Universal Stability Model for Globular Proteins

Boris Fačkovec

Institute of Organic Chemistry and Biochemistry
AS CR, vvi.

March 23rd 2012

Introduction: Demand for Stability Models

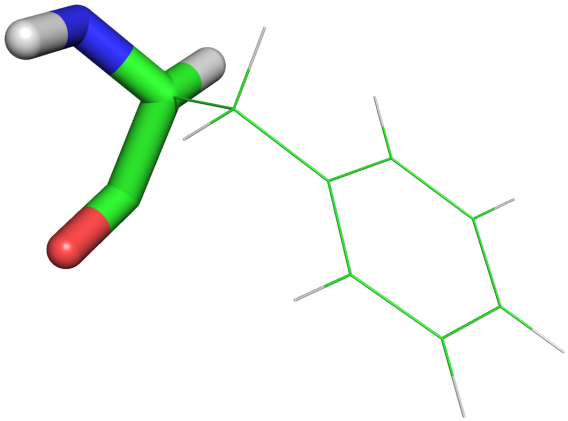
- definitions
  - **stability** - free energy difference between folded and unfolded state
  - **energy function** - mapping from protein geometry and environment information to stability
  - **protein stability model** - physical model of protein molecule which facilitates stability prediction
- energy functions for protein structure
  - structure prediction
  - protein engineering - mutagenesis
- molecular modeling of proteins
  - problem with electrostatic interactions
  - denatured state representation

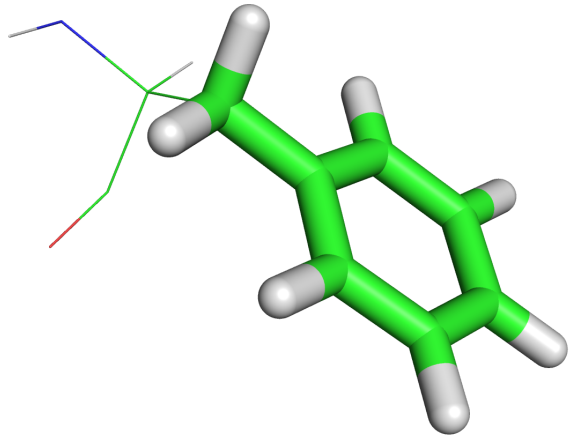
Introduction: Current Stability Models

- properties of current protein stability models
  - based of solvent accessible area calculations - 1-body, composition-dependent
  - almost exclusively rely on additivity of free energy contributions
- debates on the driving force of protein folding and **contribution** of particular redisue-residue interactions (hydrogen bonds, salt bridges etc.) to protein stability
- interaction energy calculations $\rightarrow$ performance of current force fields
- stability predictors
  - statistical force field - FoldX
  - physical force field Medusa $\rightarrow$ ERIS

Methods: Interaction energy approach

- proteins (N amino acids) split into 2 N fragments





- 4 types of fragments
  - BB - backbone disregarding amino acid type
  - CH - charges sidechains (D,E,K,R,H)
  - PO - polar sidechains (Y,W,N,Q,T,S)
  - NO - non-polar sidechains (A,L,I,V,C,M,P,F)

Methods: Interaction Energy Matrix Approach

- using additive force field they contain all the information about energy of native structure in sequential context

NPNP IEM	ALA 1	ALA 2	GLN 3	SER 4	VAL 5
ALA 1	0.00	0.01	0.00	0.00	0.00
ALA 2	0.01	0.00	0.00	0.00	-0.12
GLN 3	0.00	0.00	0.00	0.00	0.00
SER 4	0.00	0.00	0.00	0.00	0.00
VAL 5	0.00	-0.12	0.00	0.00	0.00
ASP 6	0.00	0.00	0.00	0.00	0.00
GLN 7	0.00	0.00	0.00	0.00	0.00
LEU 8	0.00	0.00	0.00	0.00	-0.54
ILE 9	0.00	-0.22	0.00	0.00	-0.32
LYS 10	0.00	0.00	0.00	0.00	0.00

Figure 1: Example of interaction energy matrix for non-polar side-chain fragments

