Evolutionary design of energy functions for protein structure prediction

Natalio Krasnogor
nxk@cs.nott.ac.uk

Paweł Widera, Jonathan Garibaldi

The University of Nottingham

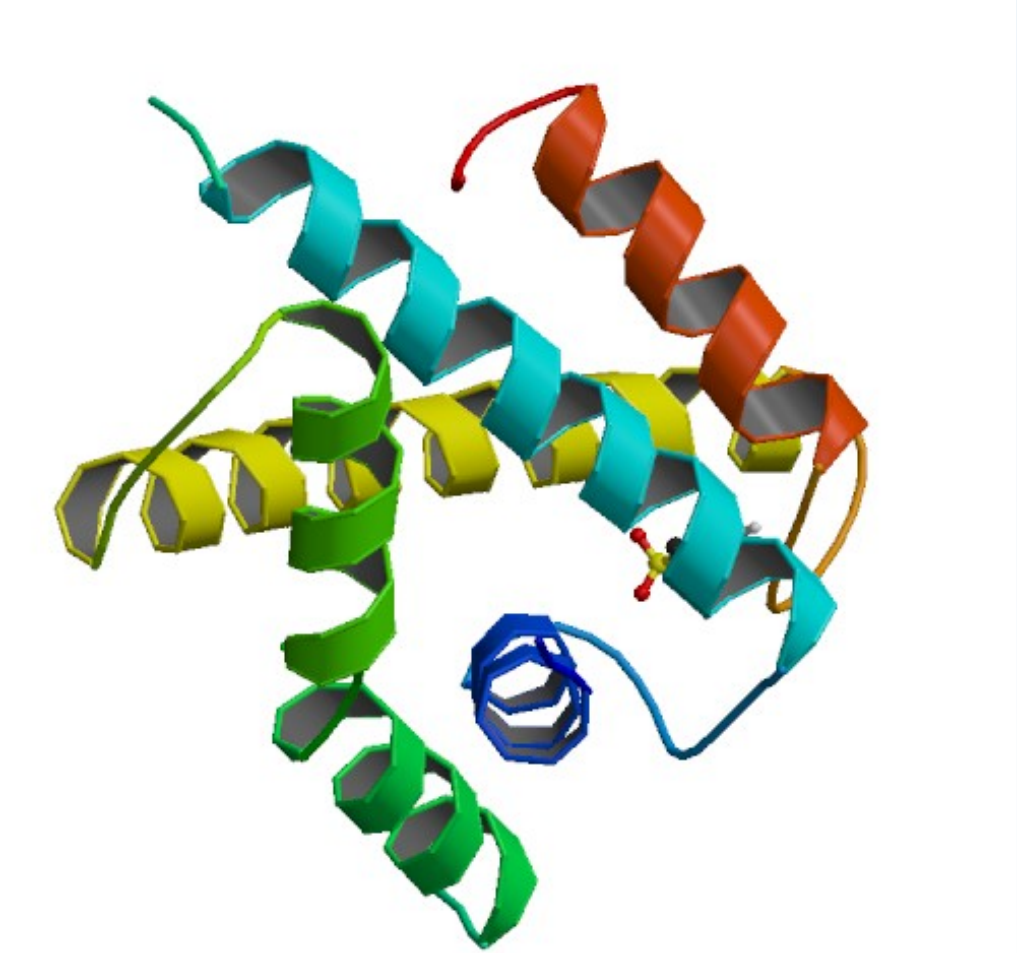
7th Annual HUMIES Awards

2010-07-09

Protein structure prediction

From 1D sequence to 3D structure

LFSKELRCMMYGFGDDQNPYTESVDILEDLVIEFITEMTHKAMSIFSEEQLNRYEMYRRSAFPKAA
IKRLIQSITGTSVSQNVVIAMSGISKVFVGEVVEEALDVCEKWGEMPPLQPKHMREAVRRLKSKGQIP

