THE TANGO ALGORITHM:

SECONDARY STRUCTURE PROPENSITIES, STATISTICAL MECHANICS APPROXIMATION AND CALIBRATION

Calculation of turn and beta intrinsic propensities.

A statistical analysis of a protein structure database has been carried out to calculate a mean force potential[1], for the various side chain-side chain energy contributions ($\Delta G_{(i,i+1)}$ and $\Delta G_{(i,i+2)}$). The main limitations lie in the size and quality of the database utilized as well as the definition of the parameters investigated. The database of 3-D structures consists of 279 proteins which always share less than 50% sequence homology and have been filtered to diminish biases produced by the existence of homologous proteins. This database was obtained following the principles described by Hobohm and co-workers[2] and is actually implemented in the protein design package WHATIF[3]. A complete list of the proteins included in this database has been published elsewhere[4]. Despite the large number of aminoacids (59117), the number of data is statistically too small in several cases. We have reduced the search space in the same way as previously published[5, 6], grouping the 20 natural aminoacids by their chemical nature in 15 different groups.

To determine the energy contribution of side chain-side chain interactions to β-strand stability with this method, it is important to minimize spurious contributions due to the packing of secondary structure elements in proteins. We have followed the same approach used to determine the intrinsic secondary structure propensities of the different amino acids[4]. Side chain-side chain interactions have been analyzed in stretches of two (i,i+1 interactions) or three (i,i+2 interactions) consecutive residues with β-strand phi and psi dihedral angles[7]. An open definition reduces the influence of protein packing by including aminoacids in loops and not only in buried β-strands.

The number of times a certain pair of residues at positions i,i+1, or i,i+2 is present in the database (observed frequency) is directly obtained with the SCAN3D option of the package WHATIF. The expected number of times, if there were no interactions between them, is obtained from the multiplication of the probabilities of finding both residues independently (eq 5).

$$p_{expected} = P_{aa1} * P_{aa2} \quad (5)$$

where $P_{expected}$ is the expected probability of finding a consecutive pair of residues in β-strand conformation, P_{aa1} is the probability of finding the amino acid 1 with β-strand dihedral angles in the database (equivalent to its β-strand intrinsic propensity: number of times found in β-strand angles / total number of times found), P_{aa2} is the same for the amino acid 2 .

This approach has been extensively used. The assumption behind it is that the database of proteins reflects a system in thermodynamic equilibrium (Boltzmann device) and therefore the observed frequencies for two residues at a certain position are related to the energy of the interaction between them. The relation between frequencies and free energies of interaction is indicated in equation 6.

$$\Delta G_{SC\text{-}SC} = -RT \ln(f_{observed}/p_{expected}) \quad (6)$$

where $\Delta G_{SC\text{-}SC}$ is the free energy of the interaction between two residues at positions i,i+1or i,i+2. $f_{observed}$ is the observed frequency of a pair (i,i+1) or triplet (i,i+2), of consecutive amino acids in β-strand dihedral angles. $P_{expected}$ is the expected probability for the same residues. To diminish statistical error, we have used the following procedure. A single case has been added ($f_{observed+1}$), or subtracted ($f_{observed-1}$), to the observed number of cases. Calculated interaction energies ($\Delta G_{SC\text{-}SC+1}$ and $\Delta G_{SC\text{-}SC\text{-}1}$), are used to determine the average energy ($\Delta G_{SC\text{-}SC.aver}$). The final value of the interaction ($\Delta G_{SC\text{-}SC.fin}$), is calculated following eq. 7

$$\Delta G_{SC\text{-}SC.fin} = \Delta G_{SC\text{-}SC.aver} + (\Delta G_{SC\text{-}SC.aver} - (\Delta G_{SC\text{-}SC+1})) \quad (7)$$

$\Delta G_{SC\text{-}SC.fin}$ values smaller than 0.04 kcal.mol^{-1} where considered null. The present

parameterization assigns equal interaction energy for all the amino acids within a chemically defined group. In β-turns since residues i and i+3 could adopt different conformations and are not fixed in the turn, we have applied a general empirical entropy penalty term of 0.3 Kcal/mol at 298K.

