Thermal adaptation of proteins: bioinformatics and physical model

What do bacteria need to survive at 100°C ?

Konstantin Zeldovich, Igor Berezovsky,
Eugene Shakhnovich

Department of Chemistry and Chemical Biology,
Harvard University

kzeldov@fas.harvard.edu

PLOS Comp Biol 3(1) e5 (2007)
PLOS Comp Biol 3(3) e52 (2007)

Temperature ranges of modern life

Psychro-, meso-, thermo-, hyperthermophilic bacteria/archaea

-10°C (Antarctic ice, permafrost in Siberia and Canada)

Colwellia spp, *Psychrobacter spp*

+110°C (deep sea hydrothermal vents, hot springs)

Pyrococcus spp, *Methanococcus spp*

>300 sequenced genomes

Simplest eukaryotes: up to ~60°C (nematode from hot springs)`

Cold-blooded animals:

Notothenia spp. Antarctic fish: -1.8°C habitat,

dies of overheating at +6°C = 40°F

Desert iguana: up to +60°C

Very few complete genomes!

Is habitat temperature reflected in the genomes?

Existing knowledge

- What is presumably related to thermostability?
  - G+C in DNA increases with temperature (wrong)
    - DNA stabilization by pairing
  - Fraction of charges (DEKR) in proteins increases
    - Hydrophobic interactions weaken with temperature
  - Fraction of polar residues decreases
    - ?

Limitations of the previous work:
based on a few (dozen) individual proteins, or
a limited number (~20) of completely sequenced genomes

Here: high-thoroughput analysis, 204 genomes

Our approach: complete enumeration

Model: habitat temperature is correlated with a linear combination of the fractions of (certain) amino acids in the proteome

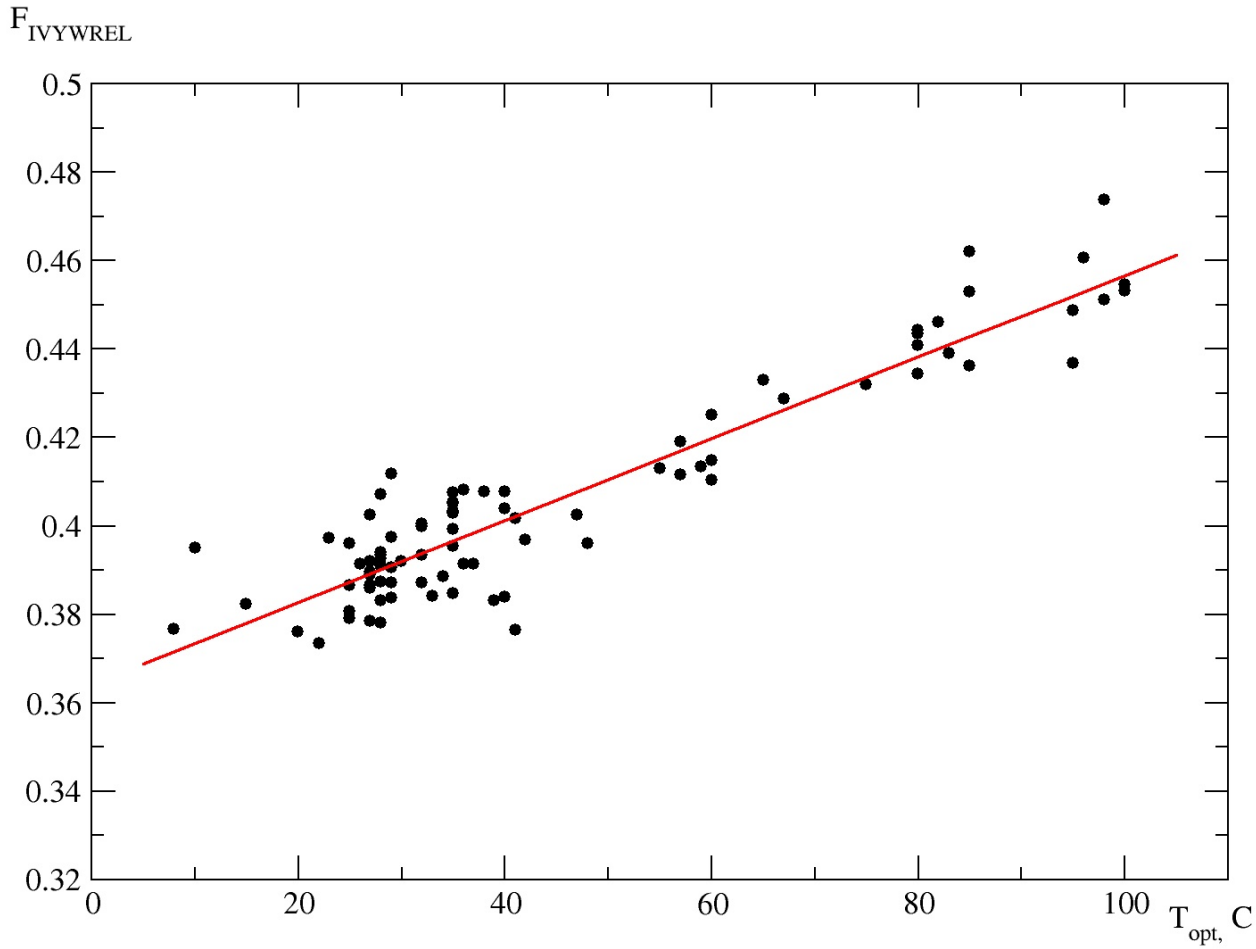
2^{20} combinations of 20 amino acids, 2^{19}-1 linearly independent combinations of fractions of a.a. f_i

