Computational and Mathematical Biology in the Genomics Age: Predicting protein structures

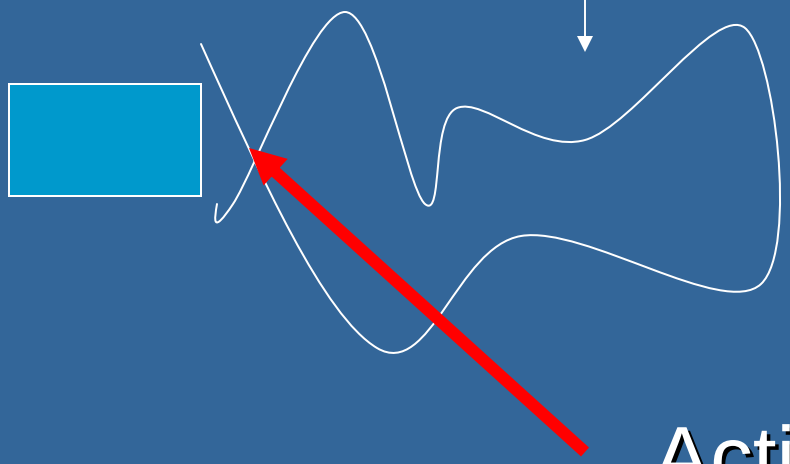
Ron Elber, Cornell

Crash course on proteins

- Proteins are one-dimensional polymers
- Made of 20 types of monomers (amino acids) with different side chains (ACDEFG…) but the same backbone
- Fold into a well defined 3D shape that includes secondary structure elements (helices, sheets)
- They are the machines of the smallest living entities (cells)

Why protein structures? Sequence determines 3D shape. Shape determines function.

ACDEFGHIJKLMNPQ



Drug design….

Active site!

Approaches to determine protein structure

- Experiment (X-ray, NMR): months

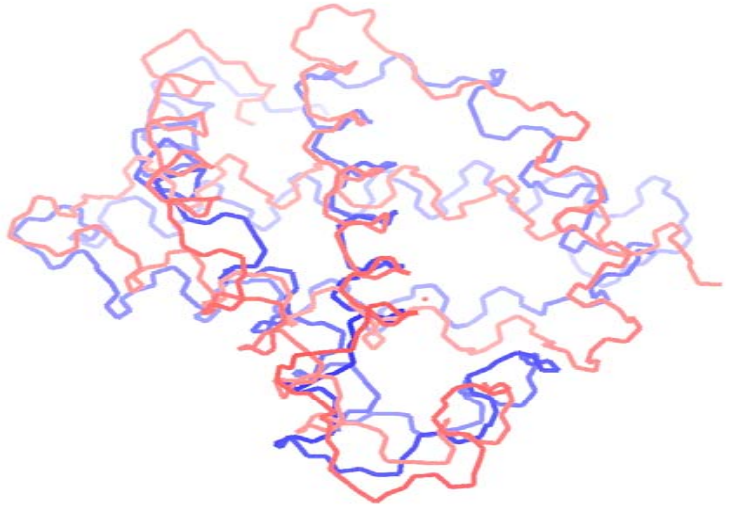
- Modeling the chemical physics weeks

- Homology based modeling: hours

Structures Are Evolutionary Templates

High degree of structural similarity is often observed in proteins with diverse sequences and in different species (below noise level – 15 percent sequence identity).

Oxygen Transport Proteins



Leghemoglobin in Plants

Myoglobin in Mammals

Three steps in homology modeling

- Identify a structural template to unknown sequence

  ACEFGH.... ⟷

- Align the unknown sequence to the structural template

A	-	C	D	W	L	K
A	R	C	-	F	L	R

- Build an atomic model based on the template

Measures of tertiary structure fitness

Instead of direct sequence comparison

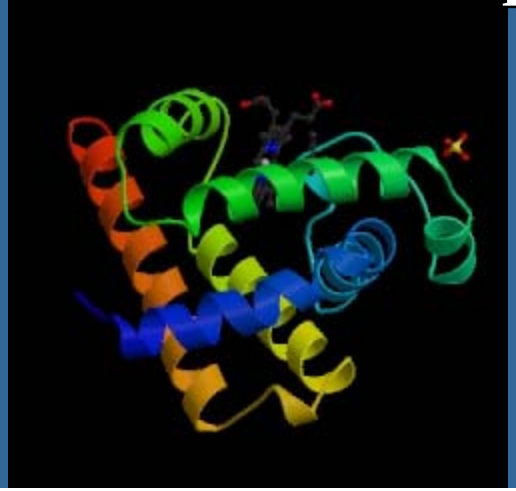
```
1BIN:A 2/3 AFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLANG-----VDPTNP
1MBC:_ 1/2 VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE-

1BIN:A 57/58 KLTGHAEKLFALVRDSAGQLKASGTVV—ADAALGSVHAQKAVTDPQFVVVKEALLKTIK
1MBC:_ 60/61 DLKKHGVTVLTALGAILKKK---GHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLH

1BIN:A 115/116 AAVGDKWSDELSRAWEVAYDELAAAIKKA
1MBC:_ 117/118 SRHPGDFGADAQGAMNKALELFRKDIAAK
```

