. These scales do not compare well with those extracted from statistical analysis of three-dimensional structure databases. Possible explanations for this could be the limited size of the databases used, the definitions of intrinsic propensities, or the theoretical approach. Here we report a statistical determination of α-helix and β-strand propensities derived from the analysis of a database of 279 three-dimensional structures. Contrary to what has been generally done, we have considered a particular residue as in α-helix or β-strand conformation by looking only at its dihedral angles (ϕ–ψ matrices). Neither the identity nor the conformation of the surrounding residues in the amino acid sequence has been taken into consideration. Pseudoenergy empirical scales have been calculated from the statistical propensities. These scales agree very well with the experimental ones in relative and absolute terms. Moreover, its correlation with the average of the experimental scales for α-helix or β-strand is as good as the correlations of the individual experimental scales with the average. These results show that by using a large enough database and a proper definition for the secondary structure propensities, it is possible to obtain a scale as good as any of experimental origin. Interestingly the ϕ–ψ analysis of the Ramachandran plot suggests that the amino acids could have different β-strand propensities in different subregions of the β-strand area.**

Key words: dihedral angles, intrinsic propensities, protein stability, secondary structure, protein folding

INTRODUCTION

The fact that the 20 natural amino acids have different intrinsic propensities to populate the secondary structure elements is generally accepted. Quantitation of the thermodynamic scales for α-helix or β-strand formation is one important step in the direction of solving the protein folding problem.[1] Methods based on the statistical analysis of the distribution of the 20 different amino acids in the different secondary structure elements, using databases of three-dimensional (3-D) protein structures, have been described to calculate these propensities.[1–6] The secondary structure elements are normally defined in terms of angles and hydrogen bonds and it is required that some consecutive residues adopt the same conformation.

Recently, several model systems have been developed to experimentally calculate the intrinsic properties for the natural amino acids to form α-helices[7–11] and β-strands.[12,13]

The theoretical scales vary widely, probably as a result of the databases used as well as of the different definitions of the secondary structure elements. Something similar occurs for the experimental scales, in this case probably as a result of different context effects in the model systems utilized. However, comparison between experimental and theoretical scales shows in general a much greater discrepancy.[7–10,12,13] So far, two reasons have been postulated to explain this discrepancy: either the statistical approach is inadequate or the databases used are too small to obtain statistically meaningful data.

In this work we report the use of a database of 279 3-D structures[14] to calculate statistical ϕ–ψ matrices for the 20 naturally occurring amino acids. These matrices are then used to determine the intrinsic propensity of the amino acids to be in α-helix and β-strand dihedral angles. The amino acid propensities are translated to free energies using a sort of mean-force potentials. Values obtained from this approach are directly compared with the available experimental scales.

Received June 3, 1994; revision accepted August 8, 1994.
Address reprint requests to EMBL, Structures and Biocomputing, Meyerhofstrasse 1, Heidelberg, Germany.

METHODS

Determination of the Statistical Scale of Helical and β-Strand Propensities Using ϕ–ψ Matrices

The database of 3-D structures has been obtained following the principles described by Hobohm and co-workers.[15] This database has been filtered for quality of the data and consists of 279 proteins with less than 50% sequence homology for a total of 59,117 amino acidic residues and it is currently included in the program WHATIF.[16] The following proteins have been considered in the database: 1aai; 1aaj; 1aak; 1aap; 1abk; 1abm; 1ads; 1arb; 1aso; 1atn; 1avh; 1ayh; 1baa; 1bbh; 1bbk; 1bbp; 1bbt; 1bmv; 1bop; 1bov; 1btc; 1caj; 1cbn; 1cbx; 1ccr; 1cd8; 1cdt; 1cid; 1clm; 1cmb; 1cox; 1cpc; 1cpl; 1cse; 1d66; 1dfn; 1dhr; 1dri; 1eaf; 1eco; 1end; 1etu; 1ezm; 1fas; 1fba; 1fc1; 1fc2; 1fdd; 1fha; 1fia; 1fnr; 1fxi; 1gky; 1gla; 1gly; 1gmf; 1gox; 1gp1; 1gpr; 1grc; 1grd; 1gst; 1hc6; 1hdd; 1hge; 1hil; 1hlh; 1ifc; 1isu; 1izb; 1lap; 1lig; 1lpe; 1lts; 1lz3; 1mam; 1mbd; 1mdc; 1min; 1mrm; 1ms2; 1mup; 1nip; 1nxb; 1ofv; 1omf; 1omp; 1ova; 1ovb; 1paf; 1pbx; 1pda; 1pfk; 1pgd; 1phg; 1phh; 1plc; 1ppb; 1ppf; 1ppl; 1ppn; 1ppt; 1prc; 1pya; 1pyp; 1r09; 1r1a; 1rbp; 1rcb; 1rhd; 1rnd; 1rve; 1s01; 1sas; 1sdh; 1sgt; 1sha; 1snc; 1spa; 1tab, 1ten; 1tfg; 1tgs; 1tho; 1tie; 1tlk; 1tmd; 1tnf; 1trb; 1tro; 1ttb; 1ula; 1utg; 1vaa; 1vsg; 1wsy; 256b; 2aaa; 2ach, 2avi; 2aza; 2bpa; 2ccy; 2cdv; 2cmd; 2cro; 2cts; 2cyp; 2dnj; 2gls; 2had; 2hip; 2lbp; 2ltn; 2mad; 2mev; 2mhr; 2msb; 2pf2; 2pia; 2plv; 2pmg; 2por; 2ren; 2rn2; 2scp; 2sga; 2sic; 2sn3; 2snv; 2stv; 2tbv; 2tmv; 2zta; 3adk; 3b5c; 3cd4; 3chy; 3cla; 3dfr; 3gbp; 3grs; 3il8; 3ink; 3pgk; 3rub; 3sc2; 3sgb; 3sod, 3tgl; 451c; 4blm; 4bp2; 4cpa; 4enl; 4fgf; 4fxn; 4gcr; 4gpd; 4icd; 4rcr; 4rxn; 4sbv; 4sgb; 4tms; 4ts1; 5fbp; 5nn9; 5p21; 7api; 7tim; 7xia; 8abp; 8adh; 8atc; 8cat; 8i1b; 9ldt; 9rnt; 9rub; 9wga; 1aba; 1bbt; 1bib; 1c2r; 1col; 1gmp; 1hge; 1ipd; 1lmb; 1mba; 1nsb; 1pii; 1pya; 1rnb; 1thg; 1yat; 2hhm, 2lh7; 2mcm; 2rsp; 3rub; 4dfr; 1ak3; 1alc; 1cc5; 1ctf; 1cy3; 1dtx; 1fx1; 1gal; 1gsr; 1hip; 1hrh; 1ith; 1mda; 1mpp; 1paz; 1ppb; 1pts.