$$F^{(j)} = \sum_{i=1}^{20} a_i f_i^{(j)}$$, a_i=0 or 1, all sets of $\{a_i\}$

Perform linear regression between $T_{opt}^{(j)}$ and $F^{(j)}$

Find the set $\{a_i\}$ among the 2^{20} possibilities that maximizes the correlation coefficient *R*

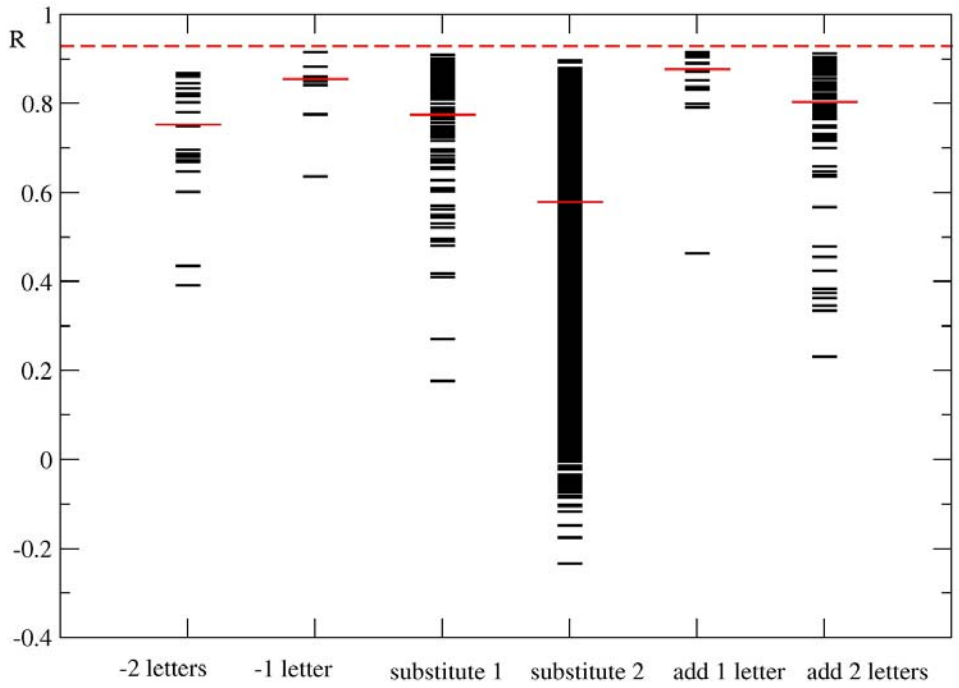
IVYWREL, or LIVEWYR



T_{opt}=937$F_{IVYWREL}$-335 , R=0.93, rmsd T_{opt}=8.9°C

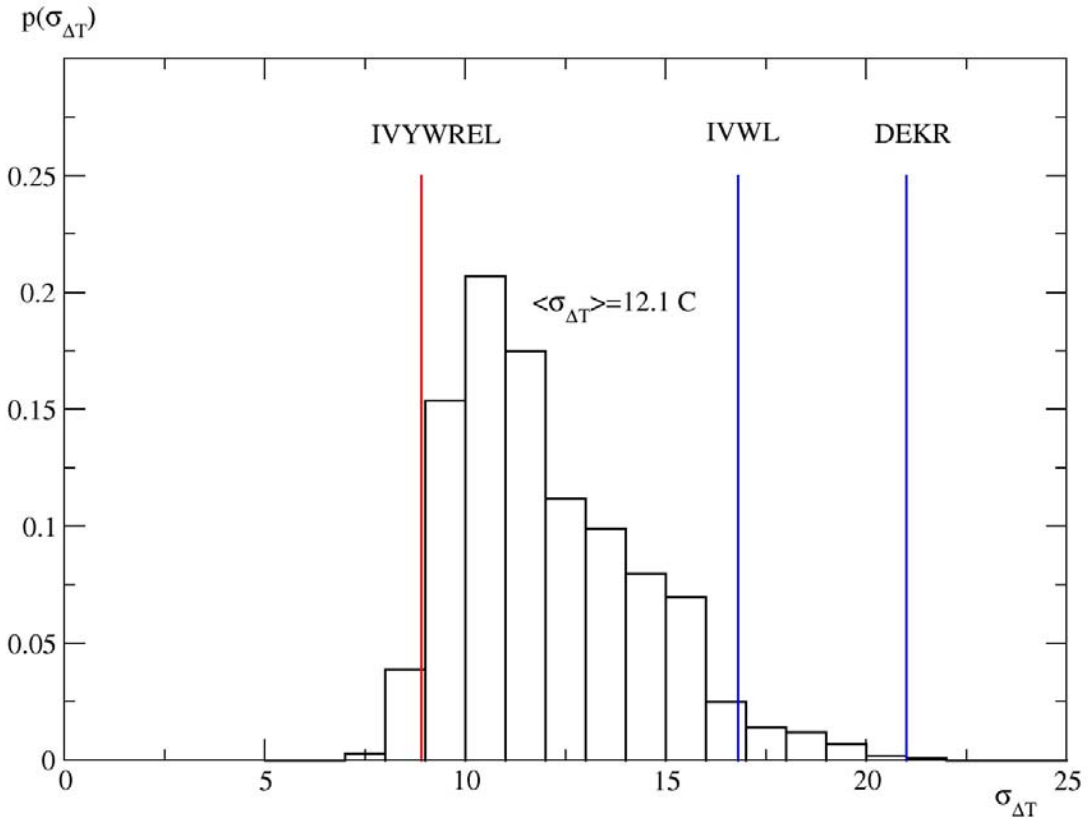