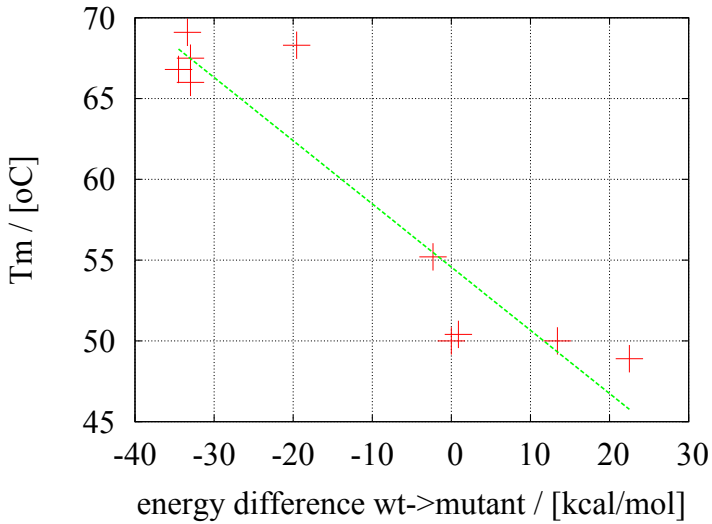


Figure 2: Naive model for stability change upon amino acid substitution

- improvement can be made introducing scaling factors

Methods: Our Model of Protein Stability

- Thermodynamic cycle $\Delta G = \Delta G_1 + \Delta G_2 + \Delta G_3$

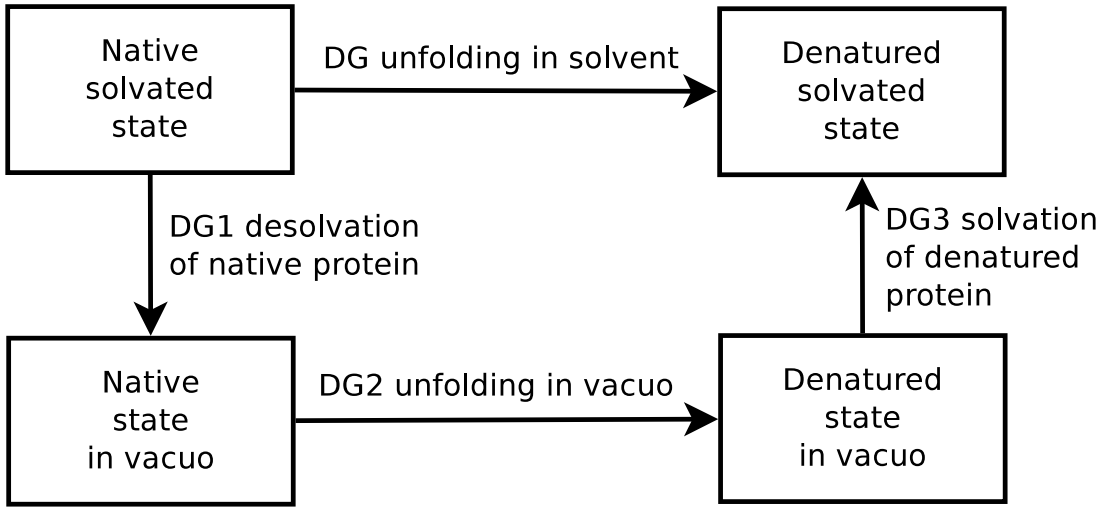


Figure 3: Unfolding free energy as a sum of the free energies for 3 processes

- ΔG_1: polar and non-polar SAS (2 parameters)
- ΔG_2: interresidual 2-body interactions (10 params), torsional restraints (1 param), configurational entropy (20 params)
- ΔG_3: solvation of individual aminoacids (20 params)

Methods: Scaling Factors - the Core of the Model

- functional form for a structure

$$\Delta G = \sum_{i=1}^{20} c_i n(AA_i) + \sum_{i=21}^{30} c_i IE_j + c_{31} SAS(np) + c_{32} SAS_{po} + c_{33} E_{tor}$$

- contains 33 parameters
- not enough experimental data!

$$\Delta G \approx 0 kcal/mol$$

- we can use our set of 1287 calculated IEMs (structures: X-ray, resolution 2 Å or better, single chain, no ligands, 70% sequence identity removed)

Methods: Optimization Procedure

- genetic algorithm in Octave (population of 1000 vectors, 500 generations)
- fitness function - RMSD

$$F = \left(\sum_{i=1}^{M} f_i^2 \right)^{\frac{1}{2}}$$

of Δ G compensation

$$f = \frac{\sum_{i=1}^{20} c_i n(AA_i) + \sum_{i=21}^{30} c_i IE_j + c_{31} SAS(np) + c_{32} SAS_{po} + c_{33} E_{tor}}{\sum_{i=1}^{20} c_i n(AA_i)}$$