Match unknown sequence to a known structure of a sequence

AFTEKQDALVSSSFEAFKANIPQYSVVFYTSILE
KAPAAKDLFSFLANGVDPTNPKLTGHAEKLFA
LVRDSAGQLKASGTVVADAALGSVHAQKAVT
DPQFVVVKEALLKTIKAAVGDKWSDELSRAW
EVAYDELAAAIKKA



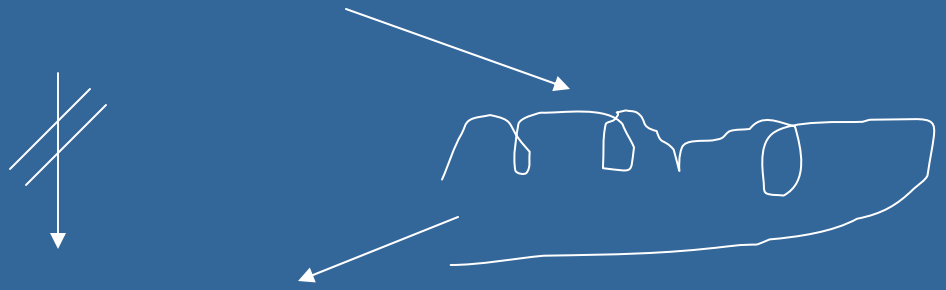
Sequence → structure → function

- Testing folds

ISTHISMYSHAPE

- Find homologs

ANYRELATIVES



PERHAPSIAM

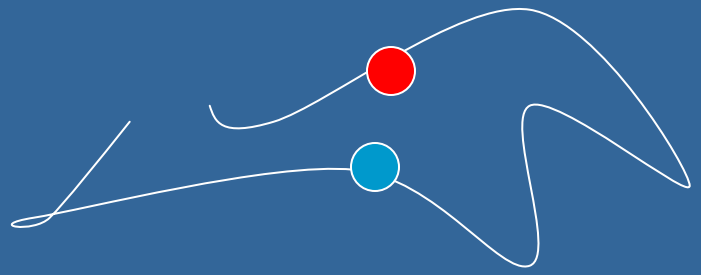
A Machine Learning Algorithm to Match a Protein Sequence to a Homolog Structure

- Potential design: Formulation and application
- Generating and learning alignments
- Applications

Potential design

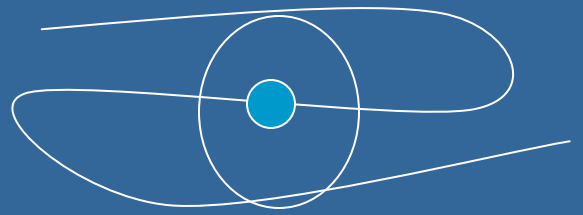
Pair or Contact potetial

$$E = \sum_{i>j} u_{ij}\left(r_{ij}; P\right)$$



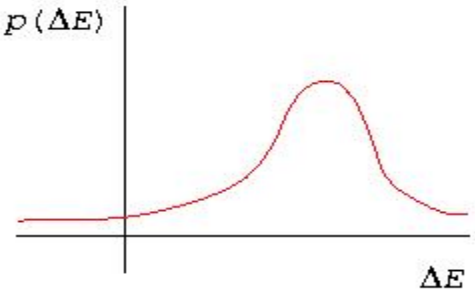
Profile potential

$$E = \sum_{i} u_{i}\left(x_{i}; P\right)$$



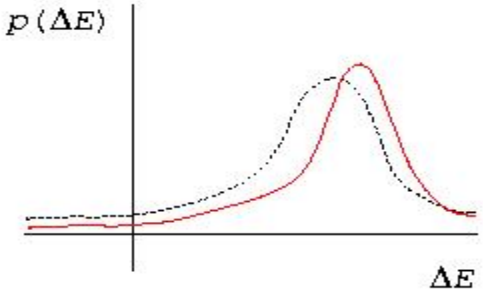
DESIGNING ACCURATE FOLDING POTENTIALS

Statistical potentials



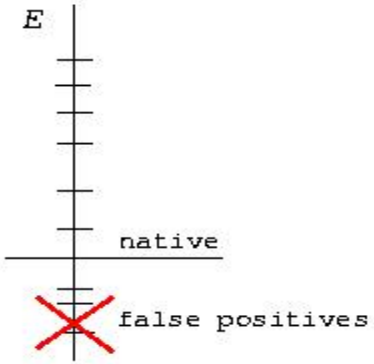
$\varepsilon_{ab} = -\ln(p_{ab}/p_a p_b)$