86 genomes

Robustness and stability of IVYWREL predictor



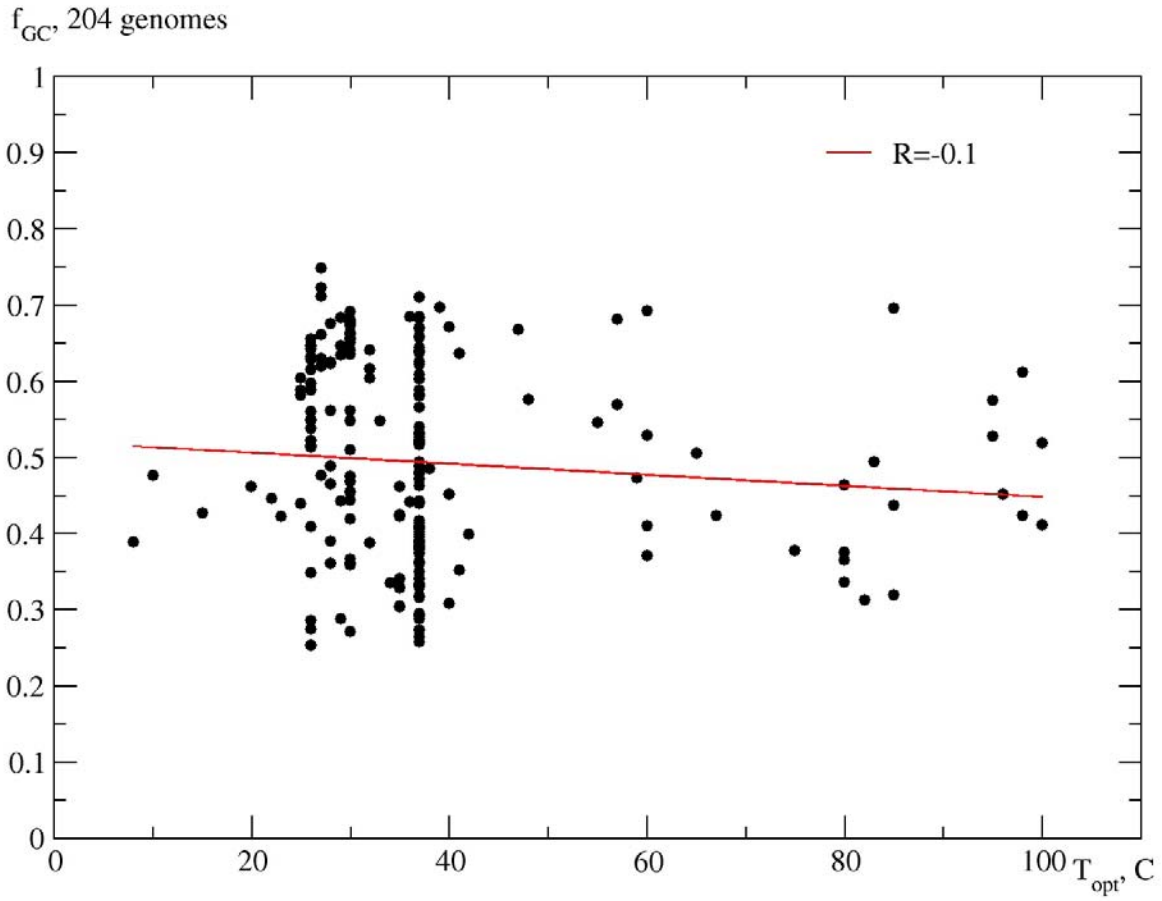
Type of change	R_{median}	Worst case		Best case	
		R_{min}	Change	R_{max}	Change
IVYWREL+1	0.877	0.47	+A	0.921	+M
IVYWREL+2	0.804	0.24	+AQ	0.917	+FP
IVYWREL-1	0.855	0.64	-I	0.921	-W
IVYWREL-2	0.754	0.40	-IE	0.874	-WE
IVYWREL subst. 1	0.776	0.18	E→A	0.914	W→H
IVYWREL subst. 2	0.580	-0.23	VE→AQ	0.902	WR→GP