To build up the ϕ–ψ matrix for every amino acid, we considered a representation of the ϕ–ψ space equivalent to the Ramachandran plot and 20 discrete intervals of 18° for both ϕ and ψ dihedral angles producing matrices of 20×20 values. Each of the defined values consists of a square of $18 \times 18°$ in the ϕ–ψ space and may be addressed using the matrix nomenclature (row, column). A row corresponds to the index of the ψ interval from 1 (180 to 162°) to 20 (−162 to −180°) and a column to the index of the ϕ interval from 1 (−180 to −162°) to 20 (162 to 180°) (Fig. 1). As an example the interval (1,1) covers a region of the Ramachandran plot corresponding to ϕ (−180 to −162) and ψ (180 to 162). The search in the database was done with the module SCAN3D of WHATIF. This module consists of a special search method based on property profiles that rapidly find sequence–structure relations.[17] The SEQUEN option of WHATIF is used to calculate the total number of hits for a particular amino acid in the database. To calculate the number of times that a particular amino is found in an interval of ϕ, ψ dihedral angles the option PHI-PSI is utilized. The result of this search is 20 ϕ–ψ matrices (one for each amino acid), of 20×20 squares that may be viewed as statistical replicates of the Ramachandran plot with specific energy levels for each amino acid.

The tendency of a certain amino acid to populate dihedral angles belonging to any of the defined regions arises from the ratio between the number of hits in the sum of intervals comprised in the region being considered (Σ Intervals) and the total number of hits for the specific amino acid in the database [Eq. (1)]:

$$X_{\text{propensity}} = (\Sigma \text{ Intervals})/\text{Total number}. \quad (1)$$

If the protein database is large enough and it does not include proteins with a high degree of homology, it is possible to assume that it reflects a system in thermodynamic equilibrium. Under these principles we consider that the distribution of the dihedral angles for the different amino acids follows a Boltzmann's distribution. We define two states, one restricted to the defined ϕ–ψ region and the other consisting of the whole Ramachandran plot where the amino acid may explore freely all the conformational space. The free energy required for the transition from the free state to a fixed state, corresponding to the defined ϕ–ψ region, is directly calculated from Eq. (2) using the statistical propensities.

$$\Delta G_{X\text{stat}(i)} = -RT \ln(X_{\text{propensity}}) \quad (2)$$

where $\Delta G_{X\text{stat(i)}}$ is the difference in free energy necessary to fix the amino acid X in the i ϕ–ψ region being considered, R is 0.00198 kcal mol^{-1} K^{-1}, and T is the temperature in Kelvin. The free energies calculated in this way are empirical pseudoenergies but can be directly compared with experimental free energies (see Results).

Determination of the Statistical Scale of Helical and β-Strand Propensities Using Secondary Structure Definitions

An alternative way to determine the intrinsic propensities is to look for the relative preferences of the amino acids to be in a defined secondary structure element. For this we have used the Kabsch and Sander[18] definition of secondary structure in combination with the WHATIF program. In the case of α-helices we have searched for the secondary structure motif HHHHHHHHH, and looked to the statistical preferences at the three central positions, ---XXX---. In this way we eliminate the preferences of certain residues to be at the first, or last, three positions of an α-helix,[1,19] because of the interaction

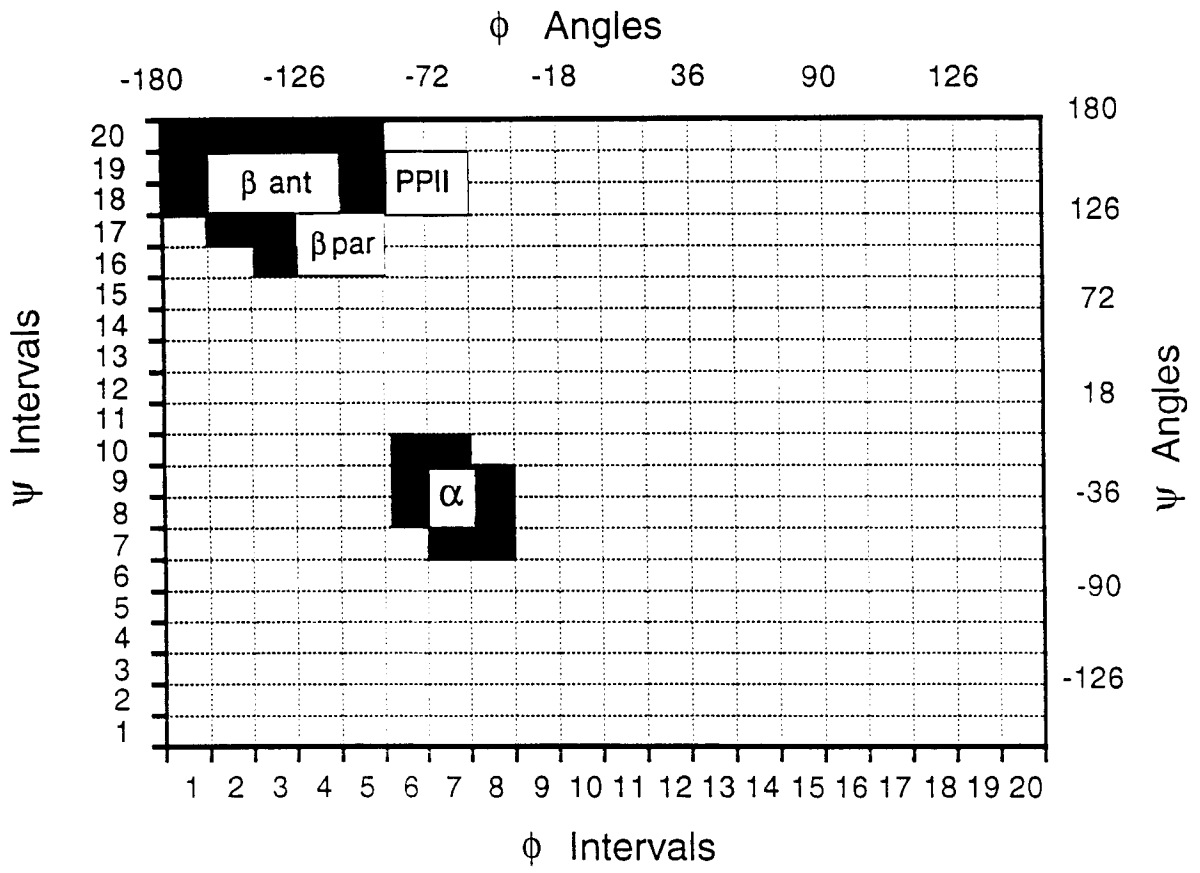

Fig. 1. Ramachandran plot showing its division in 20 dihedral angle intervals. Some of the selected regions for calculating the β-strand and α-helical propensities are shown in white with the corresponding labeling: α for α-helix, β-par for the β-parallel region, and β-ant for the β-antiparallel region.[24] The maximum area explored in the α-helix or β-strand region (excluding the PPII region) is shown in black.

with the helix dipole. In the case of the β-strand preferences we have just looked for those residues which are in a β-strand conformation following the Kabsch and Sander[18] definition. To calculate the differences in free energy the same procedure described above was used. In this case, statistical propensities are calculated dividing the number of times a residue is found in α-helix or β-strand by the number of times it is found in the database. The α-helix scale calculated this way is named from now on in this paper, helix scale, and the β-strand scale (beta).