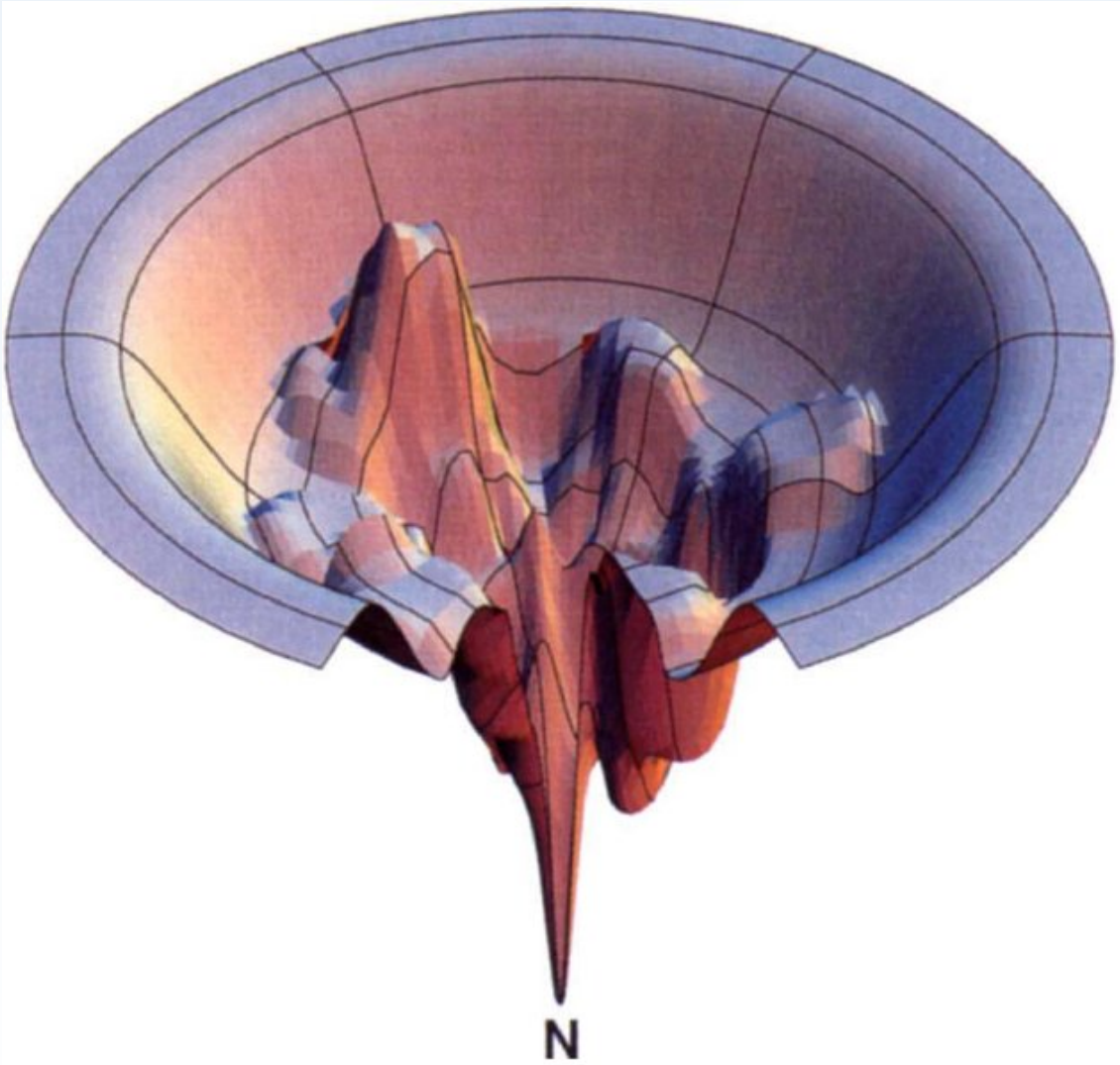


Protein basics

- 20 amino acid alphabet
- sequence encodes structure
- structure determines activity
- ratio $\frac{structures}{sequences} = 0.2\%$

The algorithm of folding

Anfinsen's thermodynamic hypothesis [Anfinsen, 1973]



[Dill and Chan, 1997]