Jackknife test, precision



Distribution of rmsd error of temperature prediction in the 43-genome test sets, 1000 test/train splits

DNA: any temperature, any G+C

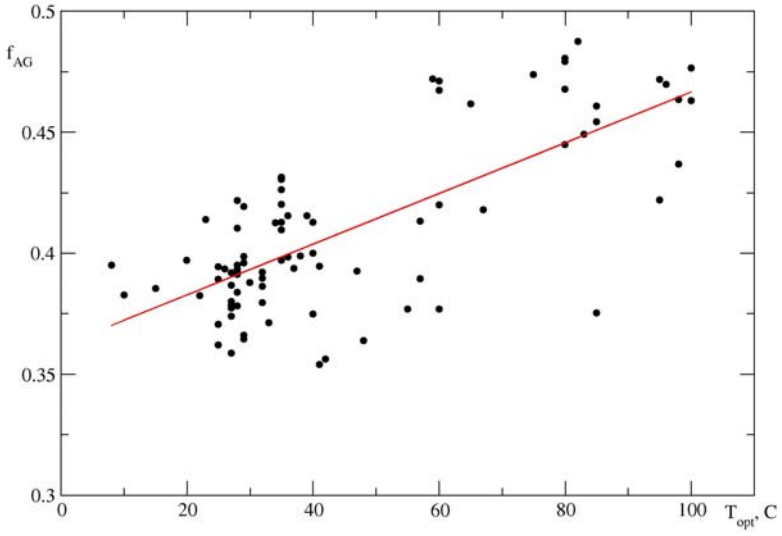


Base pairing is not the bottleneck of thermal adaptation.