To compare previously published statistical scales with the experimental ones and with those presented here it is necessary to translate them to pseudoenergy scales. The translation has been done introducing directly the observed frequencies in Eq. (2).

Normalization of the Experimental Scales

Each experimental scale has its own problems derived from the system being used and the way in which the data have been analyzed. In some cases the flanking residues have not been mutated to Ala, in other cases the systems have been analysed at high ionic strength, and so on. However, the results are related, indicating that they are measuring the same phenomena and that the differences could be due to context effects and/or the theoretical background behind the analysis. In principle, when using several experimental scales their average value should diminish the context effects and result in closer values to the real ones. The range of free energy differences between the amino acids is quite different in some experimental scales. In the case of the α-helices, the range of values in the scale of Chakrabartty et al.[10] is larger than in the other experimental scales (1.88 vs. ~0.8 kcal mol^{-1} difference in energy between Ala and Gly, respectively).[7–9,11] For the β-strand scales we find the same problem (1.2 kcal mol^{-1} difference between Ala and Gly in the scale of Minor and Kim,[13] versus 0.44 kcal mol^{-1} in the scale of Kim and Berg[12]). To obtain an average scale, to which we can compare our ϕ–ψ derived scales, it is necessary to normalize the maximum difference between two residues between different scales. Normalization precludes biases produced on the average scale by those with

higher differences in energy. To normalize the different scales we have subtracted the free energy value of the amino acid which is more favorable for α-helix or β-strand formation, in each scale, from all the amino acids. Then we have divided the resulting values by the intrinsic value of the amino acid which is less favorable in the particular scale (Pro has not been considered, since the difference in energy with *the rest* of the amino acids is too high and will bias the normalisation). In this way we obtain scales in the range 0 to 1. The average scale is calculated from the normalized scales.

RESULTS

α-Helical Propensities

The definition for the specific α-helix dihedral angles is ambiguous and may strongly influence the results obtained from this kind of analysis. For this reason we have selected several different regions of the Ramachandran plot corresponding to the α-helical region and calculated the intrinsic propensities in free energy terms as indicated in methods (Fig. 1). The most restrictive region consists of a single square with angles ϕ (−72 to −54) and ψ (−54 to −36), interval (7,13). This dihedral angle region contains the majority of the amino acids found in α-helices, by the method of Kabsch and Sander.[18] The next most populated interval is 7,12 and the more relaxed case, still having a significant number of Ala residues, comprises the intervals 6,11; 7,11; 6,12; 7,12; 8,12; 6,13; 7,13; 8,13; 7,14; and 8,14. Correlation analysis with the five experimental scales as well as with the normalized average scale indicates that the best region of the Ramachandran plot corresponds to the intervals 7,13 and 7,12 (except for Pro; data not shown). The number of hits for each amino acid in this region is shown in Table I. The absolute free energy values for the different amino acids are obtained as explained in the Methods section (see Table II). For alanine using this restricted region the absolute value is 0.62 kcal mol^{-1}. Considering that the enthalpy of hydrogen bond formation in an α-helix is approximately 1.0±0.2 kcal mol^{-1},[20] the elongation parameter, *s*, will be between 1.35 and 2.7. This range of values is close to the experimental ones found in the different propensity scales,[7–9,11,12] as well as in a designed peptide in which the nucleation process of helix formation was eliminated.[21]

In Table I we also show the number of hits for each amino acid using the second approach described in methods (helix scale). The absolute free energy terms for each amino acid calculated as indicated in methods are shown in Table II.

Comparison With the α-Helix Experimental Scales

Five experimental scales derived from different model systems: two from site-directed mutagenesis in proteins[8,9] and three from peptides[7,10,11] have been used to evaluate the quality of the statistical data. Correlation analysis of the different experimental scales (without including Pro) indicates that all of them are related, the most dissimilar scales being those derived from barnase[8] and T4 lysozyme[9] ($r<0.71$) (Table III). The correlation slopes are similar in all the cases (data not shown), except when comparing the other scales with that of Chakrabartty et al.[10] In this case the free energy differences between the different amino acids are larger.

Comparison of the pseudoenergy scales, derived from the statistical scales, of Richardson and Richardson[1] or Chou and Fassman,[2] with the experimental ones shows a poor correlation ($r<0.8$). The same happens with the scale derived from the data of Sander et al.,[22] (data not shown). The statistical scale derived from the distribution of the amino acids at the three central positions of a nine residue

TABLE I. Number of Hits in the Database for the Natural Amino Acids Using Definitions for α-Helices and β-Strands Based on Different Criteria

	Total*	α-Helix†	α-ϕ–ψ‡	β§	β-ϕ–ψ**	β-res††
Ala	5022	2424	1774	812	964	471
Cys	1024	233	158	297	384	148
Asp	3527	812	732	416	612	144
Glu	3524	1243	1128	565	695	318
Phe	2420	735	472	761	902	395
Gly	4850	682	488	752	453	149
His	1282	331	224	278	378	156
Ile	3154	1214	719	1178	1356	565
Lys	3454	1223	920	611	836	359
Leu	4869	2115	1397	1286	1336	515
Met	1232	582	356	313	360	186
Asn	2745	595	437	364	586	181
Pro	2796	105	435 (138)‡‡	258	34	4
Gln	2102	771	573	357	490	220
Arg	2607	1115	731	505	694	328
Ser	3732	779	727	706	993	490
Thr	3603	810	609	960	1379	679
Val	4101	1246	841	1631	1939	949
Trp	897	280	193	251	289	139
Tyr	2176	585	396	708	845	416

*Total number of amino acids in the protein database.
†Number of hits per amino acid in the database, when looking for the three central residues of a nine helix defined by the Kabsch and Sander method.[19]
‡Number of hits per amino acid in a region of the Ramachandran plot corresponding to two of the ϕ–ψ intervals (7,12 and 7,13).
§Number of hits per amino acid in the database when looking for those residues in a β-strand conformation using the definition of Kabsch and Sander.[19]
**Number of hits per amino acid in the β region of the Ramachandran plot corresponding to the intervals 1,1 to 5,1; 1,2 to 5,2; 1,3 to 5,3; 2,4 to 5,4; 3,5 to 5,5.
††Number of hits per amino acid in the β region of the Ramachandran plot corresponding to the intervals, 2,2 to 2,4 and 3,2 to 3,4.
‡‡The number of proline residues when looking at a more restricted region of the ramachandran plot (interval 7,13).

