TANGO

1-Description and basis of the algorithm

The model used by the TANGO algorithm is designed to predict cross-beta aggregation in peptides and denatured proteins and consists of a phase-space encompassing the random coil and 4 possible structural states: β-turn, α-helix, β-sheet aggregation and α-helical aggregation. Every segment of a peptide can populate each of these states according to a Boltzmann distribution, i.e. the frequency of population of each structural state for a given segment will be relative to its energy. Therefore, to predict cross-beta aggregating segments of a peptide TANGO simply calculates the partition function of the phase-space. Here we first describe how we determine the propensity for each of the different structural states, how we sample phase-space and which assumptions are embedded in these choices.

α-Helical propensities

The parameters used in the latest version of AGADIR (AGADIR-1s11), have been used to determine the helical propensity of the amino acid sequences. The only modification has been the implementation of a multiple partition function (see below).

β-Turn propensities

beta-turn propensity is calculated by considering four energy contributions: (1) an amino-acid specific cost in conformational entropy for fixing that residue in a beta-turn compatible conformation, (2) interactions of each amino acid with the turn structure in a position dependent manner, (3) in some cases side chain-side chain, or side chain-main chain interactions within the turn and (4) a single H-bond between the main chains of residues i and i+3 of the turn. We have only considered 4 types of turns for which we could obtain significant statistical data, Types I, I, II and II. The entropic cost of fixing a particular amino-acid in turn dihedral angles, has been obtained using statistical f,y matrices, as previously published. Since residues i and i+3 could adopt different conformations and are not fixed in the turn, we have applied a general entropy penalty term of 0.3 Kcal/mol at 298K. The interaction of the amino acids with the turn has been obtained by statistical analysis of the protein database (see methods section), assuming that counts for observed interactions higher than the expected value represent favorable interactions and the opposite is true.

Cross β-aggregation

To estimate the aggregation tendency of a particular amino acid sequence, we have taken the following assumptions: (1) In an ordered beta-sheet aggregate the main secondary structure is beta-strand. (2) The regions involved in the aggregation process are fully buried, thus paying full solvation costs and gains, full entropy and optimize their H-bond potential (that is the number of H-bonds made in the aggregate is related to the number of donor groups that are compensated by acceptors). An excess of donors or acceptors remains unsatisfied. (3) Complementary charges

in the selected window establish favorable electrostatic interactions and overall net charge of the peptide and net charges near the aggregating region (two residues before or after the choosen window), disfavor aggregation.

Estimation of β-propensity.

We have included three energy contributions: a residue-specific cost in conformational entropy for fixing that residue in a beta-strand conformation and side chain-side chain interactions of residue i with residues at positions i+1 and i+2. Formation of a beta-strand requires, in general, less conformational entropy cost than formation of an alpha-helix of equivalent length, because the beta-strand region of the Ramachandran plot is larger than the alpha-helical region while the depth of the energy well is similar. On the other hand, a single beta-strand does not have main chain-main chain hydrogen bonds that counteract the loss in conformational entropy. In the absence of other contributions the beta-strand will not be populated over the random coil. However, a factor not generally considered is the existence of intra-strand side chain-side chain interactions that when favorable could promote beta-strand population. The unique side chains that are close in space in an extended conformation (beta-strand) are those between positions i and i+2. Residues at positions i and i+1 could also influence the formation of the beta-strand since they are on average more distant than in the random-coil. This phenomenon has energetic implications that we denominate (i,i+1) beta-interactions. On this basis, favorable (i,i+1) b-interactions reflect repulsions between residues i and i+1 while (i,i+1) beta-unfavorable interactions reflect attractions of the these side chains when they are not in a beta-strand conformation. These side chain-side chain interactions introduce energetic coupling in the b-strand-coil transition, producing some cooperativity. The entropic cost of fixing a particular aminoacid in beta dihedral angles, has been obtained using statistical f,y matrices, as previously published.

The other two terms participating in the equation, interaction between residues i,i+1 and i+2, are relative to the energy contribution of side chain-side chain interactions. They have been determined using a mean-force potential.

Desolvation costs of aggregated segments.