Refolding experiment

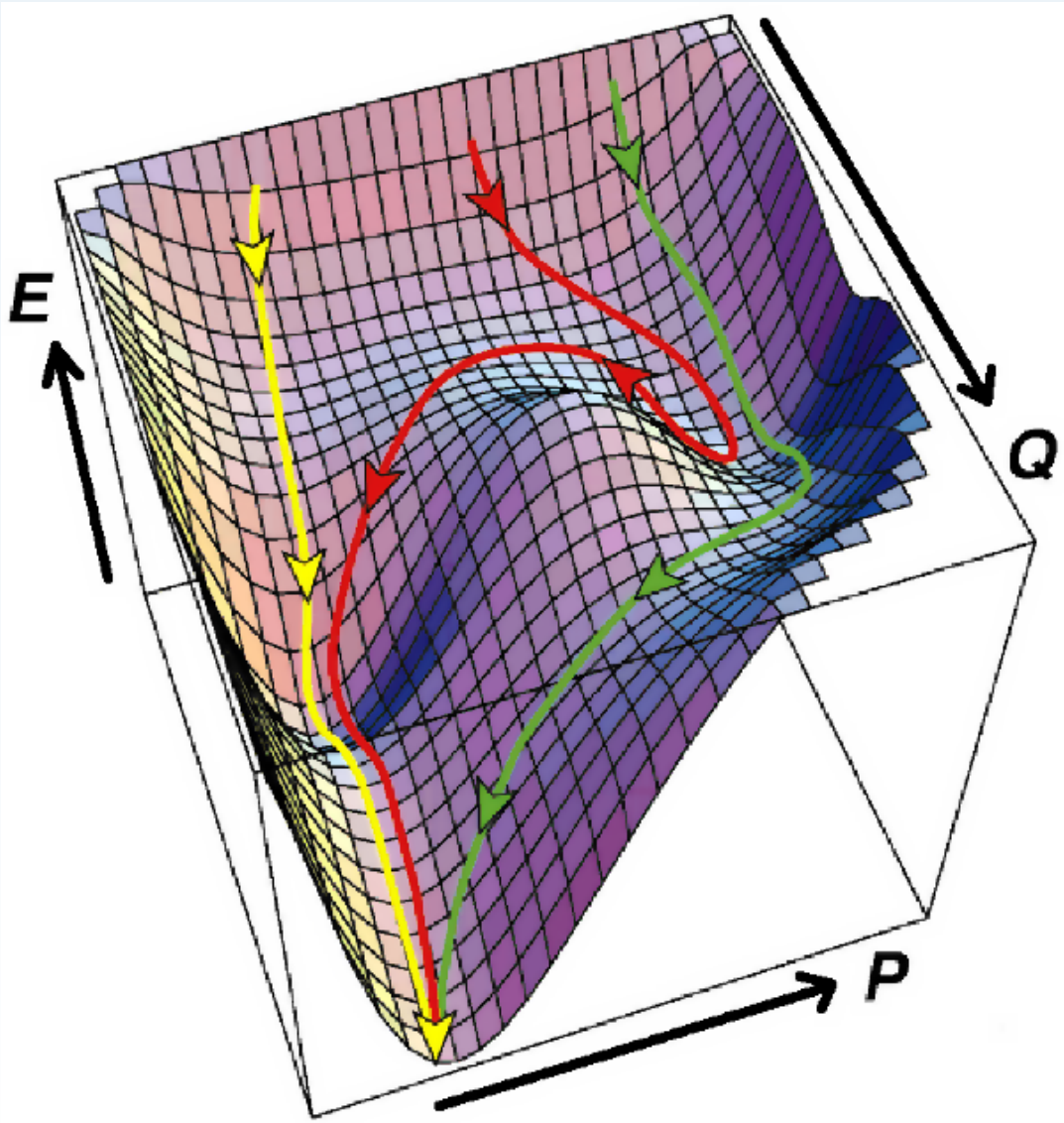
- folds to the same native state
- native state is energetically stable

Energy funnel

- roll down free energy hill
- avoid local minima traps

The two aspects of folding

Towards practical prediction



[Dill and Chan, 1997]