TABLE II. α-Helical and β-Strand Propensities for the Different Amino Acids in kcal mol^{-1}

	α-Helix*	α-helix ϕ–ψ†	β‡	β ϕ–ψ§	β-res**
A	0.432	0.617	1.080	0.978	1.40
C	0.877	1.107	0.733	0.573	1.14
D	0.870	0.932	1.266	1.038	1.89
E	0.167	0.675	1.085	0.962	1.42
F	0.706	0.968	0.685	0.585	1.07
G	1.162	1.361	1.104	1.405	2.06
H	0.802	1.034	0.906	0.724	1.25
I	0.566	0.876	0.583	0.502	1.02
K	0.615	0.784	1.026	0.841	1.34
L	0.494	0.740	0.789	0.766	1.33
M	0.444	0.736	0.812	0.729	1.12
N	0.906	1.089	1.197	0.915	1.61
P	1.945	1.780††	1.412	2.613	3.90
Q	0.594	0.770	1.050	0.863	1.33
R	0.503	0.753	0.976	0.784	1.23
S	0.928	0.969	0.987	0.784	1.20
T	0.884	1.053	0.784	0.569	0.99
V	0.706	0.939	0.546	0.444	0.87
W	0.690	0.910	0.755	0.671	1.10
Y	0.778	1.009	0.665	0.560	0.98

*Free energy in absolute terms required to fix the different amino acids in α-helical conformation following the definition of Kabsch and Sander.[18]

†Free energy in absolute terms required to fix the different amino acids in the α-helical region defined by two intervals of the Ramachandran plot (7,12 and 7,13).

‡Free energy in absolute terms required to fix the different amino acids in β-strand conformation following the definition of Kabsch and Sander.[18]

§Free energy in absolute terms required to fix the different amino acids in the β-strand region defined by the intervals, 1,1 to 5,1; 1,2 to 5,2; 1,3 to 5,3; 2,4 to 5,4; 3,5 to 5,5.

**Free energy in absolute terms required to fix the different amino acids in the β-strand region defined by the intervals 2,2 to 2,4 and 3,2 to 3,4.

††In the case of Pro we have used a more restrictive area (13,8). To the intrinsic value of Pro in α-helices or β-strands we should add the cost of not making a hydrogen bond (~1 kcal + mol^{-1}).[20]

helix using the Kabsch and Sander[18] definition (helix scale) results in a much better correlation with the experimental data (r>0.8), with the exception of the barnase[8] and O'Neil and DeGrado[7] scales (Table III). This means that the poor correlation of the previous statistical scales was the result of insufficient data or poorly choosen protein databases. When the helix scale was derived from considering all the residues in the database that are in α-helices, instead of the three central ones in a nine residue stretch in helical conformation, the correlation with the experimental data was not as good (data not shown). In this case Glu becomes too favorable when compared to the experimental scales.

The restricted ϕ–ψ scale results in better correlations with the experimental scales, than the helix scale. If we do not include Glu in the analysis, the correlation with the experimental scales is much better (Table III). The poor result of this scale for Glu could be due to the large statistical preference of Glu at the beginning of α-helices (Richardson and Richardson[1] and Dasgurpta and Bell[19]), which cannot be eliminated as in the case of the search of the three central residues in a nin-residue helix (see above). In Figure 2 we show the correlation analysis of the ϕ–ψ scale with the five experimental ones and the normalized average scale.

β-Strand Propensity

To derive a ϕ–ψ scale of β-strand propensities and due to the broad distribution of the different aminoacids in the β-region of the Ramachandran plot, we have explored the whole region using rectangular areas including different discrete intervals and a very large area covering the region of β-strand defined by Rooman et al.[23] (see Fig. 1). The region of the Ramachandran plot covered by the intervals (2,2; 2,3; 3,2; 3,3; 4,2, and 4,3) corresponds very closely to that of the antiparallel β-sheet conformation of L-alanine[24] (ϕ −142±13 and ψ 145±13) (β_{ant}, Fig. 1), while that covered by the intervals (4,4, 4,5, 5,4, and 5,5) corresponds approximately to the parallel β-sheet conformation of L-alanine[24] (ϕ −118±13, ψ 112±13) (β_{par}, Fig. 1). In Table I we show the number of hits for each aminoacid when considering the larger area covering the region of β-strand defined by Rooman et al.[23] (intervals 1,1 to 5,1; 1,2 to 5,2; 1,3 to 5,3; 2,4 to 5,4; 3,5 to 5,5) (Fig. 1). In the same table we also show the number of hits for each amino acid when looking in the database for those residues adopting a β-strand conformation as defined by Kabsch and Sander.[18] The absolute free energy values are shown in Table II.

Comparison With the β-Strand Experimental Scales

In the last 2 years, three groups using two completely different model systems have experimentally determined the thermodynamic scale of β-strand propensities for the 20 natural amino acids. The scale of Kim and Berg[12] has been obtained from substitutions in a zinc-finger peptide. The other two scales were obtained by Minor and Kim[13] and Smith et al.[14] from mutations in a central solvent-exposed position of a β-strand of the IgG-binding domain of protein G. The three scales show discrepancies, as occurred for the α-helical scales. The main differences between the two scales derived from the protein G are in the amino acids at positions $i-2$ and $i+2$ of the substituted residue. In the first case there were two Ser,[13] while in the second case there were two Thr.[14] This could explain the differences in free energy for some of the mutations (i.e., aromatics). The correlation between the Kim and Berg scale with the other two renders similar coefficients, ~0.77 (0.89 if we exclude Asp, which is clearly out of the correlation; Table IV), thus indicating that both

TABLE III. Correlation Analysis Between the Different Helical Propensity Scales in Terms of ΔG (kcal·mol^{-1}).

	Bar*	T4†	Agadir‡	Bald§	Average**	Phi-Psi††	R&R‡‡	Ch-Fas§§	Helix***
DGi	0.86	0.71	0.89	0.82	0.93	0.77 (0.84)	0.59	0.61	0.78
Bar		0.61	0.84	0.80	0.90	0.75 (0.78)	0.47	0.50	0.66
T4			0.89	0.85	0.86	0.84 (0.92)	0.78	0.67	0.92
Agadir				0.94	0.98	0.92 (0.95)	0.73	0.74	0.92
Bald					0.94	0.87 (0.91)	0.68	0.66	0.85
Average						0.90 (0.94)	0.68	0.69	0.88
ϕ–ψ†††							0.73 (0.84)	0.87 (0.86)	0.92 (0.94)
R&R								0.67	0.73
Ch-Fas									0.78

*Horovitz et al.[8]
†Blaber et al.[9]
‡Muñoz and Serrano[11] (The intrinsic values are slightly different from the published ones, since in a new more advanced version of the algorithm we have separated the helix dipole from the capping effect as well as introduced the temperature and pH dependence[28]).
§Chakrabartty et al.[10] The values used for the aromatic residues are those noncorrected for their presumed effect on the far-UV CD spectra of helical peptides. The correlation analysis with all the other scales, when using the corrected values for Tyr and Trp was worse in all of the cases (data not shown). For Glu we have used the intrinsic value when it is charged and for His when it is neutral.
**Average scale obtained from the normalized five experimental scales (see methods).
††Helical propensities derived from the ϕ–ψ matrices using the intervals (7,12 and 7,13). The results in brackets correspond to the correlations when Glu was not included (see results).
‡‡Richardson & Richardson.[1]
§§Chou & Fassman.[2]
***Helix propensities derived from a search in the database using a template of nine residues in helical conformation and looking at the statistical distribution of amino acids at the three middle positions.
†††O'Neil & DeGrado.[8]
Pro has not been included in any of the correlation analyses.

are measuring the same phenomenon and that the relative differences are likely due to context effects. More remarkable is the high slope found in the correlation (4 when correlating Kim and Berg's against Minor and Kim's). This is an indication of large differences in the absolute energy values and might be due to the location of the residue being mutated as well as to the system used. In one case the residue is located in a central position of a β-strand, which is forming a parallel β-sheet on one site an antiparallel on the other and whose main chain groups are involved in hydrogen bonds.[13,14] In the second case the mutated residue is in an edge β-strand belonging to a β-hairpin and its main chain groups do not make hydrogen bonds with the rest of the protein.[12] This allows a higher degree of flexibility to the residues placed in that position and consequently tend to blurr the energy differences.