As explained above 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:

$$\Delta G = \Delta G_{solv} + \Delta G_{VdW} + \Delta G_{Hbond} + \Delta G_{entropy} + \Delta G_{electrostatics}$$

were Dsolv and Dvdw are obtained from the FOLD-EF forcefield (Reference) assuming maximum burial. DHbond 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). 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 just count the total number of donors and acceptors and we take the minimum number of the two. In the case of Pro 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 Pro inside a segment is penalized by 10 Kcal/mol.
Dentropy assumes full entropy cost and is the sum of the main chain entropy due to the residues being in an extended conformation and side chain entropy (as described by ABGYAN). The model used to calculate the electrostatic contribution to helix stability was previously described in Viguera, Lacroix, Serrano). In the following paragraph we describe how electrostatic contributions Delectrostatic to beta-aggregates are computed.

Electrostatic contribution.

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 involves charged side chain groups, free N-terminal and C-terminal main chain groups, and the succinyl blocking group if the peptide is succinylated) to compute the electrostatic environment of the amino acids in the random coil and in helical segments, taking into account the ionic strength, temperature and the pKa (see below).

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 5A. For the rest of the polypeptide chain TANGO calculates 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 one has been introduced to reflect that on average some of the compensated charges 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.

α-Helix aggregation.

Some peptides and proteins aggregate in a helical conformation. This is typically observed in proteins with a tendency to form coiled-coil structures or Leu-zippers (references). Since formation of dimers or higher order helical aggregates will compete with beta-sheet aggregation we have included this structural state in the TANGO algorithm in a very simple manner. As for

beta-sheet aggregation we assume full burial upon aggregation, but only for one face of the helical structure. Thus, we assume than in a helical aggregate residues i,i+1, i+4,i+5, i+8, i+9 etc will be fully buried. For those residues we applied the same considerations as for burial of residues in beta-sheet aggregates. The energy required to fold the segment into a helical conformation, however, is directly derived from AGADIR.

pH, ionic and temperature dependence

The effect of pH, temperature and ionic strength on electrostatic interactions was taken into account as described in AGADIR2-1s11. Similarly the dependence of entropy, H-bonds and hydrophobic interactions on temperature and ionic strength are taken into consideration as described in AGADIR2-1s11.

Multiple partitions Function.

To calculate a partition function a multiple window approximation as used for the AGADIRms algorithm and described in Munoz et al. has been implemented. Basically we consider that overlapping windows and windows up to two residues from the beginning or the end of the window been analyzed can compete. Windows separated by more than two residues from the one been considered are not included in the partition function.

A simplification is that we do not consider aggregation intermediates. We consider aggregates as a single molecular species or structural state in competition with b-turn and a-helical conformations again for the sake of simplifying the partition function. This simplification can be translated in the assumption that the aggregating segment has an 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 depend on the concentration of the peptide as well as of its association constant, this means that the aggregation probabilities we are obtaining are only relative. Thus they allow comparison inside the same polypeptide chain, or with mutants of the polypeptide chain, but not between different polypeptide chains.

Third, like 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, that is an aggregate can have a helical segment as long as it is out of the aggregated segment (there is experimental evidence for this, like in lysozyme were helical regions still persist in the amyloid aggregate).

Under these assumptions and the definition of the random coil state as those conformations which are not helical, or turn or involved in aggregation, the multiple sequence partition function of one window becomes the sum of the statistical weights for all the possible combinations of structured 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).

2-Input and output formats

Running TANGO from the command line.

1. Open a command line window and go to the directory where you have put the executable.
2. Call tango as in the following example (use spaces, not tabs):

    Tango P05100 ct="N" nt="N" ph="7.4" te="303" io="0.05" seq="DNEWGYIAYHVSQDP"

    ct: Protection at the C-terminus: can be N for no or Y for amidated

    nt: Protection at the N-terminus: can be N for no, A for acetylated or S for succinilated.

    ph: pH

    te: Temperature in Kelvin

    io: Ionic strength

Running TANGO with an input file.

To be run with an input file, Tango needs a text file that can have any name the user wants as long as it has less than 25 characters. Inside the file the user can place as many sequences to be analyzed as long as the number is less than 1000 sequences.