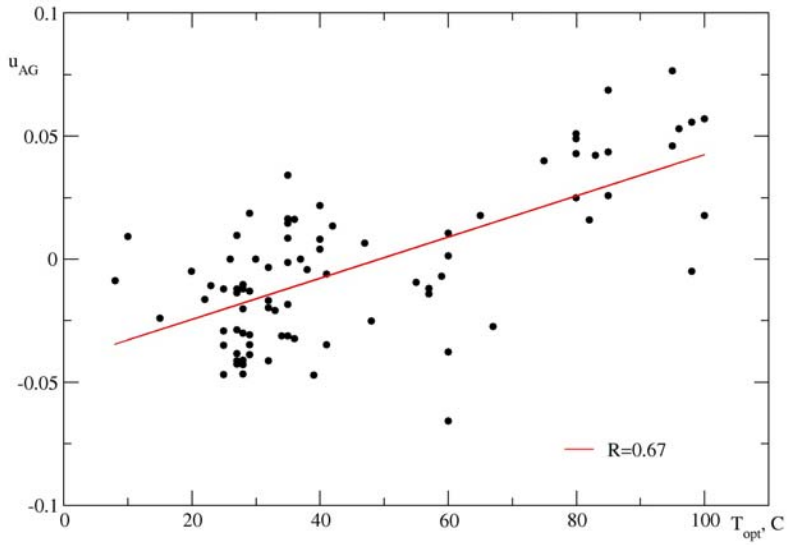
204 genomes

DNA adaptation via codon bias

Fraction of A+G



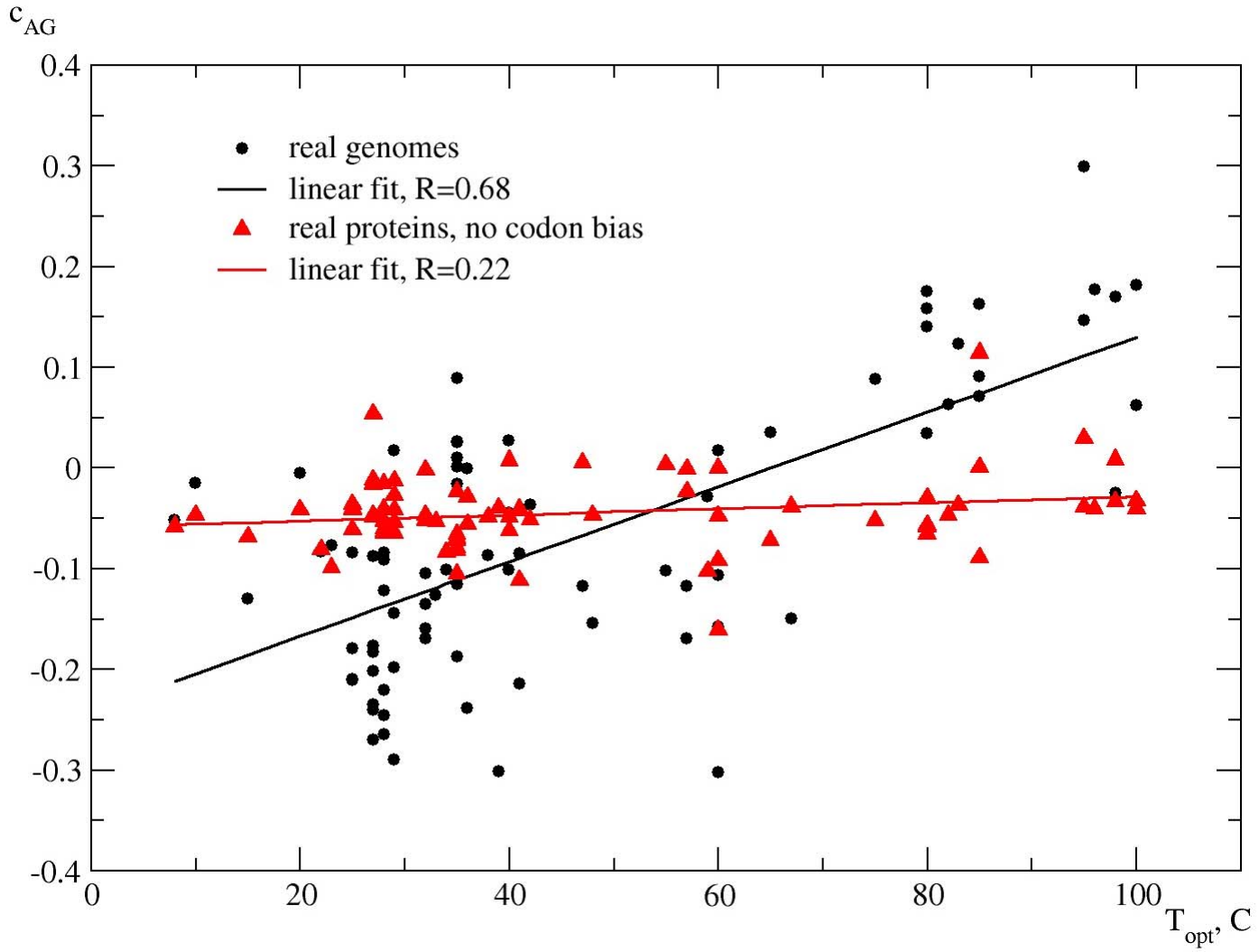
Autocorr. function of A,G



Fractions of A, G nucleotides are changing with temperature

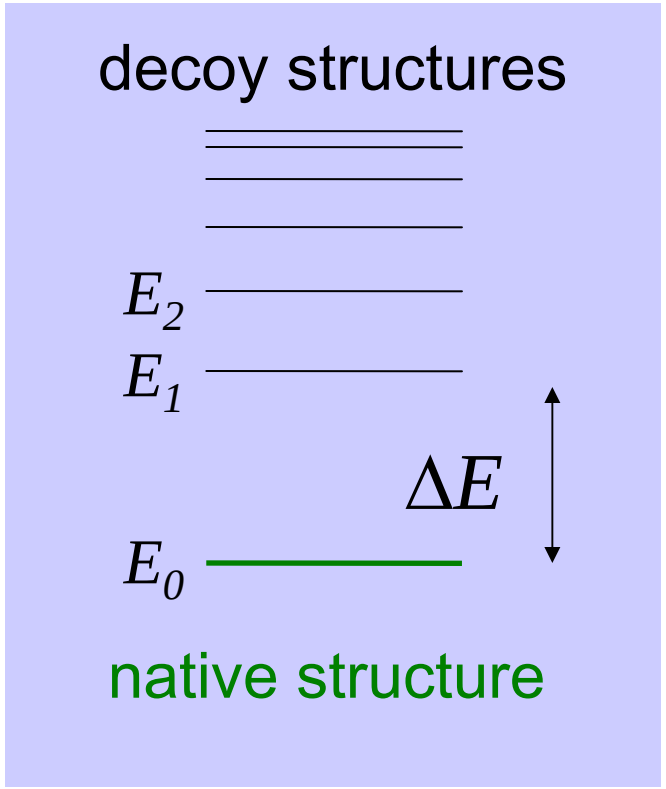
Thermal adaptation of proteins and DNA are independent processes.