Energy landscape

- all-atom force field
- statistical potential

Search method

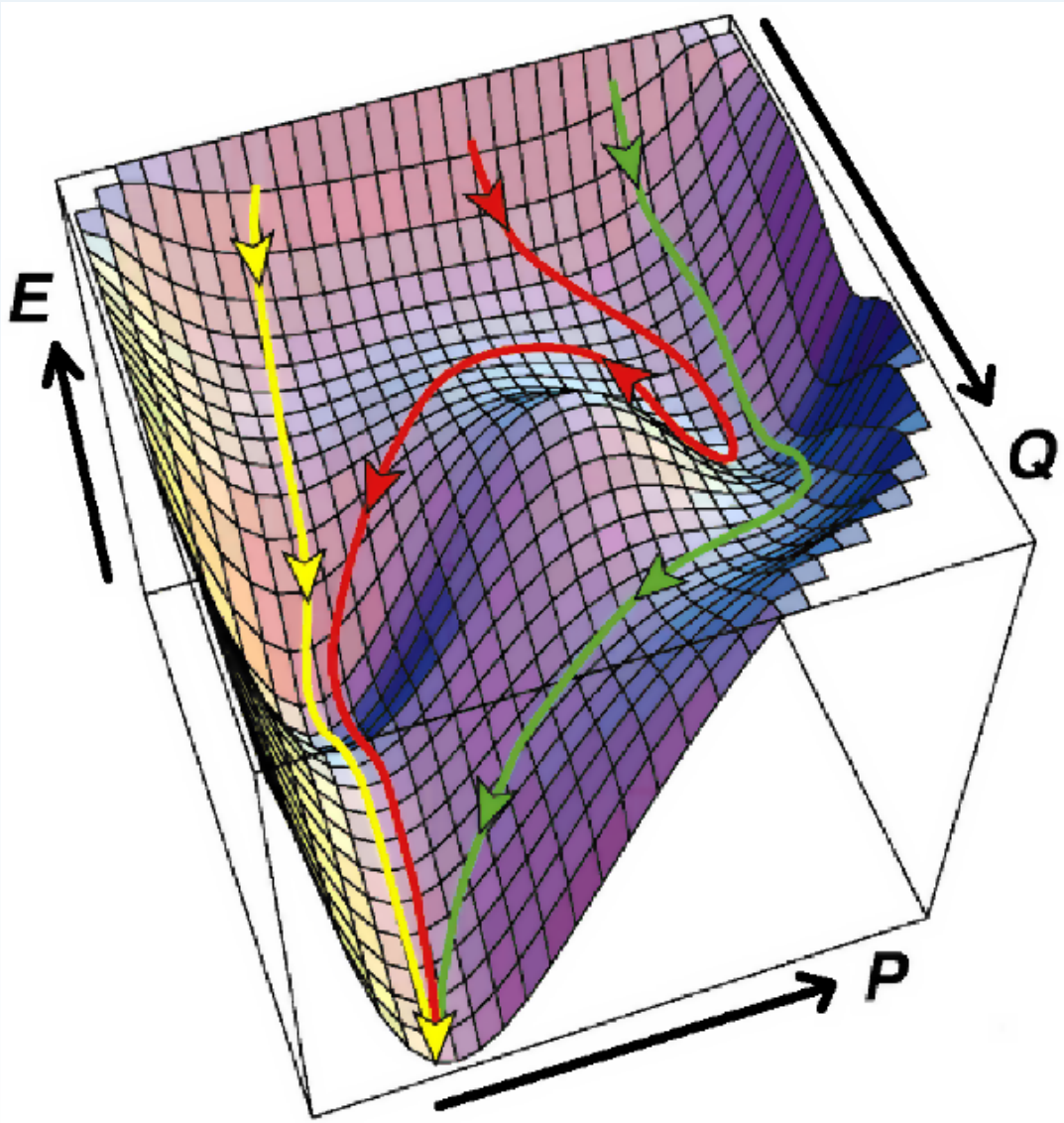
- random walk
- structure optimisation

Folding@home 8.5 peta FLOPS

- 10 000 CPU days for 10μs of folding

The two aspects of folding

Towards practical prediction



[Dill and Chan, 1997]