Estimation of desolvation costs in β-sheet aggregates.

We assume that the residues forming the core of the ordered aggregate must be fully buried. This implies full desolvation and minimum degrees of freedom. The energetic cost of burying a sequence stretch is defined by the following equation:

$$\Delta G = \Delta solv + \Delta vdw + \Delta Hbond + \Delta entropy + \Delta \varepsilon electrostatic \quad (1)$$

where Δsolv and Δvdw are obtained from the FOLD-EF forcefield [8] assuming maximum burial. ΔHbond is equal to the number of H-bonds made by the buried segment multiplied by the H-bond contribution (the same value used in AGADIR1s [9]). The number of H-bonds is equal to the number of donors, or acceptors, in the polypeptide chain that could pair with an acceptor or donor, respectively. For the backbone this is always 2 per residue, and for the side chains we count the total number of donors and acceptors and we take the minimum number of the two. In the case of proline residue we consider that if it is N-terminal to the segment we loss only one backbone H-bond, while if it is C-terminal we loss two. A proline residue inside a sheet breaks one H-bond and distorts a regular β-sheet. Thus it is incompatible with been in the center of a β-sheet and therefore it is highly penalized (10 Kcal/mol).

Δentropy assumes full entropy cost and is the sum of the main chain entropy due to the residues being in an extended conformation and full side chain entropy cost [10]. The model used to calculate the electrostatic contribution is the same that the one used to determine helix stability previously described in Lacroix *et al.* [11]. The electrostatic interactions obviously change with the degree of ionization and consequently with the pH of the solution, while the pKa of ionizable groups in a peptide change from their standard values depending on the electrostatic environment. In TANGO we considered all

electrostatic interactions (this includes charged side chain groups, free N-terminal and C-terminal main chain groups, and the succinyl blocking group if the peptide is succinylated), taking into account the ionic strength, temperature and pKa (see below) as described for AGADIR [11, 12].

TANGO distinguishes between charges in the segment under consideration (internal charges) which are considered fully buried, charges within two residues outside the N-or C-terminus of the segment (neighbouring charges) which are considered solvent exposed and the rest of the charges in the polypeptide chain (external charges). External charges are also considered to be solvent exposed but in addition their contribution is corrected with chain length. For buried charges we use a dielectric constant of (332/(8.8 * exp(-0.004314 * (temp-273.0)))), while for exposed charges it is (332/(88 * exp(-0.004314 * (temp-273.0)))).

The net charge for the segment under consideration plus its neighbouring residues is calculated assuming an average distance between charges in the aggregate of around 5Å in the aggregate. For the rest of the polypeptide chain TANGO calculate the net charge and divide it by the number of residues introducing a higher average distance for longer polypeptide chains

There are two types of electrostatic interactions: repulsive interactions due to a net charge and attractive interactions due to compensated charges. The latter has been introduced to reflect that on average some of the compensated charges in the aggregate nucleus will make salt bridges and thus contribute to the stability of the aggregate. In the case of the attractive compensated charges we correct the favorable electrostatic interaction calculated by dividing it by 3. This arbitrary correction factor is introduced since as explained above this term reflects the formation of internal salt bridges which of course cannot be formed by all compensated charges.

Statistical mechanics approximation

Ideally to calculate a partition function a multiple window approximation as used for the AGADIRms algorithm and described in Munoz *et al.*[9] should have been implemented. However, since we are taking into consideration 4 possible structural states, the calculation of the partition function would be computationally too demanding.

Therefore we have opted for a two-window approximation which assumes that the probability of finding more than two ordered segments in the same polypeptide chain is too low to be considered (the simple one window will deviate too much from reality for peptides with > 50 residues). Our assumption is that in the same polypeptide chain there could be one or two non-overlapping (separated by 5 unstructured residues or more, see Munoz & Serrano [9]) structured segments. This simplification could result in deviations for large proteins containing three or more strongly overlapping regions with a high tendency to adopt a particular state.