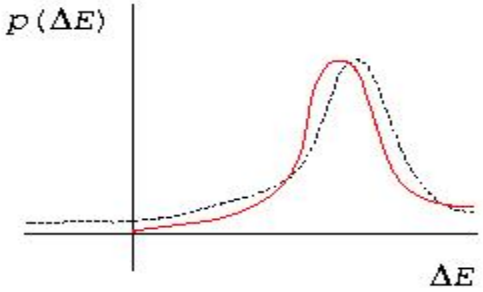
Z–score optimization



$\langle \Delta E \rangle / \sigma \rightarrow \max$

Linear Programming approach



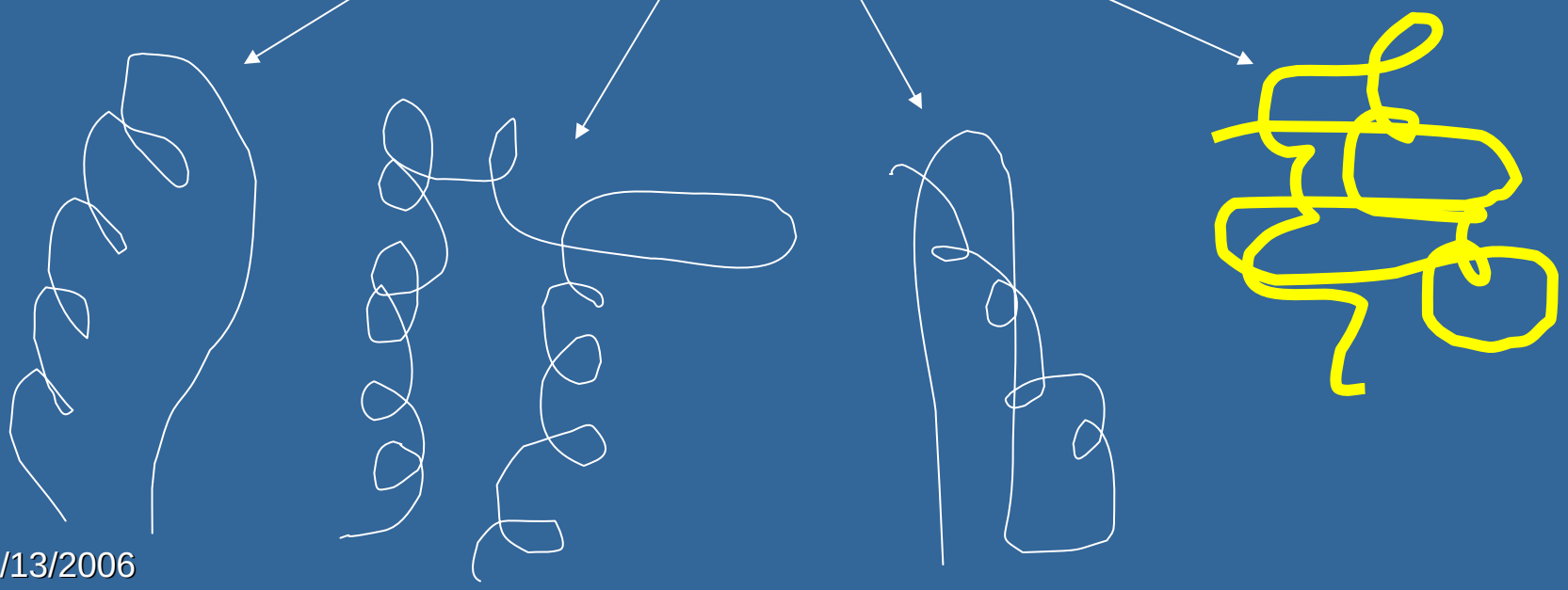


$\Delta E = E_{decoy} - E_{nat} > 0$ **for all alternative (decoy) structures**

Learning the fold that matches a sequence from the set of all known structures

$$E(S_n, X_i; P) - E(S_n, X_n; P) > 0$$

$$a_1 a_2 a_3 \ldots a_n$$



Learning folds: Find a potential that recognizes the native fold

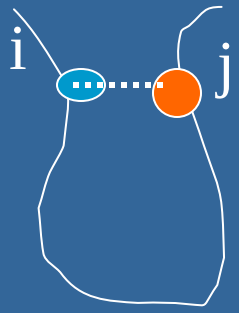
$$E(S_n, X_i; P) - E(S_n, X_n; P) > 0$$

$$E(X) = \sum_i p_i f(X)$$

$$E_{contact} = \sum_\alpha n_\alpha p_\alpha$$

Mathematical Programming approach to potential design (contact energies)