Energy landscape

- all-atom force field
- statistical potential

Search method

- random walk
- structure optimisation

Folding@home 8.5 peta FLOPS

- 10 000 CPU days for 10μs of folding

Community wide prediction experiment

Critical Assessment of techniques for protein Structure Prediction

CASP facts

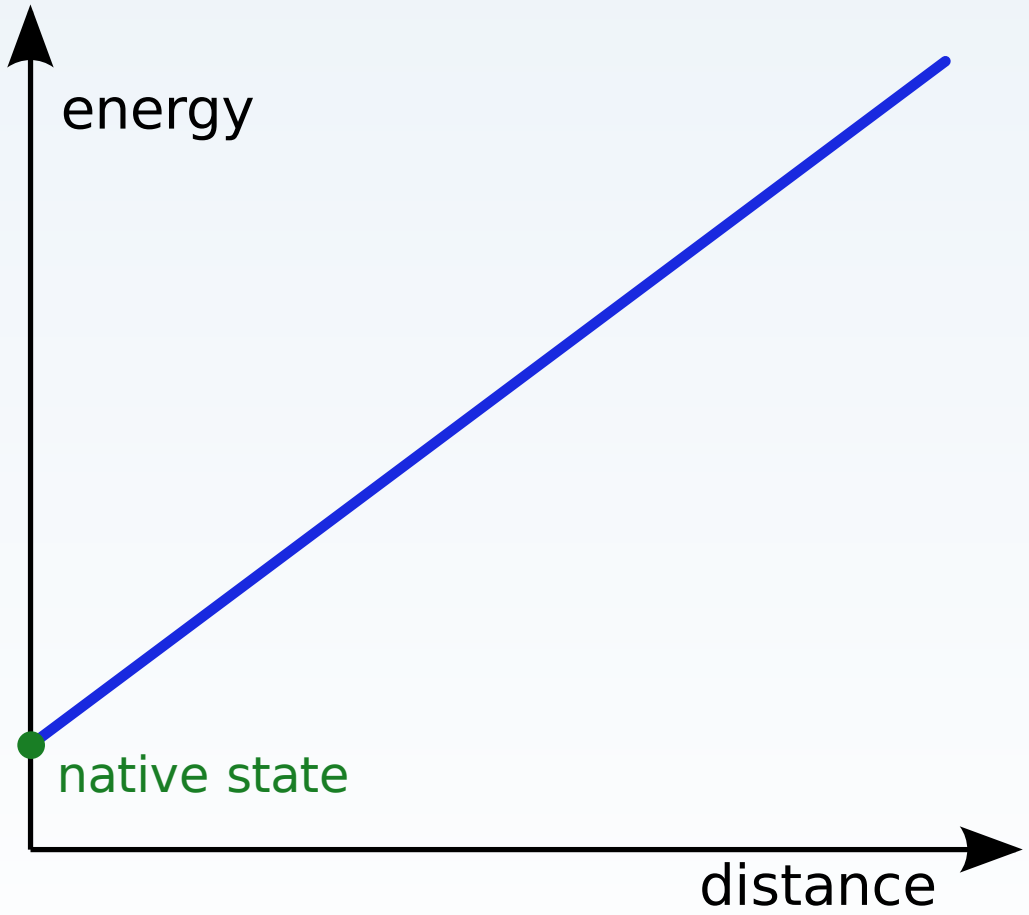
- biannual competition started in 1994
- parallel prediction and experimental verification
- model assessment by human experts

9th edition of CASP

- 150 human groups
- 140 server groups

How to find good quality models?