Codon bias determines correlation of ApG pairs with temperature



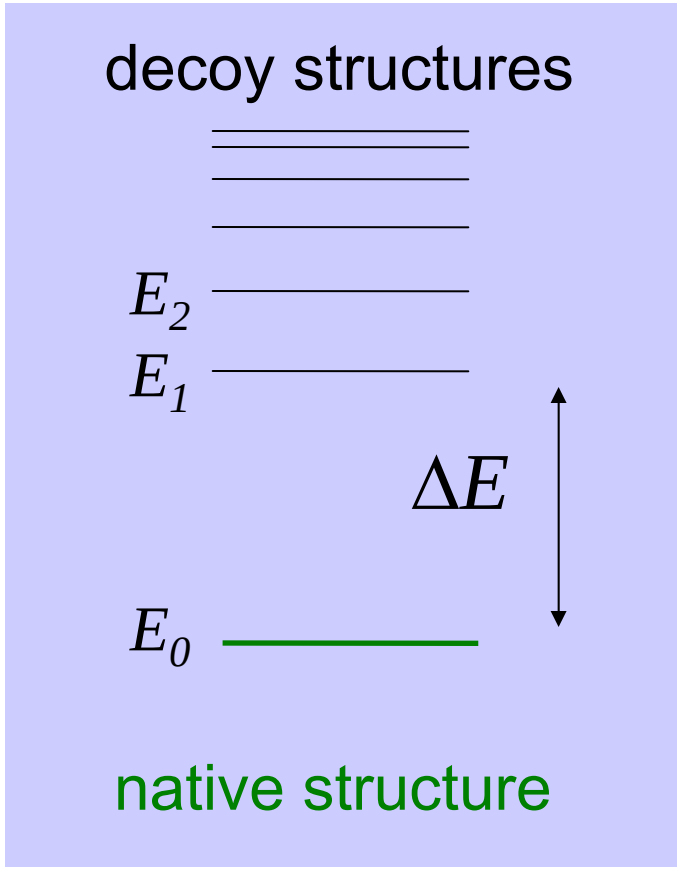
Physical mechanism of protein stability: energy gap

Proteins from mesophiles



$k_B T$

Proteins from thermophiles

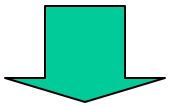


$$P_{nat}(T) \approx \frac{1}{1+e^{-\Delta E / k_B T}}$$

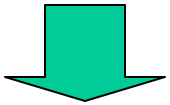
Why IVYWREL?

- Possible physical mechanism of protein thermostability:

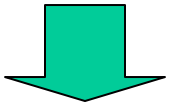
Elevated environmental temperature



Proteins need an enhanced energy gap to retain stability



Most diverse (hydrophobic and hydrophilic) residues
are recruited. An all-hydrophobic or all-charge
protein is essentially a homopolymer and is thus unstable

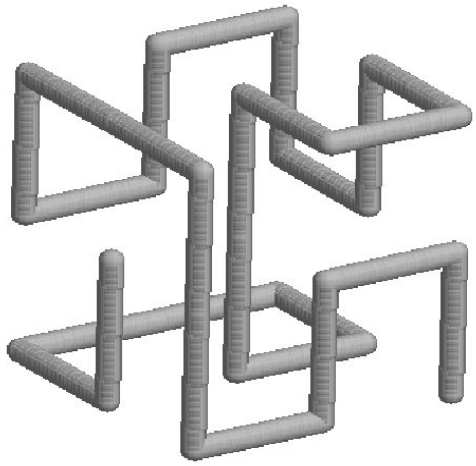


Increase of fractions of (some of) hydrophobic and hydrophilic residues.
The Nature's choice is IVYWREL

Physical model

- Complete set of compact 27-mers, 3x3x3 lattice
- Miyazawa-Jernigan pairwise contact potential

sequence



103346 structures

E_i

Probability of being in the native state

$$P(T) = \frac{e^{-E_0/T}}{\sum_{i=0}^{103345} e^{-E_i/T}}$$

E.Shakhnovich, A. Gutin, J Chem Phys 93, 5967 (1990)
S. Miyazawa, R.Jernigan, J Mol Biol 256, 623 (1996)