$$E = \sum_{i>j} p_{ij} = \sum_{\alpha} n_{\alpha} p_{\alpha}$$

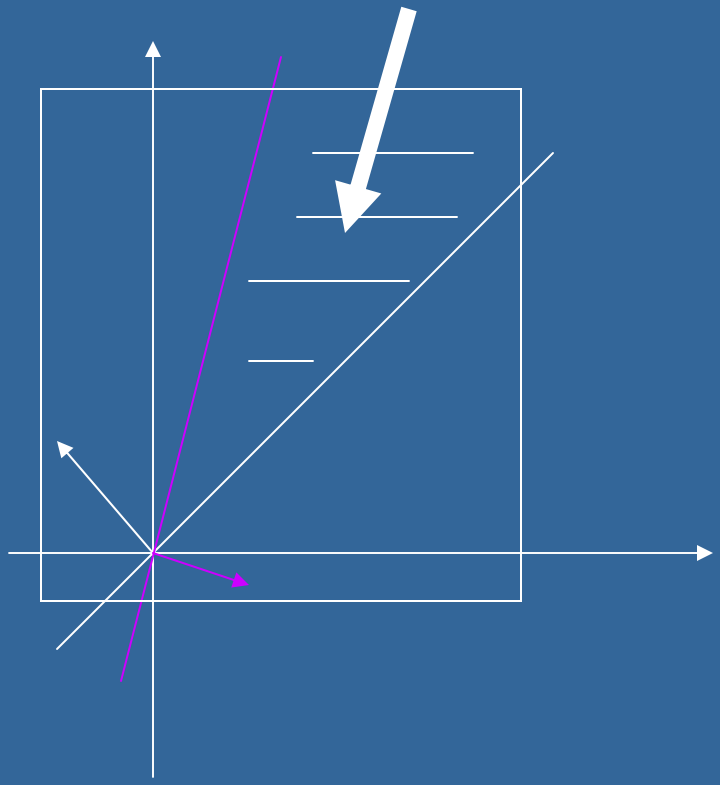


$$\Delta E_{i,nat} = E_i - E_{nat} > 1$$

$$\Delta E_{i,nat} = \sum_{\alpha} \left(n_{\alpha}^{i} - n_{\alpha}^{nat} \right) p_{\alpha} = \Delta n \cdot p > 1$$

$$\text{subject to } \|p\|^2 = \min$$

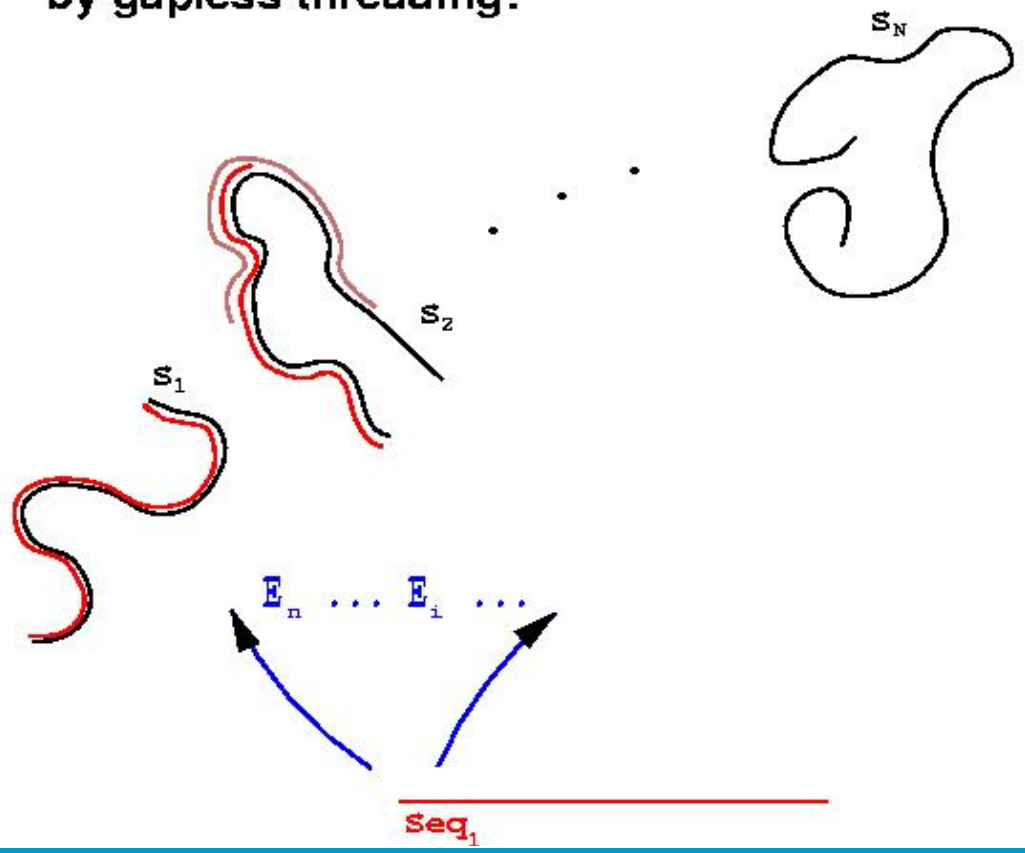
Interior point, SVM



p is the unknown

Creating decoy structures (inequalities)