The format of the sequences to be run is as follows:

Name Cter Nter pH Temp Ionic Sequence

Name = name of the sequence (less than 25 characters)

Cter = status of the C-terminus of the peptide (amidated Y, free N)

Nter = status of the N-terminus of the peptide (acetylated A, succinilated S and free N)

pH = pH

Temp = temperature in Kelvin

Ionic = ionic strength in M

sequence = sequence of the peptide in one letter code.

Example

Sup1 N N 7 298 0.1 AMAPVLYLQDKSS

sup2 N N 7 298 0.1 AMASVLYLQDKSS

sup3 N N 7 298 0.1 AMAPVLYLQSKSS

sup4 N N 7 298 0.1 AMASVLYLQSKSS

sup5 N N 7 298 0.1 AMAPVLYLQPKSS

sup6 N N 7 298 0.1 AMARVLYLQDKSS

sup7 N N 7 298 0.1 AMAPVLYLQRKSS

The programme window first asks if the user wants to have the aggregation content by residue. If the user types Y, then he/she will get a file for each sequence in the text file.

TANGO output.

The output of TANGO is in text format with the extension .out. You will get two classes of outputs. One with the name of the file you run that will contain the average aggregation per residue for every sequence you had in your file. The other will be a series of files with the names of the sequences you run that will contain the prediction at the residue level.

Those files will have the following columns:

Sequence Number
Amino acid in one-letter code
Percentage of β-strand conformation
Percentage of β -turn conformation
Percentage of α-helical conformation
Percentage of Aggregation
Percentage of Helical Aggregation.

Please be aware that the latest is calculated independently of the first four and therefore you could get a number higher than 1 if you sum the 5 columns.

Example of sequence output:

```
01, M, 0.14, 0.08, 0.00, 0.00, 0.00
02, R, 0.24, 0.11, 0.00, 0.00, 0.00
03, S, 0.44, 0.11, 0.00, 0.00, 0.00
04, L, 0.45, 0.27, 0.00, 0.87, 0.00
05, E, 0.36, 0.19, 0.00, 0.87, 0.00
06, T, 1.14, 0.16, 0.00, 1.40, 0.00
07, F, 1.08, 0.16, 0.00, 1.61, 0.00
08, V, 1.08, 0.62, 0.00, 1.61, 0.00
09, G, 1.03, 0.66, 0.00, 0.74, 0.00
10, D, 0.18, 0.67, 0.00, 0.74, 0.00
11, Q, 0.14, 0.67, 0.00, 0.74, 0.00
12, V, 0.39, 0.05, 0.00, 3.80, 0.00
13, L, 0.64, 0.00, 0.00, 3.80, 0.00
14, E, 0.73, 0.00, 0.00, 3.31, 0.00
15, I, 0.75, 0.00, 0.00, 3.31, 0.00
16, V, 0.48, 0.00, 0.00, 3.31, 0.00
17, P, 0.23, 0.22, 0.00, 3.04, 0.00
18, S, 0.12, 0.24, 0.00, 0.62, 0.00
19, N, 0.00, 0.35, 0.00, 0.00, 0.00
20, E, 0.00, 0.38, 0.00, 0.00, 0.00
21, E, 0.00, 0.17, 0.00, 0.00, 0.00
22, Q, 0.33, 0.15, 0.00, 0.00, 0.00
23, I, 0.39, 0.14, 0.00, 0.00, 0.00
24, K, 0.40, 0.12, 0.00, 0.00, 0.00
25, N, 0.40, 0.12, 0.00, 0.00, 0.00
26, L, 0.11, 0.12, 0.00, 0.00, 0.00
27, L, 0.34, 0.03, 0.00, 0.00, 0.00
28, Q, 0.59, 0.02, 0.00, 0.00, 0.00
```