Correlation between energy and distance to the native structure

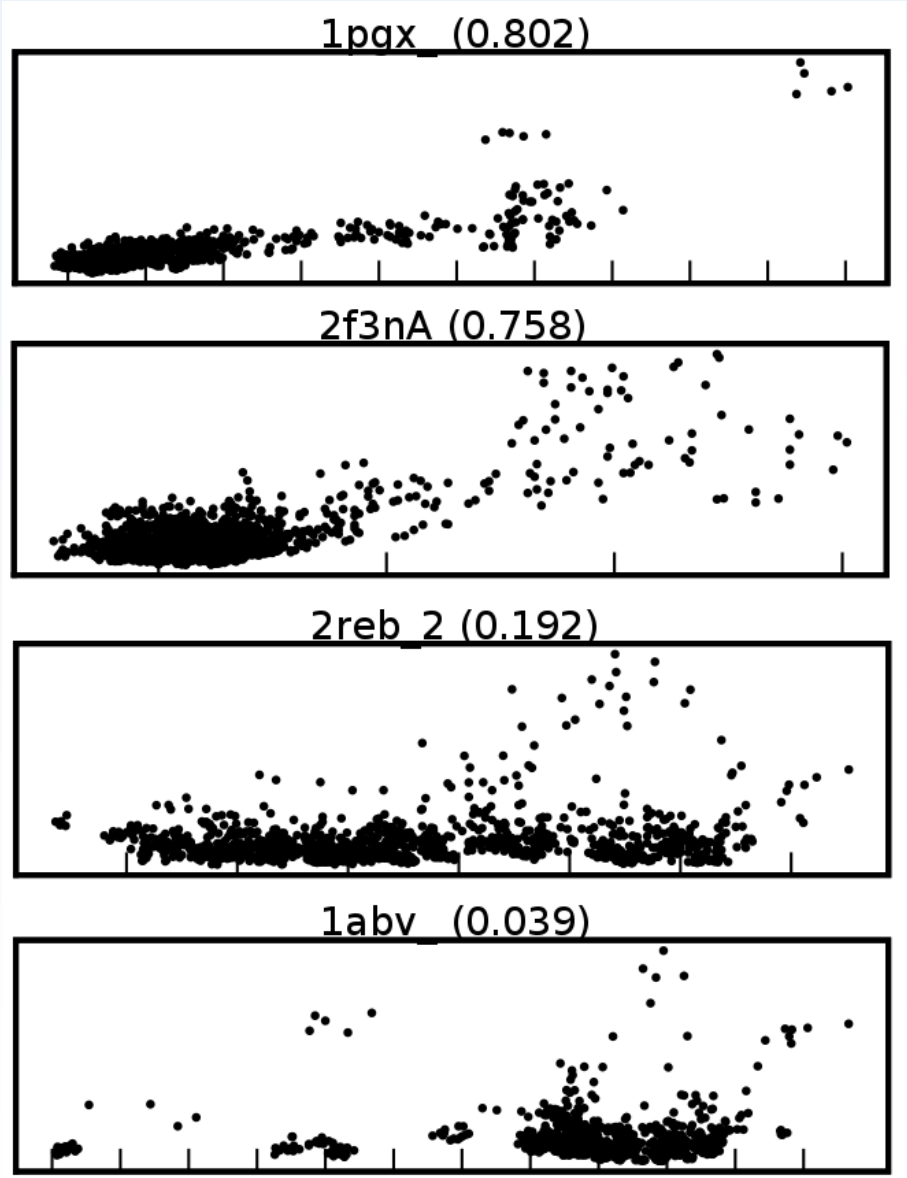


Requirements

- energy reflects distance
- distance reflects similarity

How the best of CASP do it?

Energy of models vs. distance to a target structure



Similarity measure

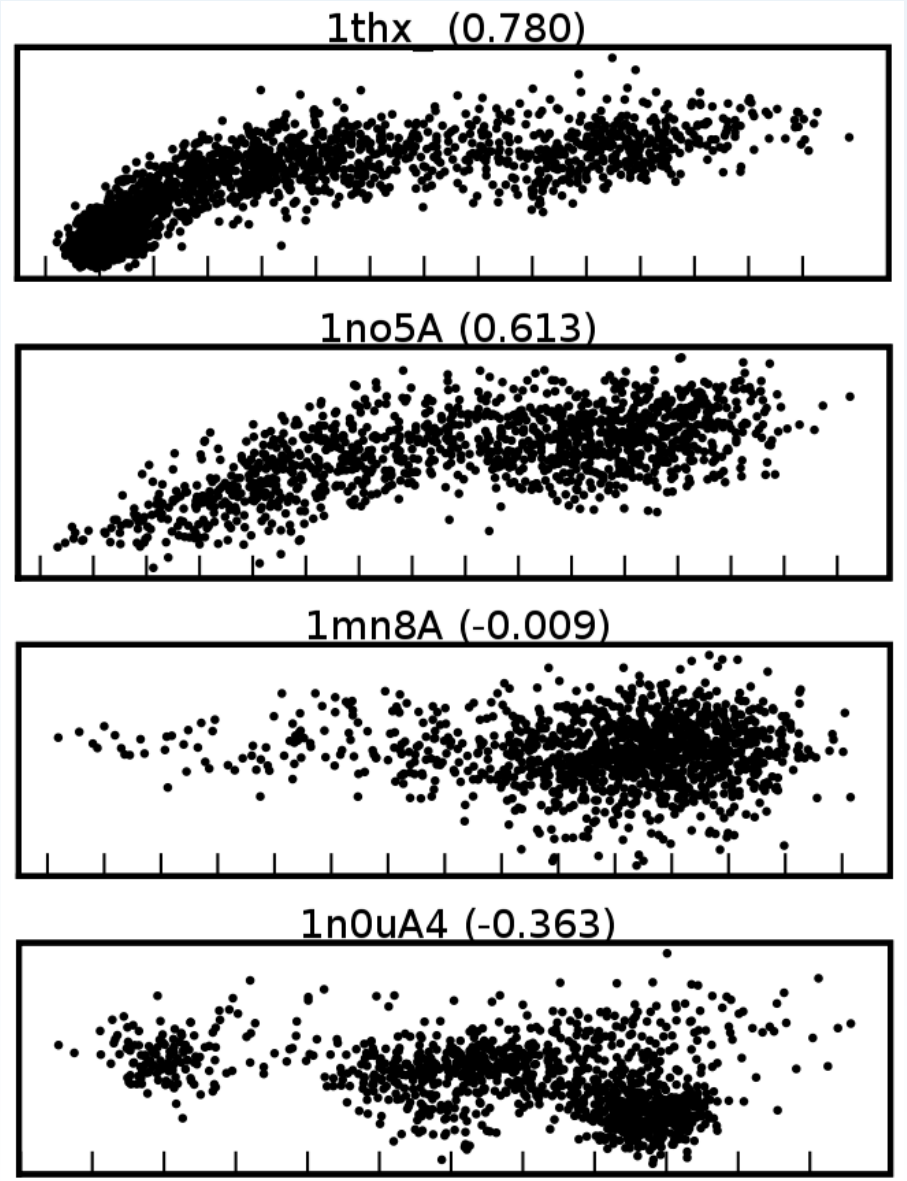
$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$