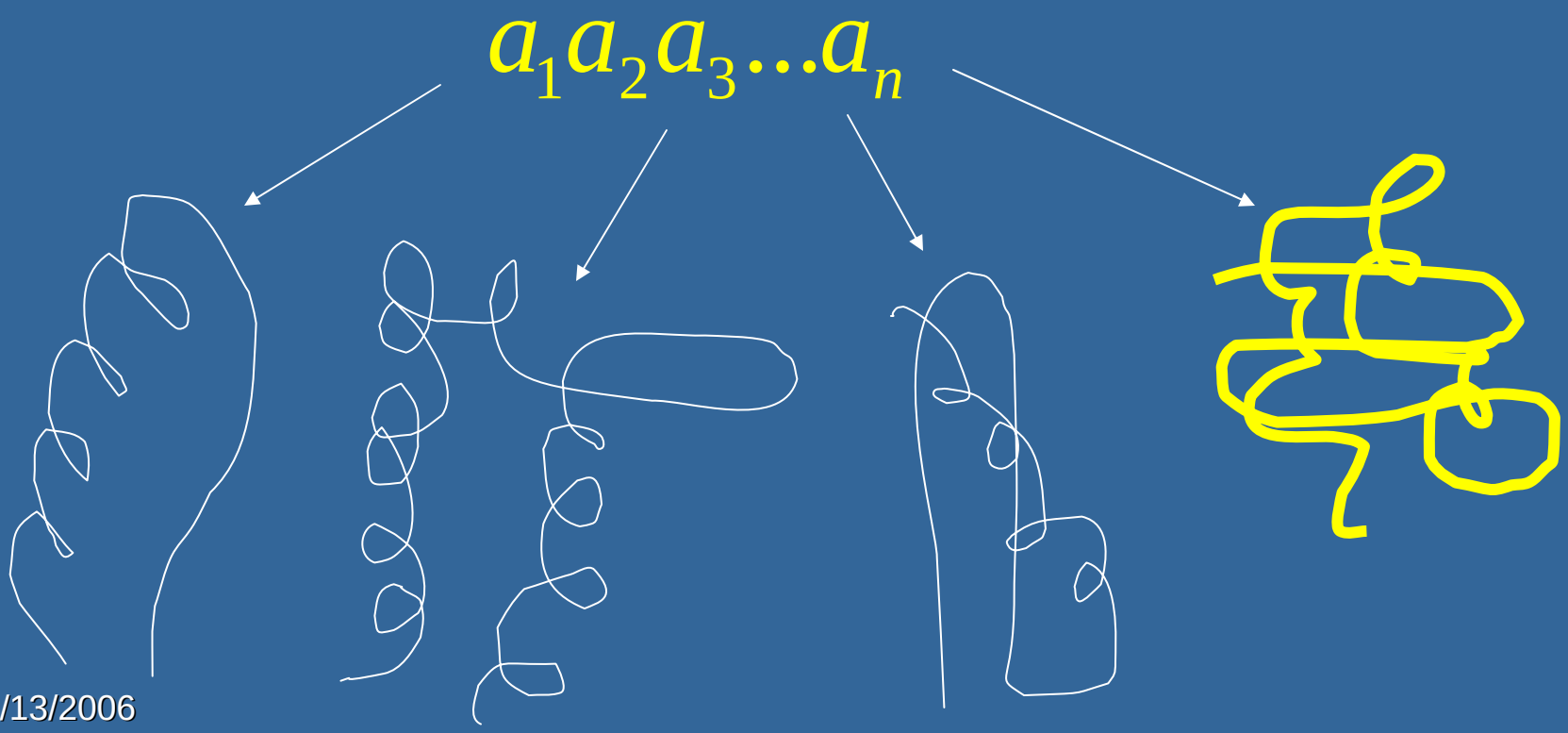
by gapless threading:



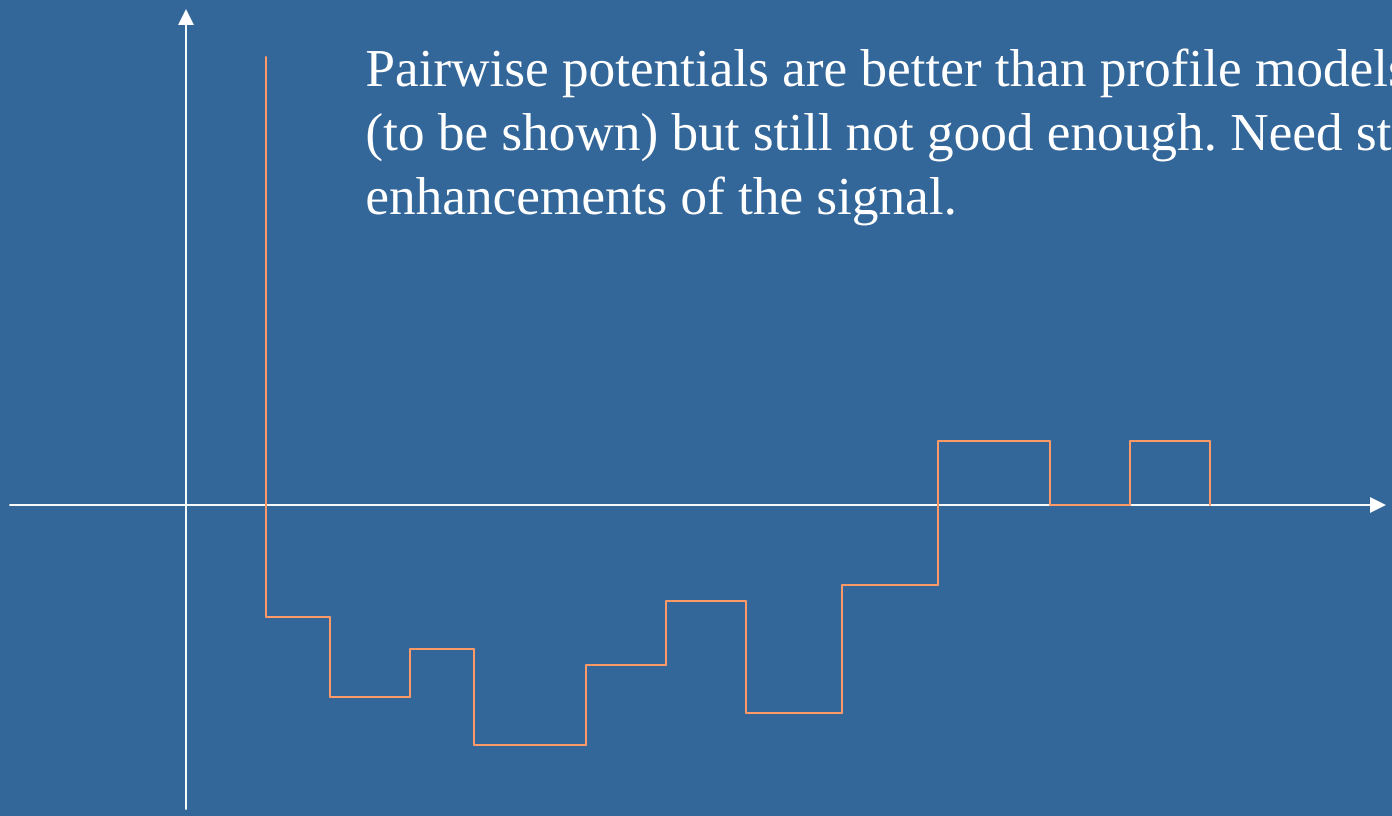
$$\Delta E_{i,nat} = \sum_{\alpha} \left(n_{\alpha}^{i} - n_{\alpha}^{nat} \right) p_{\alpha} = \Delta n \cdot p > 1$$

Learning the correct fold using 60 million comparisons between native and wrong structures

$$E(S_n, X_i) - E(S_n, X_n) > 0 \qquad \text{i=1,...,60000000}$$



General pairwise potentials are **insufficient** to recognize correct protein fold for a large set of protein-like structures (13 steps optimized independently lead to infeasibility):
Tobi & Elber, Proteins 41,40-46(2000)



Pairwise potentials are better than profile models (to be shown) but still not good enough. Need statistical enhancements of the signal.