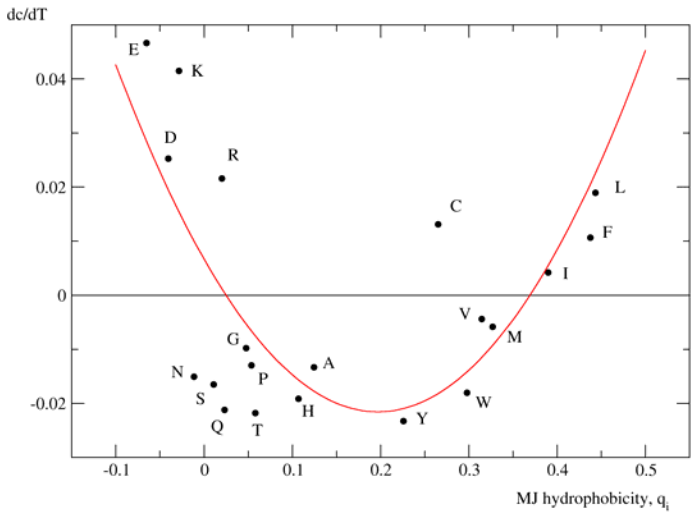
Design of thermophilic model proteins

Monte-Carlo optimization of P_{nat} in sequence space

$$P_{nat}(T) = \frac{e^{-E_0/T}}{\sum_{i=0}^{103345} e^{-E_i/T}}$$

Input: T (environmental)

Output: a.a. composition

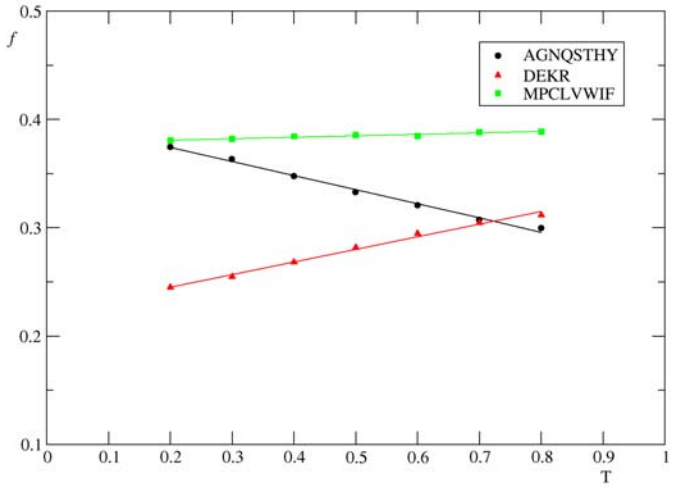


Both ends of hydrophobicity scale go up!