- searched space - boundaries of SFs
  - 0 .. 1 - for 2-body interactions and torsion, without loss of generality
  - 0 .. RIE - for 1-body interactions
  - -50 .. 50 - for SAS SFs

Methods: Optimization Algorithm

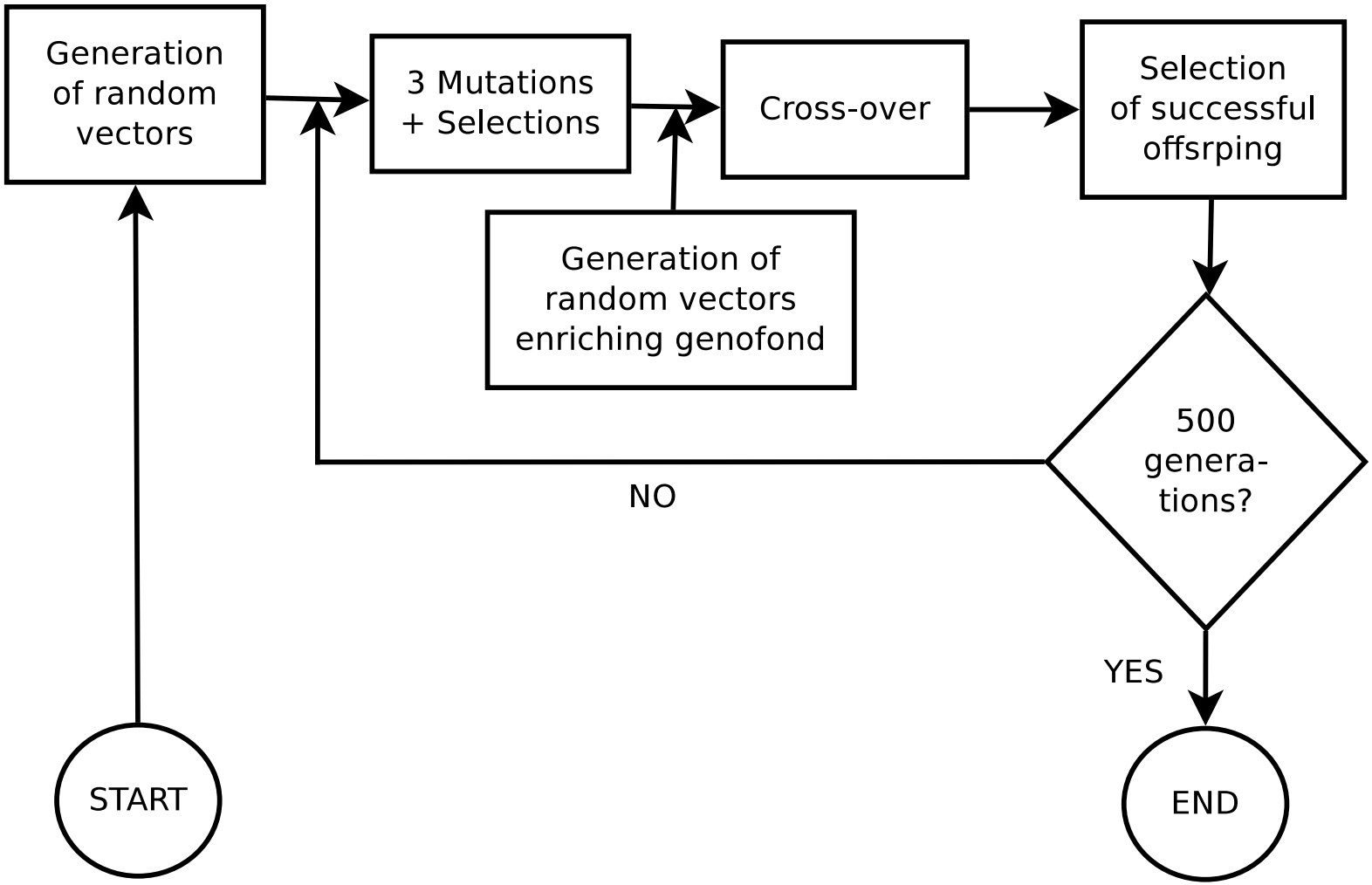


Figure 4: Evolution diagram

Results: Contribution of One-Body Interactions to Overall Stability

amino acid	scaling factor lowest	scaling factor highest			
			ARG	25.7	31.3
CYS	7.3	9.5	ASP	25.6	31.0
ILE	7.2	9.6	HIS	25.3	30.5
PRO	7.1	9.4	ASP	25.4	30.8
PHE	7.5	9.4			
ALA	7.3	9.4	SER	14.8	18.6
GLY	6.9	9.3	GLN	15.3	19.3
LEU	7.3	9.5	TRP	14.9	19.3
VAL	7.4	9.4	ASN	15.1	19.8
MET	7.3	9.6	TYR	14.8	20.2
			THR	15.0	18.5