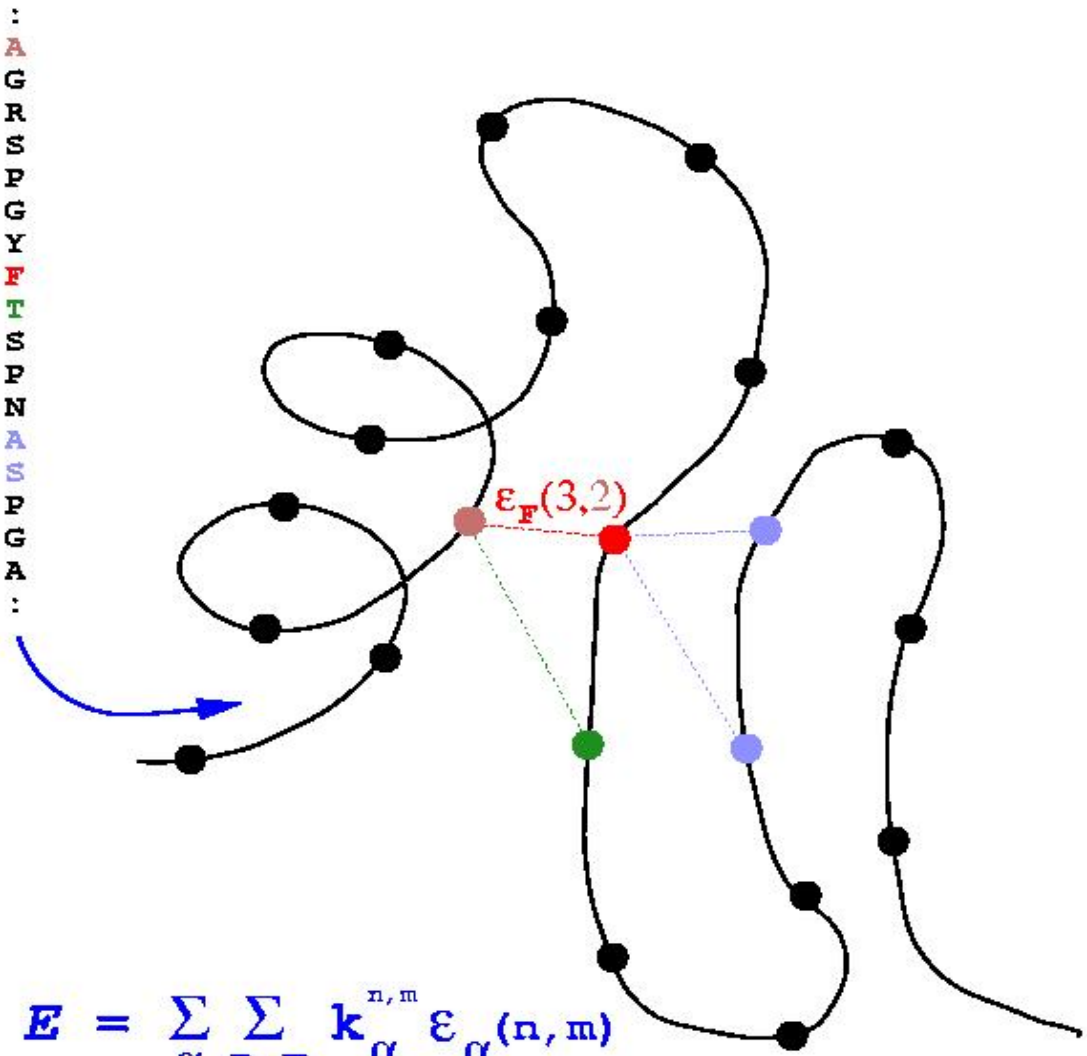
Threading Onion Model (THOM2)

An improved profile model that mixes the accuracy of pairwise energies and the efficiency of profile energies.

Defining effective pair energies in terms of structural fingerprints of sites in contact …

Threading Onion Model

with the first and second contact shells (THOM_2)



$$E = \sum_{\alpha} \sum_{n,m} k_{\alpha}^{n,m} \varepsilon_{\alpha}(n,m)$$

Contact between a site of n neighbours
and occupied by an amino acid of type α
with a site of m neighbours contributes $\varepsilon_{\alpha}(n,m)$

THOM2 yields effective pair interactions, maintaining the efficiency of profile models.

- Comparable performance to contact potentials (with 300 parameters) in terms of self-recognition
- LP derived optimal parameters (interior point algorithms!)
- Optimal alignments with gaps found using dynamic programming
- Need for gap penalties for family recognition …

Alignment

Even if we identify a homolog, the problem of structural modeling is not solved. An accurate alignment is crucial for successful modeling.