29, L, 0.63, 0.04, 0.00, 0.00, 0.00
30, E, 0.62, 0.04, 0.00, 0.00, 0.00
31, A, 0.33, 0.06, 0.00, 0.00, 0.00
32, Q, 0.09, 0.12, 0.00, 0.00, 0.00
33, E, 0.06, 0.10, 0.00, 0.00, 0.00
34, H, 0.16, 0.10, 0.00, 0.00, 0.00
35, L, 0.44, 0.09, 0.00, 0.00, 0.00
36, Q, 0.56, 0.04, 0.00, 0.00, 0.00
37, L, 0.78, 0.03, 0.00, 0.00, 0.00
38, D, 0.82, 0.03, 0.00, 0.00, 0.00
39, F, 0.70, 0.00, 0.00, 0.00, 0.00
40, W, 0.76, 0.00, 0.00, 0.00, 0.00
41, K, 0.54, 0.00, 0.00, 0.00, 0.00
42, S, 0.38, 0.25, 0.00, 0.00, 0.00
43, P, 0.20, 0.25, 0.00, 0.00, 0.00
44, T, 0.05, 0.25, 0.00, 0.00, 0.00
45, T, 0.04, 1.47, 0.00, 0.00, 0.00
46, P, 0.04, 1.36, 0.00, 0.00, 0.00
47, G, 0.04, 1.48, 0.00, 0.00, 0.00
48, E, 0.01, 1.48, 0.00, 0.00, 0.00
49, T, 0.14, 0.27, 0.00, 0.00, 0.00
50, A, 0.20, 0.12, 0.00, 0.00, 0.00
51, H, 0.49, 0.00, 0.00, 0.00, 0.00
52, V, 2.69, 0.00, 0.00, 0.00, 0.00
53, R, 2.57, 0.00, 0.00, 0.00, 0.00
54, V, 2.50, 0.00, 0.00, 6.27, 0.00
55, P, 2.21, 0.00, 0.00, 6.27, 0.00
56, F, 0.00, 0.00, 0.00, 11.61, 0.00
57, V, 0.00, 0.00, 0.00, 11.93, 0.00
58, N, 0.00, 0.00, 0.00, 11.93, 0.00
59, V, 0.00, 0.01, 0.33, 11.69, 0.00
60, Q, 0.00, 0.01, 0.33, 8.01, 0.00
61, A, 0.00, 0.01, 0.33, 7.08, 0.00
62, V, 0.00, 0.01, 0.33, 6.80, 0.00
63, K, 0.00, 0.00, 0.00, 0.00, 0.00
64, V, 0.00, 0.00, 0.00, 4.17, 0.00
65, F, 0.00, 0.00, 0.00, 4.39, 0.00
66, L, 0.00, 0.09, 0.00, 4.39, 0.00
67, E, 0.00, 0.19, 0.00, 4.39, 0.00
68, S, 0.00, 0.38, 0.00, 4.53, 0.00
69, Q, 0.00, 0.40, 0.00, 5.02, 0.00
70, G, 0.00, 0.31, 0.00, 12.27, 0.00
71, I, 0.00, 0.21, 0.00, 64.25, 0.00
72, A, 0.00, 0.02, 0.00, 68.13, 0.00
73, Y, 0.00, 0.01, 0.00, 70.91, 0.00
74, S, 0.00, 0.01, 0.00, 70.91, 0.00
75, I, 0.00, 0.01, 0.00, 70.91, 0.00
76, M, 0.00, 0.01, 0.00, 67.35, 0.00
77, I, 0.00, 0.12, 0.00, 62.97, 0.00

```
78, E, 0.00, 0.12, 0.00, 4.98, 0.00
79, D, 0.00, 0.12, 0.00, 1.93, 0.00
80, V, 0.00, 0.12, 0.00, 1.85, 0.00
81, Q, 0.00, 0.00, 0.00, 0.00, 0.00
```

3-Interpretation of the data

The user must be aware that TANGO considers that the polypeptide sequence is fully denatured and solvent exposed. Thus a globular protein with high stability could have a strong aggregating sequence inside and will aggregate little if it folds fast and under diluted conditions, while the same sequence in a small unfolded peptide will readily aggregate.

Strong aggregation regions in globular proteins could be problematic if the protein is exposed to denaturing conditions, suffers a point mutation that destabilizes it or is at sufficient high concentration that the small percentage of the denatured form can start aggregation.

The user must be aware that aggregation is a concentration dependent process and therefore sequences that at 0.1 mM will be soluble they will precipitate at 1 mM. TANGO assumes a fixed concentration of 1 mM (We have other versions with concentration and stability dependence which are available).

As a rule of the thumb any segment with an aggregation tendency above 5% over 5-6 residues is a potential aggregating segment.