The comparison of the pseudoenergy scale, derived from the statistical scale, of Chou and Fassman[2] with the two experimental ones and their normalized average, shows very weak correlations ($r<0.7$). Similar poor results are obtained with a scale derived from the work of Sander et al. (data not shown)[22] (Table IV). Only if we exclude glycine from the analysis do we obtain a good correlation coefficient ($r\sim0.8$) (data not shown). The correlation coefficients improve when we use the scale derived from the analysis of the new protein database,[15] considering the residues in a β-strand conformation using the Kabsch and Sander definition[18] (Table IV). A particular interesting case is that of Gly. In the statistical thermodynamic scales derived from looking at those residues which are part of β-strands in proteins, Gly is much more favorable than in the experimental scales. Comparison of a β-scale derived from the side chain hydrogen exchange blocking effects,[25] with all the other experimental and theoretical scales, results in similar correlation coefficients ($0.7>r<0.79$), with the exception of that from Minor and Kim[13] and Smith et al.[14] ($r\sim0.62$).

In Table IV we show the correlation coefficients of the ϕ–ψ scales mentioned above, with the different theoretical and experimental scales. The ϕ–ψ scale correlates very well with all the different scales, with the exception of the Chou and Fassman scale ($r=0.61$, eliminating Gly from the correlation raises r to 0.71). Moreover the ϕ–ψ scale is as similar to the normalized average scale as the three experimental ones. In Figure 3 we show the correlation analysis of the comparison between the ϕ–ψ scale and the experimental data of Kim and Berg[12] (Fig. 3A), Minor and Kim[13] and Smith et al.[14] (Fig. 3B), the β scale using the Kabsch and Sander definition[18] (Fig. 3C), and the average normalized scale (Fig. 3D).

DISCUSSION

The reason why different residues have different helical or β-strand propensities is not clear. In the classical helix-coil transition theory it was postulated that the nucleation factor σ was the same for all the amino acids, and that only the elongation

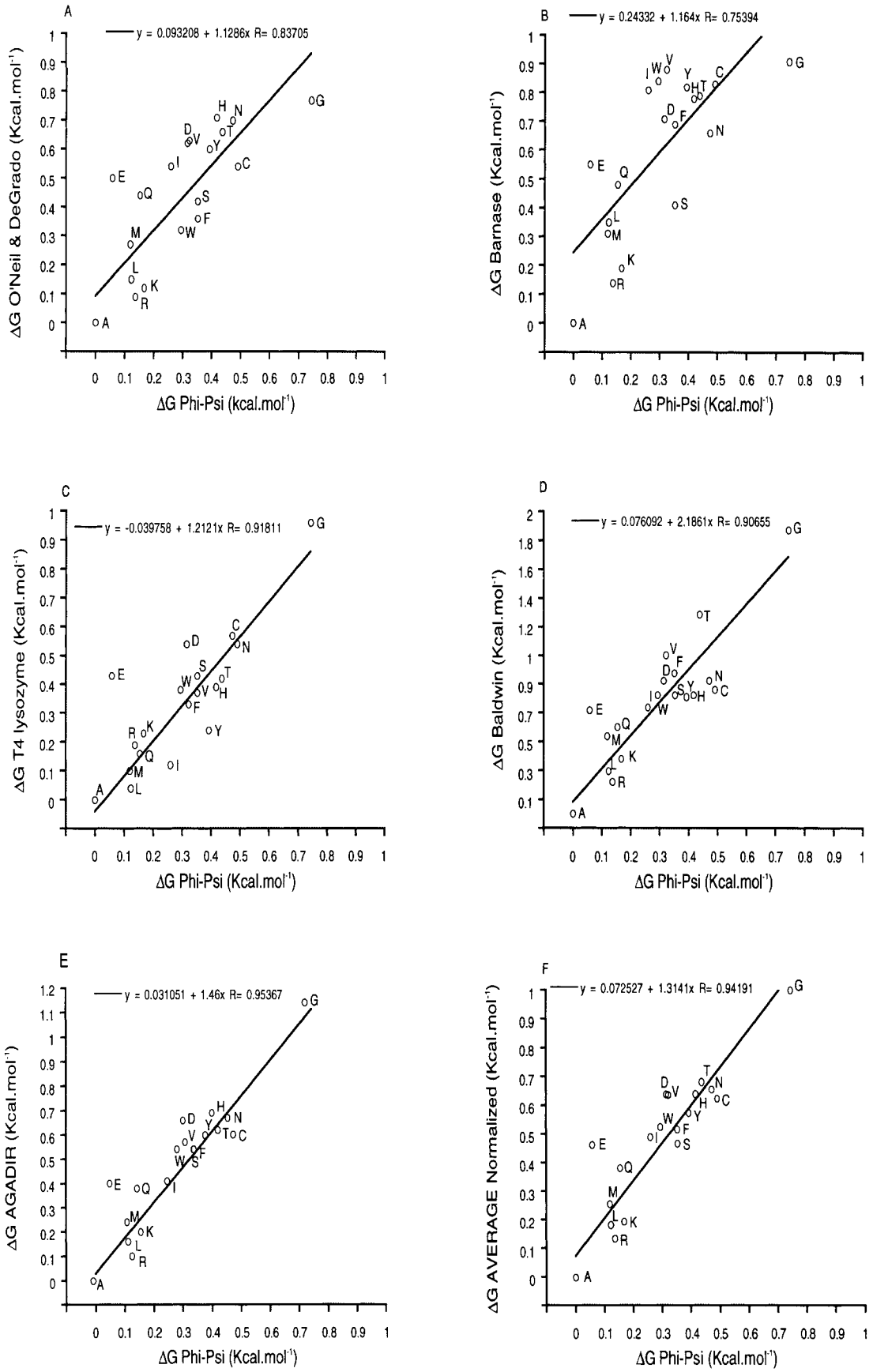

Fig. 2. Correlation analysis between the experimental scales for helical propensities and the one derived from the ϕ–ψ matrices. (**A**) O'Neil and DeGrado.[7] (**B**) Horovitz et al.[8] (**C**) Blaber et al.[9] (**D**) Chakrabartti et al.[10] (**E**) A refined version of the algorithm AGADIR that calculates effectively the helical content of more than 420 monomeric peptides in squeous solution[11] (Muñoz and Serrano[28]). (**F**) Normalized average scale of the five experimental scales.