A second simplification is that we do not consider aggregation intermediates. We consider aggregates as a single molecular species or structural state in competition with β-turn and α-helical conformations again for the sake of simplifying the partition function. This simplification can be translated in the assumption that the aggregating segment has infinite concentration, or in other words, that once formed it immediately aggregates with infinite association constant. Since in reality the aggregation kinetics and the extent of aggregation will both depend on the concentration of the peptide and of its association constant, this means that the aggregation probabilities we are obtaining are only relative. Thus they allow quantitative comparison inside the same polypeptide chain, or with mutants of the polypeptide chain, but only qualitative comparison between different polypeptide chains.

In a third simplification, as in the multiple window approximation of AGADIR, we have assumed that there is no energetic coupling between the two non-overlapping segments (independent of their conformation) that are simultaneously present in the same molecule. This assumption seems rather reasonable for monomeric peptides in which there are no long or medium range interactions. Finally, we assume that all possible states can coexist by pairs in the same polypeptide molecule, i.e. is an aggregate can have a helical segment as long as it is out of the aggregated segment (for example: in lysozyme amyloids, helical regions still persist [13]).

Under these assumptions and the definition of the random coil state as those conformations which are not helical, turn or involved in aggregation, the two-window sequence partition function becomes the sum of the statistical weights for all the possible combinations of structured segments (from one to two non-overlapping segments) plus

the statistical weight for the random coil state (the set of molecular conformations which do not include any structured segment). The weight for the random coil is 1 (arises from the product of the weights of all the residues in the random coil state). As a result of the third assumption, the statistical weight of molecular conformations with more than one structured segment are simply the product of the weights of all the structured segments included on it (see Munoz *et al.*[9]).

Calibration of the parameters

Since we are using different states and energy contributions it is important not to overfit the data. Out of all the parameters included in TANGO we have only fitted the electrostatic contribution in the β-sheet aggregate and the intrinsic propensity of Pro inside an aggregate. For the β-turn state the parameters are those obtained from statistical analysis of the database and the H-bond term is that of AGADIR [5, 6, 9]. In the case of the α-helical state AGADIR is used without any modification. The energy of the folded state is introduced also without any weight. Finally regarding the β-sheet aggregates the H-bond term is the same one used in AGADIR and the desolvation and entropy contributions are those used in FOLD-X [8]. The only parameter that has been empirically adjusted is the electrostatic contribution to the β-sheet aggregate. We have used a dataset of 179 peptides corresponding to sequence fragments of 21 different proteins as our calibrating set as described in Supplementary Table 1.

References

1. Sippl, M.J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol* **213**, 859-883. (1990).
2. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. Selection of Representative Protein Data Sets. *Protein Science* **1**, 409-417 (1992).
3. Vriend, G., Sander, C. & Stouten, P.F.W. A Novel Search Method for Protein-Sequence Structure Relations Using Property Profiles. *Protein Engineering* **7**, 23-29 (1994).
4. Munoz, V. & Serrano, L. Intrinsic Secondary Structure Propensities of the Amino-Acids, Using Statistical Phi-Psi Matrices - Comparison with Experimental Scales. *Proteins-Structure Function and Genetics* **20**, 301-311 (1994).
5. Munoz, V. & Serrano, L. Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. *J Mol Biol* **245**, 297-308. (1995).
6. Munoz, 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* **245**, 275-296. (1995).
7. 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 Conformation in the Absence of Tertiary Interactions. *Biochem* **31**, 10226-10238 (1992).
8. Guerois, R., Nielsen, J.E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* **320**, 369-387 (2002).
9. Munoz, V. & Serrano, L. Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson- Roig formalisms. *Biopolymers* **41**, 495-509 (1997).
10. Abagyan, R. & Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* **235**, 983-1002 (1994).
11. Lacroix, E., Viguera, A.R. & Serrano, L. Elucidating the folding problem of alpha-helices: Local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. *Journal of Molecular Biology* **284**, 173-191 (1998).
12. Munoz, V., Blanco, F.J. & Serrano, L. The distribution of alpha-helix propensity along the polypeptide chain is not conserved in proteins from the same family. *Prot Sci* **4**, 1577-1586 (1995).
13. Booth, D.R. et al. Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* **385**, 787-793 (1997).