Decoys generated by

- I-TASSER
  [Wu et al., 2007]
- Robetta
  [Rohl et al., 2004]

How the best of CASP do it?

Energy of models vs. distance to a target structure



Similarity measure

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}\delta_i^2}$$

Decoys generated by

- I-TASSER
  [Wu et al., 2007]
- Robetta
  [Rohl et al., 2004]

How the energy function is designed?

Weighted sum vs. free combination of terms

$$F(\vec{T}) = w_1 * T_1 + \ldots w_n * T_n$$

[Zhang et al., 2003]

$$F(\vec{T}) = \frac{T_1 * T_3}{w_1 * log(T_2)} + sin\left(\frac{T_4 - w_2 * T_1}{T_5 * \exp(cos(w_1 * T_3))}\right)$$

Decision support

- local numerical approximation

GP input

- terminals: $T_1, \ldots, T_8$
- functions: *add sub mul div sin cos exp log*
- random ephemerals in range [0,1]

How the energy function is designed?

Weighted sum vs. free combination of terms

$$F(\vec{T}) = w_1 * T_1 + \ldots w_n * T_n$$

[Zhang et al., 2003]

$$F(\vec{T}) = \frac{T_1 * T_3}{w_1 * log(T_2)} + sin\left(\frac{T_4 - w_2 * T_1}{T_5 * \exp(cos(w_1 * T_3))}\right)$$

Decision support

- local numerical approximation

GP input

- terminals: $T_1, \ldots, T_8$
- functions: *add sub mul div sin cos exp log*
- random ephemerals in range [0,1]

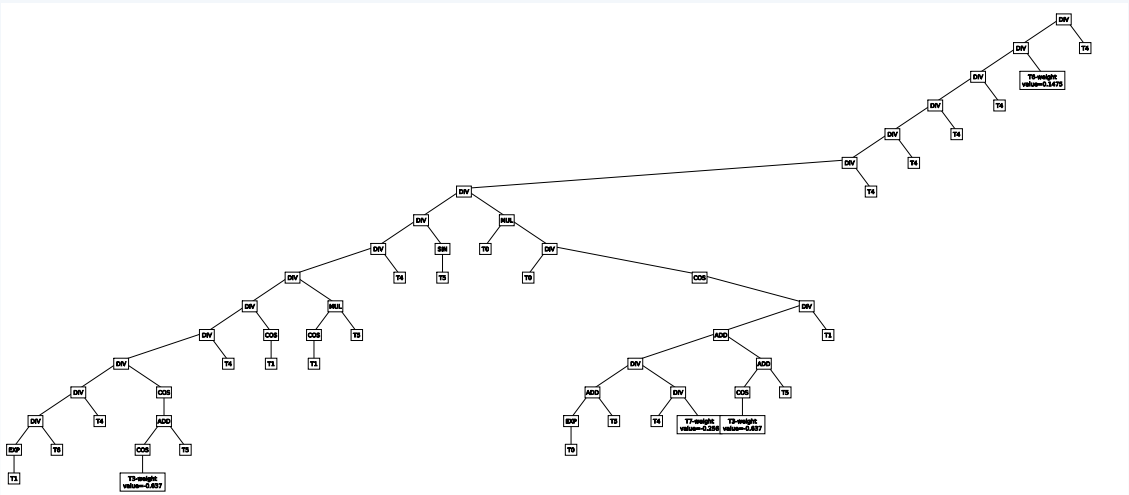
How the energy function is designed?