TABLE IV. Correlation Analysis Between the Different β-Strand Propensity Scales in Terms of ΔG (kcal·mol^{-1}).

	Beta*	Ch & Fass†	Zinc‡	Prot.G§	Average**	H_{ex}††
Phi-Psi‡‡	0.85	0.61	0.90	0.91 (0.90)	0.95	0.70
Beta		0.89	0.69	0.80 (0.76)	0.77	0.74
Ch & Fass			0.47	0.69 (0.64)	0.63	0.78
Zinc				0.79 (0.76)	0.93	0.79
Prot.G					0.96	0.62
Average						0.73

*β-strand propensities derived from a search in the database looking for those residues in a β-strand conformation under the Kabsch and Sander definition.[18]
†Chou and Fassman.[2]
‡Thermodynamic scale of Kim and Berg.[12]
§Thermodynamic scale of Minor and Kim.[13] In bracketts we show the correspondence with the thermodynamic scale of Smith et al.[14]
**Average scale obtained from the normalized scales of Kim and Berg[12] and Minor and Kim.[13]
††The hydrogen exchange scale of Bai and Englander.[25]
‡‡β-strand propensities derived from the ϕ-ψ matrices using the data from Table 2 Pro has not been included in any of the correlation analysis. In all cases the intrinsic preferences were expressed in terms of free energy with respect to Alanine.

factor s was different.[26] This means that the helical propensities of the different residues should be related to the interaction of the side chain of a particular amino acid with the nucleated helix. The same reasoning could be applied to the β-strand.

A different way of explaining these differences is that the side chain of a particular residue favors certain ϕ–ψ angles because of steric reasons or/and because it facilitates the solvation of the side or/and main chain.[25] In the original construction of the Ramachandran plot by Ramachandran and co-workers[27] and based on steric reasons alone, it was indicated that different types of amino acids should have slightly different energy distributions over the plot. This becomes more evident when the conformational energy for a particular amino acid is calculated using bond torsional potentials and London dispersion interactions.[24] For example, the energy distribution for L-alanine in the β-region of the Ramachandran plot is very similar to the probabilities of finding Ala at the different intervals we have considered in the protein database[15] (see Fig. 3A and Flory[24]). Then the intrinsic tendency of a particular residue will be independent of its sequence and secondary structure context. This interpretation is the one which has been assumed in an algorithm (AGADIR), which correctly calculates the helical content in solution of more than 320 monomeric peptides.[11]

The proper definition of α-helix is not clear and several criteria have been currently utilized. Definitions based in positions of the α-carbons[1] and in the hydrogen bond network[18] are the most popular ones. In both definitions, for a particular residue to be considered as helical, it is necessary that at least either the three preceding and/or following residues should be in a helical conformation. This necessarily introduces a bias in the statistical analysis since the presence of a particular residue in an α-helix is going to depend on the intrinsic helical propensities of the flanking residues as well as on the general way in which α-helices pack against proteins. The same will happen for the β-strands. If the amino acid's intrinsic secondary structure preferences are due to their particular properties and not to the way in which they interact with a particular secondary structure element, then it makes more sense to look at their distribution by dihedral angles, ignoring their secondary structure context. This is so for two reasons: First, we do not consider those elements of secondary structure which have distorted angles, due to packing reasons, and could have different amino acid propensities. Second, in this way we select not only α-helices or β-strands, but other secondary structure elements in which a particular residue could be in α-helical or β-strand dihedral angles (loops or β-turns), but having different secondary structure contexts. This might diminish the context effect, thus resulting in a more balanced scale. Another interesting feature of using the ϕ–ψ approach is that it provides absolute free energy values for each amino acid, instead of relative ones to a particular residue.

α-Helices

In the case of the α-helix scale, no such significant differences were found when the dihedral angles or the secondary structure approach was used, except in the case of Glu, which is clearly overestimated in the dihedral angle scale. This is probably due to the fact that typical angles of α-helix are rarely found in amino acids not forming helices, so the two scales are more similar than in the case of the β-strand. However, if we exclude Glu, the dihedral scale correlates in general better with the experimental ones. Specially interesting is the fact that the ϕ–ψ scale (without considering the Glu, residue) correlates really well with the normalized average scale as well as with a experimental scale derived from a modified version of AGADIR[28]). In the dihedral angles approach we are considering a very limited region of the Ramachandran plot, while in the secondary structure approach distorted helices are also selected, thus biasing the scale. Looking to the distribution of the different amino acids around the α-helical region, it is clear that when moving out of the two dihedral angle intervals considered (7,12 and 7,13), the preferences for the different amino acids change (data not shown). One dramatic case is that of Pro, which is found only 138 times in the interval 7,13, and 322 times in the interval 7,12, while Ala is found 1221 times in the first interval and 553 times in the second one. All of this results in