Results: Contribution of Two-Body Interactions to Overall Stability

amino acid	scaling factor lowest	scaling factor highest
BB	0.578	0.71
BBCH	0.46	0.583
BBPO	0.669	0.84
BBNP	0.284	0.356
CHCH	0.115	0.149
CHPO	0.451	0.578
CHNP	0.401	0.561
POPO	0.476	0.631
PONP	0.471	0.603
NPNP	0.387	0.555
HphobSA	-4.388	-0.79
HphilSA	-28.371	-23.991
torsion	0.391	0.49

Results: Applicability and Reliability of the Model

- fitness function (RMSD of compensation) about 3%
  - folding free energy in order of tens of kcal/mol - probably enough to reach experimental values
  - better than expected - decomposition is good enough to include other factors
- challenges remaining
  - additivity of solvation energies in denatured state
  - additivity of intramolecular interactions
  - key positions in native structure
  - solvation free energy of native state as a linear function of polar and nonpolar SAS
  - vibrational entropy not included

Results: Applicability and Reliability of the Model

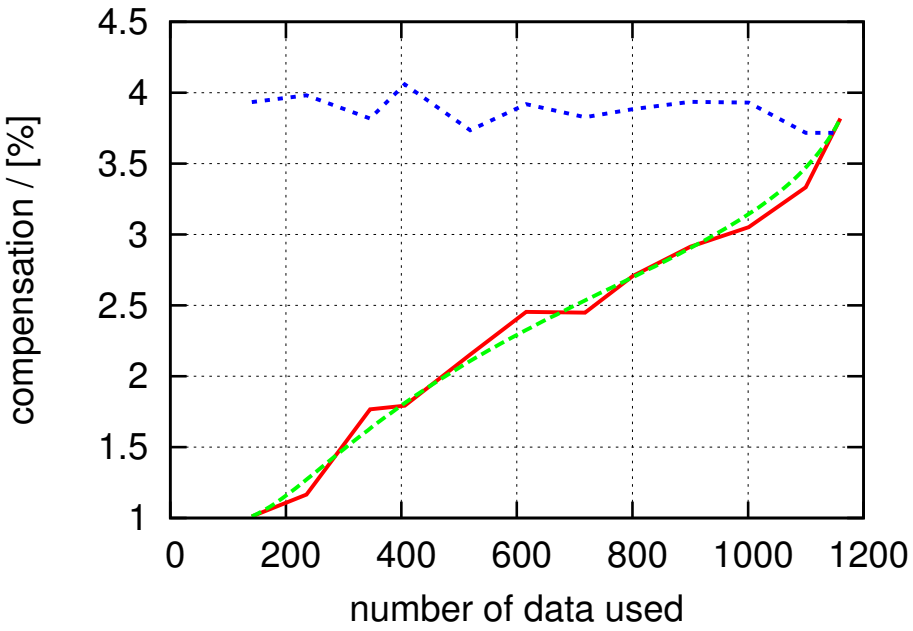


Figure 5: Fitness as a function of number of protein structures in data sample. 33 parameters can be reliably determined using just 300 proteins.

- hypothesis - native state can be reliably represented by IEM of minimum energy structure
- decomposition of protein stability into one-body and two-body contributions

Conclusion

- we have developed a new transferable and robust model of protein stability
- it can help us to
  - understand data in IEMs (proper treatment of electrostatic interactions)
  - understanding thermodynamics of folding studying contributions
  - develop more accurate energy functions
- 16 parameters are sufficient
- model is robust
- accurate enough to represent ex

Future Directions

- web application
- stability change upon mutation database
- improvement of the model - polar surface definition

Acknowledgement

- Jiří Vondrášek
- Jiří Vymětal
- Jiří Kysilka

Figure 6: IOCB Center for Complex Molecular Systems groups.

GAČR P208/10/0725 MŠMT LH11020 RP Z40550506