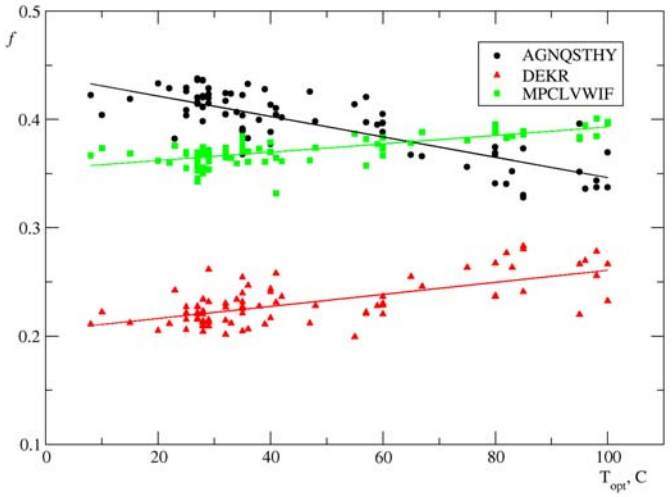
Trends of amino acid composition

Fraction of hydrophobic, charged and polar residues vs. *T*

model

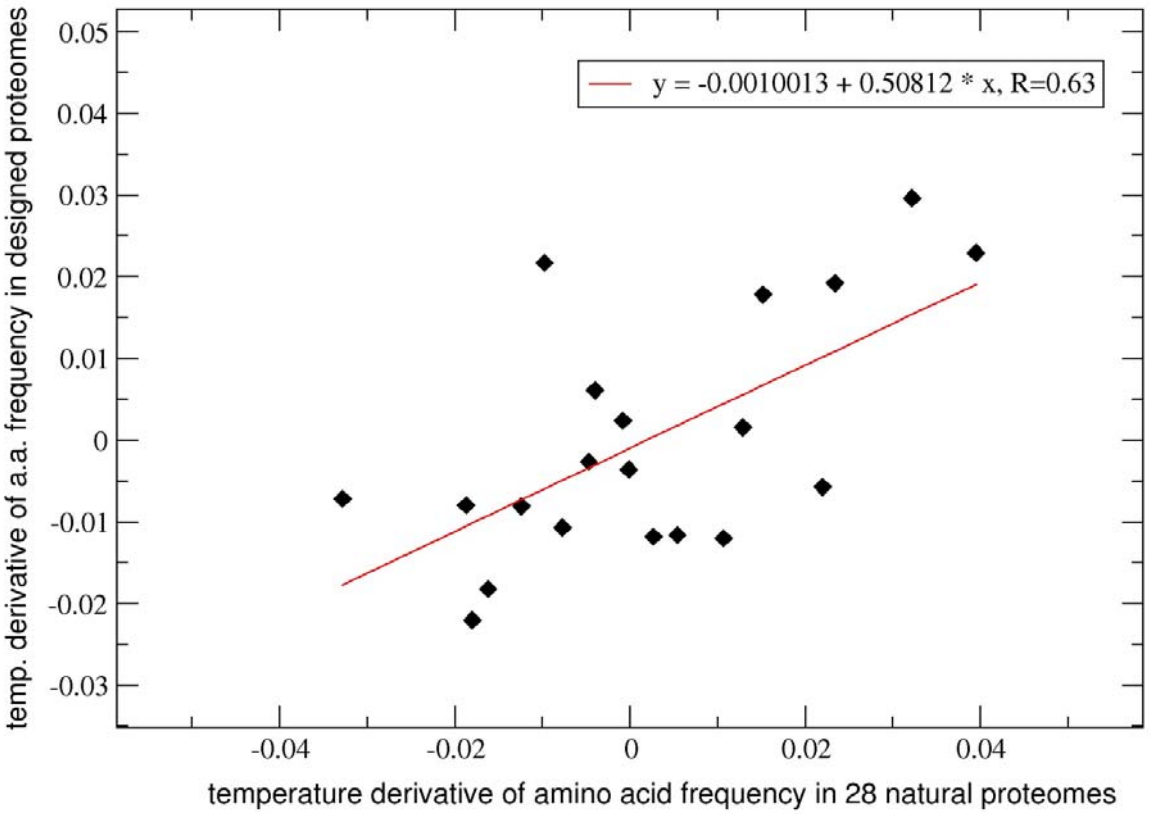


83 bacterial genomes



Temperature trends of the three major groups of amino acids are correctly predicted

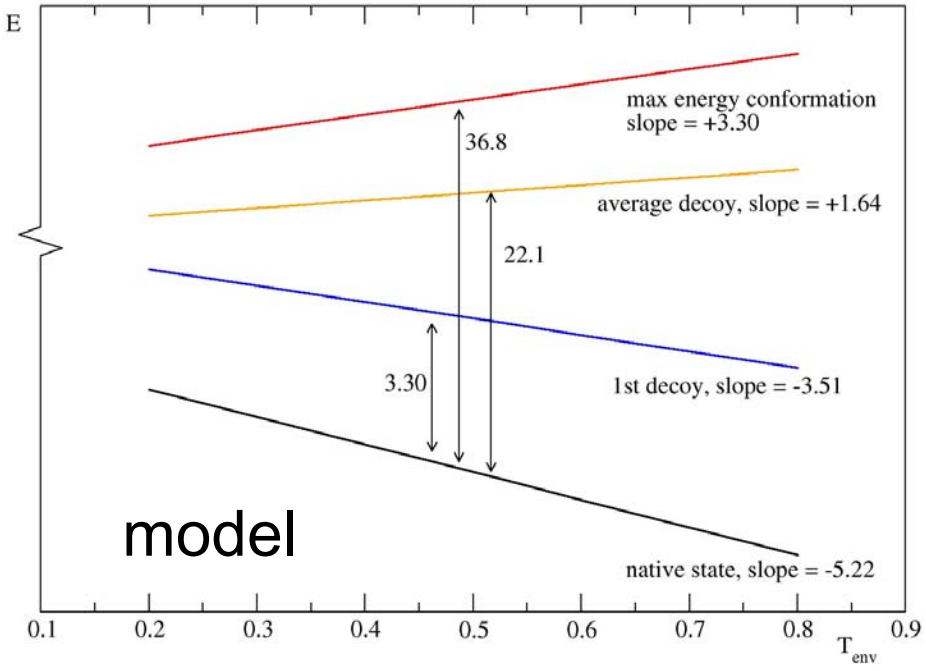
Temperature trends of individual amino acids



To a large extent, the variation of amino acid composition in prokaryotes is explained by purely physical factors

Model of thermal adaptation

Energy gap increases with environmental temperature



Charged residues form repulsive contacts in decoys
- negative design

Hydrophobic attractive interactions stabilize native state
- positive design

Positive and negative design allow for thermal adaptation

Conclusions

- Optimal growth temperature (OGT) in prokaryotes is highly correlated with amino acid composition of their proteins
- IVYWREL content predicts OGT with a ~9°C precision. Best possible discrete set of amino acids for the purpose
- Hydrophobic and charged residues ensure a larger energy gap via a combination of positive and negative design
- G+C in DNA is not correlated with OGT in 204 species
- DNA adaptation proceeds via codon bias