Weighted sum vs. free combination of terms

$$F(\vec{T}) = w_1 * T_1 + \ldots w_n * T_n$$

[Zhang et al., 2003]

$$F(\vec{T}) = \frac{T_1 * T_3}{w_1 * log(T_2)} + sin\left(\frac{T_4 - w_2 * T_1}{T_5 * \exp(cos(w_1 * T_3))}\right)$$



[Widera et al., 2010]

Decision support

- local numerical approximation

GP input

- terminals: $T_1, \ldots, T_8$
- functions: *add sub mul div sin cos exp log*
- random ephemerals in range [0,1]

Can GP improve over a weighted sum of terms?

Nelder-Mead downhill simplex optimisation

	spearman-sigmoid		correlation	
method	**d-100**	**all**	**d-100**	**all**
simplex	0.734	0.638	0.650	0.166
GP	0.835	0.714	*0.740	*0.200

Criteria for human-competitivness

CRITERION F

result >= past achievement in the field

CRITERION E

result >= most recent human-created solution to a long-standing problem

CRITERION H

result holds its own in a competition involving human contestants

Criteria for human-competitivness

CRITERION F

result $>=$ past achievement in the field

CRITERION E

result $>=$ most recent human-created solution to a long-standing problem

CRITERION H

result holds its own in a competition involving human contestants

Criteria for human-competitivness

CRITERION F

result $>=$ past achievement in the field

CRITERION E

result $>=$ most recent human-created solution to a long-standing problem

CRITERION H

result holds its own in a competition involving human contestants

Comparison to the human made solution

1. automated method to discover the best combination of the energy terms
2. human-competitive improvement to the solution of a long-standing problem
3. challenge weighted sum of terms with expert-picked weights

Potential impact

1. automated energy design using a free functional combination of terms haven't been used before
2. energy functions determines the search landscape and its smoothness is a key to the efficient prediction
3. long-term effects in protein science that the improvement in prediction quality could bring

Why this is the best entry?

1. innovates the field with a novel approach to a long-standing problem
2. could be a step towards more accurate prediction and in a long-term improve drug design and identification of disease-causing mutations
3. represent a new and difficult challange for GP
   `http://www.infobiotics.org/gpchallenge/`

References

Anfinsen, C. (1973).

Principles that Govern the Folding of Protein Chains.
Science, 181(4096):223–30.

Dill, K. A. and Chan, H. S. (1997).

From Levinthal to pathways to funnels.
Nat Struct Mol Biol, 4(1):10–19.

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004).

Protein Structure Prediction Using Rosetta.
In Brand, L. and Johnson, M. L., editors, *Numerical Computer Methods, Part D*, volume Volume 383 of *Methods in Enzymology*, pages 66–93. Academic Press.

Widera, P., Garibaldi, J., and Krasnogor, N. (2009).

Evolutionary design of the energy function for protein structure prediction.
In *IEEE Congress on Evolutionary Computation 2009*, pages 1305–1312, Trondheim, Norway.

Widera, P., Garibaldi, J., and Krasnogor, N. (2010).

GP challenge: evolving energy function for protein structure prediction.
Genetic Programming and Evolvable Machines, 11(1):61–88.

Wu, S., Skolnick, J., and Zhang, Y. (2007).

Ab initio modeling of small proteins by iterative TASSER simulations.
BMC Biol, 5(1):17.

Zhang, Y., Kolinski, A., and Skolnick, J. (2003).

TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction.
Biophys. J., 85(2):1145–1164.