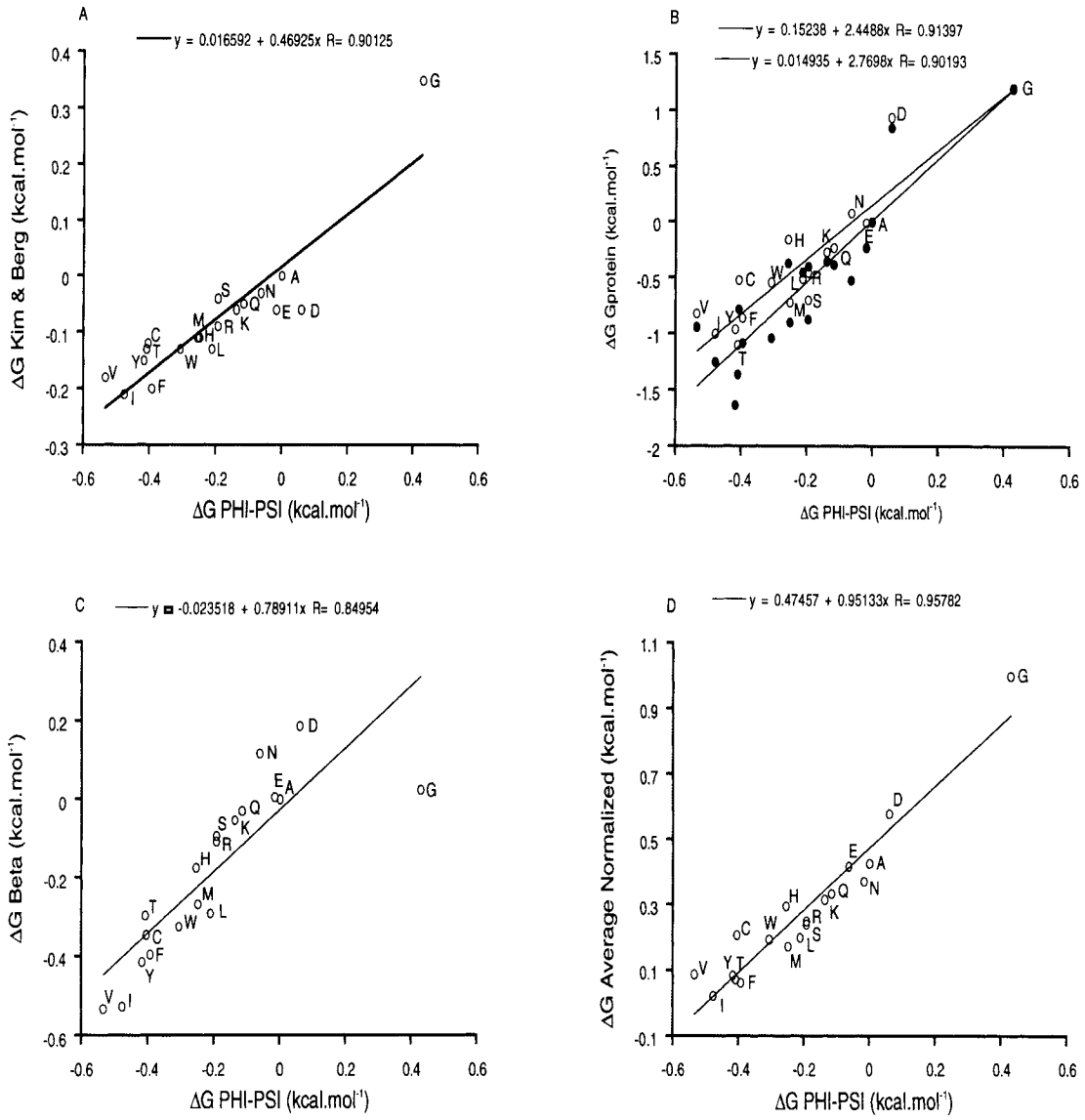

Fig. 3. Correlation analysis between the ϕ–ψ scale for β-strand propensities and the thermodynamic scales of (**A**) Kim and Berg,[12] (**B**) Minor and Kim[13] (open circles), and Smith et al.[14] (closed circles). (**C**) The β-strand scale derived from a search in the database using the Kabsch and Sander[18] definition. (**D**) The average scale obtained from the normalized scales of Kim and Berg[12] and Minor and Kim.[13]

a higher number of residues found when using the Kabsch and Sander definition, than when using the restricted phi-psi definition.

β-Strand

The correlation analysis of the different β-scales indicates that the ϕ–ψ one compares much better with the experimental ones than the one derived from the same database using the Kabsch and Sander[18] definition. In this case the number of hits for each amino acid is generally higher in the ϕ–ψ scale than in the other one, probably due to the relative abundance of residues with β-strand angles in loops and β-turns. If the β-strand intrinsic propensities are the result of the packing interactions within the β-sheets, then we should expect that the ϕ–ψ scale that also considers those residues adopting β-strand angles but belonging to loops or β-turns should compare worse with the experimental ones. The fact that we observe the opposite result validates the assumption that the intrinsic properties of the different amino acids are independent of their context. In the case of the β-strand region of the Ramachandran plot it is interesting to mention that the distribution by dihedral angles of the different amino acids is not uniform. In Figure 4 we show the

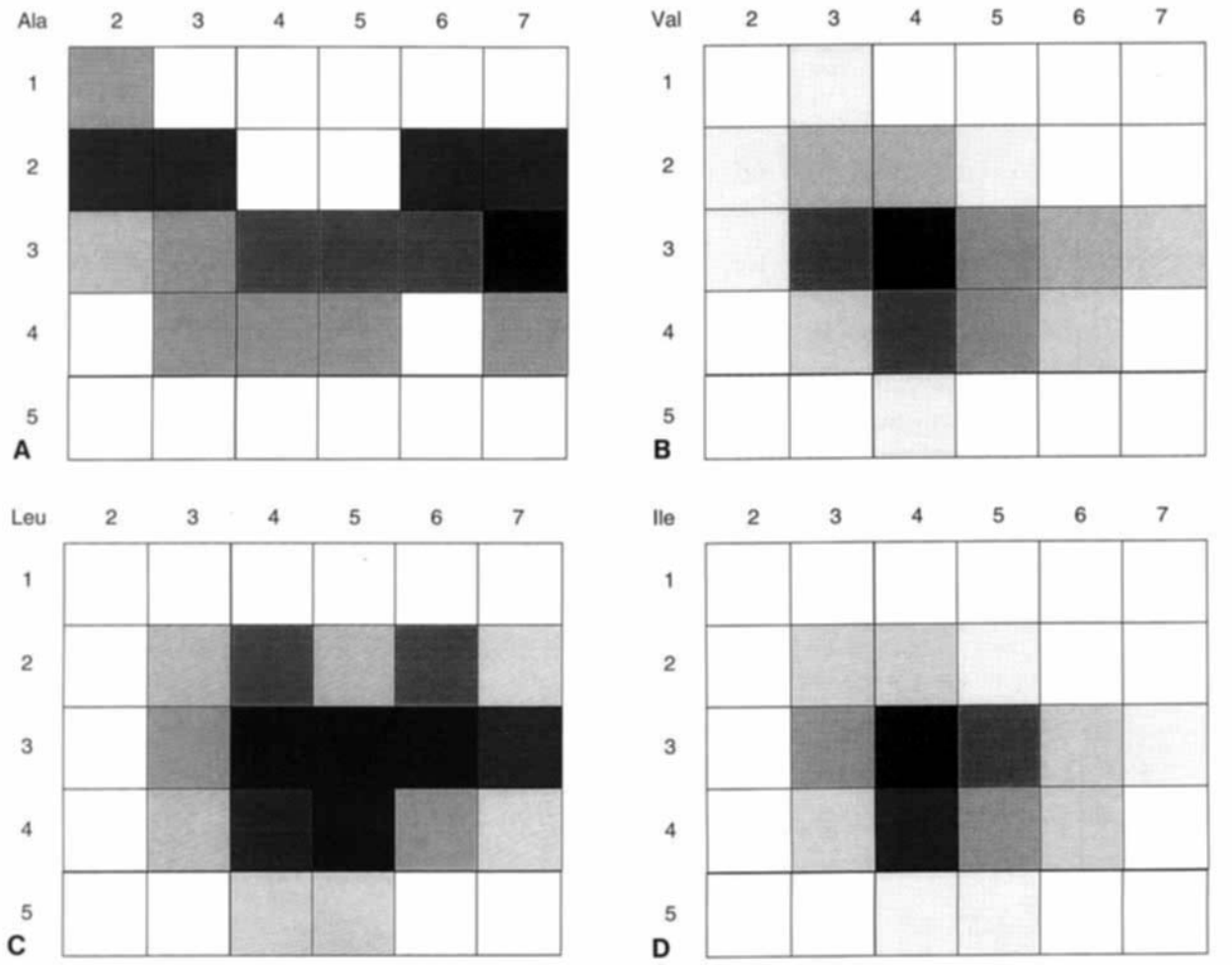

Fig. 4. β-Strand region of the Ramachandran plot (180 to 140 and −162 to −54), showing the relative propensities of four amino acids to populate different dihedral angles intervals in this region. The relative preference goes from black to white. (**A,** Ala) The intervals of relative propensity range from 0.03 to 0.00, in intervals of 0.005. (**B,** Val) The intervals of relative propensity range from 0.08 to 0.00, in intervals of 0.01. (**C,** Leu) The intervals of relative propensity range from 0.035 to 0.00, in intervals of 0.005. (**D,** Ile) The intervals of relative propensity range from 0.08 to 0.00, in intervals of 0.01.

typical distribution pattern of four hydrophobic residues, in relative terms. The distribution of the four residues is quite different. Interestingly enough as we mentioned above the distribution of Ala in the β-region of the Ramachandran plot is very similar to that calculated using conformational energies for polypeptides.[24] This irregular distribution again supports the hypothesis that the intrinsic propensities of the different amino acids are related to the free energy required to fix them in particular dihedral angles, independent of their context. In this respect it is worth mentioning that a more restricted region that includes the dihedral angles of the residue being mutated in the protein G (Tables I and II) is the one that correlates better with the data of Minor and Kim[13] or Smith et al.[14] (r~0.97; data not shown). This region does not include the dihedral angles of the position mutated by Kim and Berg[12] and it correlates worse with their data (r~0.8; data not shown). This restricted region includes as well the antiparallel β-sheet conformation of L-alanine.[24] The dihedral angles of the residue mutated by these authors are included by this region. When considering other areas of the same size overlapping the above region it results in worse correlations (data not shown). These results indicate that we should not speak of a universal scale for β-strand formation, but rather of different β-strand scales depending of the dihedral angles being considered.

CONCLUSIONS

From the results presented here, it is possible to conclude that calculation of the thermodynamic propensities to form secondary structure elements is feasible with a classical statistical approximation and a sufficiently large database of 3-D structures. It is also necessary to work out an appropriate definition of propensity to form a secondary structure

element of an amino acid, in order to eliminate biases from the packing of the proteins. Definition of the secondary structure propensity as the tendency to populate dihedral angles typical for the secondary structure seems to be a good way to circumvent it. Scales for these propensities, calculated from ϕ–ψ matrices derived from a database of 279 3-D structures, are therefore an alternative to the experimental model systems.

ACKNOWLEDGMENTS

We are specially grateful to Dr. G. Vriend for his efforts and time in programing the ϕ–ψ matrix option in WHATIF. We are very grateful to A. Ortiz for his helpful suggestions and discussions of the manuscript.

REFERENCES

1. Richardson, J.S., Richardson, D.C. Amino acid preferences for specific locations at the ends of α helices. Science 240: 1648–1652, 1988.
2. Chou, P., Fassman, G.D. Empirical predictions of protein conformation. Annu. Rev. Biochem. 47:251–276, 1978.
3. Rose, G., Seltzer, J. A new algorithm for finding the peptide chain turns in a globular protein. J. Mol. Biol. 113: 153–164, 1977.
4. Garnier, J., Osguthorpe, D.J., Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
5. Kyte, J., Doolittle, R. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105–132, 1982.
6. Sander, C., Scharf, M., Schneider, R. Design of protein structures. In: "Protein Engineering. A Practical Approach." Rees, A.R., Sternberg, M.J., Wetzel, R. (eds.). Oxford: Oxford University Press, 1992: 89–115.
7. O'Neil, K., DeGrado, W. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. Science 250:246–250, 1990.
8. Horovitz, A., Matthews, J.M., Fersht, A.R. α-Helix stability in proteins. II. Factors that influence stability at an internal position. J. Mol. Biol. 227:560–568, 1992.
9. Blaber, M., Zhang, X., Matthews, B.W. Structural basis of amino acid α-helix propensity. Science 260:1637–1640, 1993.
10. Chakrabartty, A., Kortemme, T., Baldwin, R.L. Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing sidechain interactions. Protein Sci., 3:843–852, 1994.
11. Muñoz, V., Serrano, L. Elucidating the folding problem of helical peptides using empirically derived parameters. Nature Struct. Biol., 1:399–409, 1994.
12. Kim, C.A., Berg, J.M., Thermodynamic β-sheet propensities measured using a zinc-finger host peptide. Nature (London) 362:267–270, 1993.
13. Minor, D.L., Kim, P.S. Measurement of the β-sheet-forming propensities of amino acids. Nature (London) 367:660–663, 1994.
14. Smith, C.K., Withka, J.M., Regan, L. A thermodynamic scale for the β-sheet forming tendencies of the amino acids. Biochemistry 33:5510–5517, 1994.
15. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
16. Vriend, G. WHAT IF: A molecular modeling and drug design program. J. Mol. Graphics 8:52–56, 1990.
17. Vriend, G., Sander, C., Stouten, P.F.W. A novel search method for protein sequence-structure relations using property profiles. Protein Eng. 7:23–29, 1994.
18. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
19. Dasgupta, S., Bell, J.A. Design of helix ends. Int J. Peptide Protein Res. 41:499–511, 1993.
20. Scholtz, J.M., Marqusee, S., Baldwin, R.L., York, E.J., Stewart, J.M., Santoro, M., Bolen, D.W. Calorimetric determination of the enthalpy change for the α-helix to coil transition of an alanine peptide in water. Proc. Natl. Acad. Sci. U.S.A. 88:2854–2858, 1991.
21. Kemp, D.S., Boyd, J.G., Muendel, C.C. The helical s constant for alanine in water derived from template-nucleated helices. Nature (London) 352:451–454, 1991.
22. Sander, C., Scharf, M., Schneider, R. Design of protein structures. In "Protein Engineering: A Practical Approach." Rees, A.R., Sternberg, M.J., Wetzel, R. (eds.). Oxford: Oxford University Press, 1992: 89–115.
23. Rooman, M.J., Kocher, J.P.A., Wodak, S.J. Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformations in the absence of tertiary interactions. Biochemistry 31:10226–10238, 1992.
24. Flory, P.J. "Statistical Mechanics of Chain Molecules." Oxford: Oxford University Press, 1988.
25. Bai, Y., Englander, S.W. Hydrogen bond strenght and β-sheet propensities: The role of a side chain blocking effect. Proteins: Struct. Funct. Genet. 18:262–266, 1994.
26. Zimm, B.H., Bragg, J.K. Theory of the phase transition between helix and random coil in polypeptide chains. J. Chem. Phys. 31:526–535, 1959.
27. Ramachandran, G.N., Sasisekharan, N.V. Conformation of polypeptides and proteins. Adv. Protein Chem. 24:1–95, 1968.
28. Muñoz, V., Serrano, L. Elucidating the folding problem of helical peptides using empirical parameters II. Helix macrodipole effects and rational modification of the helical content of natural peptides. J. Mol. Biol., in press.