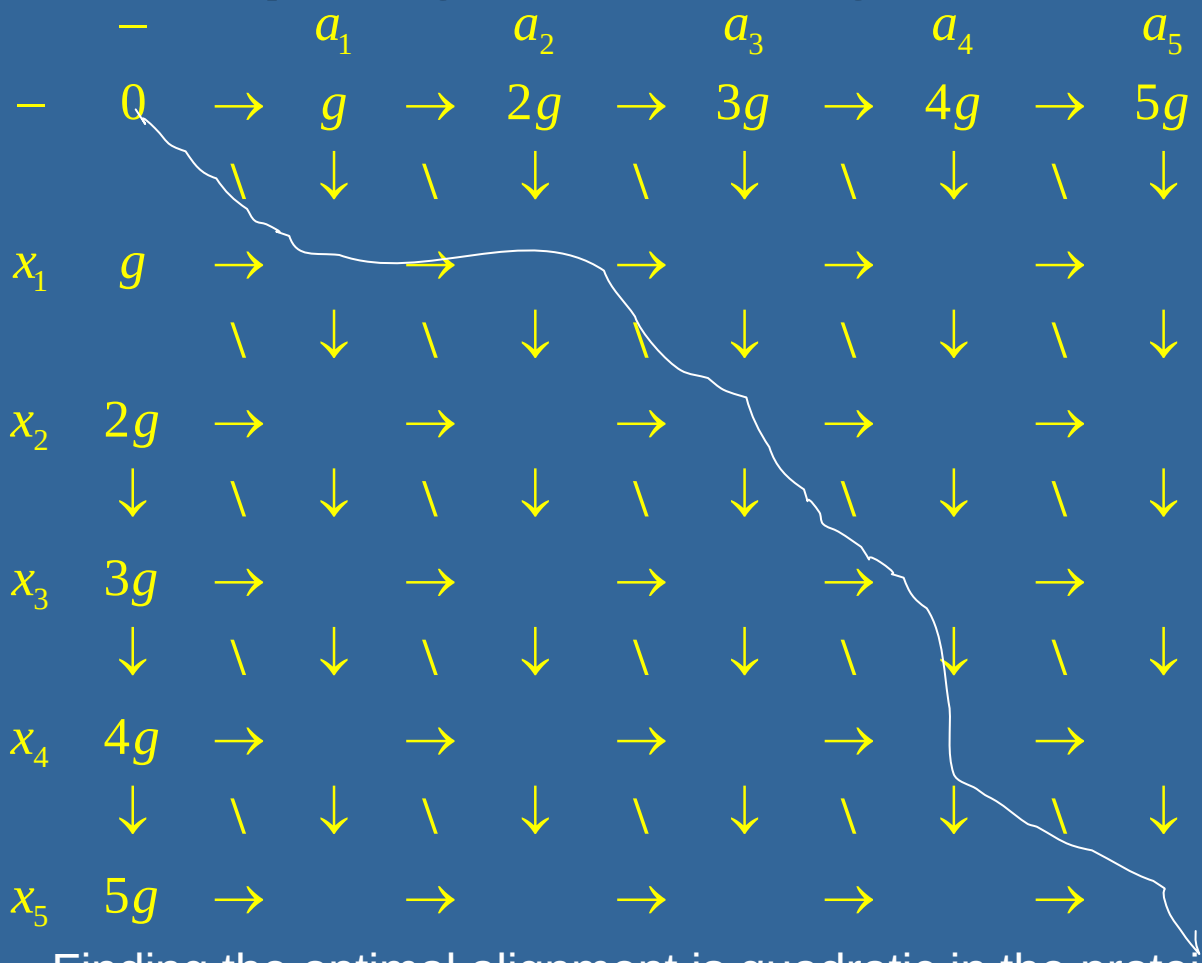
Also the presence of gaps can make the identification more difficult

$$\begin{array}{cccccc} a_1 & a_2 & - & - & a_3 & a_4 \\ x_1 & x_2 & x_3 & x_4 & x_5 & - \end{array}$$

If we need gaps we call the fitness function – score (instead of energy) and denote it by

$$T$$

An alignment is a path in a dynamic programming table



Finding the optimal alignment is quadratic in the protein length using Dynamic Programming

Dynamics programming Find optimal alignment for a given set of parameters

$T(n,m)$ The optimal score for aligning a sequence length n against a sequence length m

If we had the optimal scores for the following earlier alignments:

$$T(n-1,m-1))$$
$$T(n-1,m)$$
$$T(n,m-1)$$

can we construct the score ?

$$T(n,m)$$

Yes…

Dynamic programming: Continue

We consider three possibilities to obtain an alignment of n against m amino acids.

Option A: align n-1 against m-1 amino acids score $T(n-1,m-1)$ extend the alignment by a(n)/b(m) with a score S(an,bm)

$$T(n-1,m-1)+S(a_n,b_n)$$

Option B: align n amino acids against m-1 amino acids with a score $T(n,m-1)$ extend the alignment by -/(b(m) with a score g for a gap

$$T(n,m-1)+g$$

Option C: align n-1 amino acids against m amino acids with a score $T(n-1,m)$ Extend the alignment by a(n)/- with a corresponding score of g

$$T(n-1,m)+g$$

To decide which of the three options is optimal we need to compare the score of the three options A, B, C

Dynamic programming: Decision

$$T(n,m) = \max \begin{bmatrix} T(n-1,m-1) + S(a_n, b_m) \\ T(n,m-1) + g \\ T(n-1,m) + g \end{bmatrix}$$

How to start??

$$T(1,-) = T(-,1) = g$$

And continue (for example…) by

$$T(a_1,b_1) = \max\begin{bmatrix} T(a_1,-)+g \\ T(-,b_1)+g \\ T(0,0)+S(a_1,b_1) \end{bmatrix} = \max\begin{bmatrix} 2g \\ 2g \\ S(a_1,b_1) \end{bmatrix}$$

Here we start…

	–		a_1		a_2		a_3		a_4		a_5
–	0	→	g	→	$2g$	→	$3g$	→	$4g$	→	$5g$
		\	↓	\	↓	\	↓	\	↓	\	↓
b_1	g	→		→		→		→		→	
		\	↓	\	↓	\	↓	\	↓	\	↓
b_2	$2g$	→		→		→		→		→	
	↓	\	↓	\	↓	\	↓	\	↓	\	↓
b_3	$3g$	→		→		→		→		→	
	↓	\	↓	\	↓	\	↓	\	↓	\	↓
b_4	$4g$	→		→		→		→		→	
	↓	\	↓	\	↓	\	↓	\	↓	\	↓
b_5	$5g$	→		→		→		→		→	

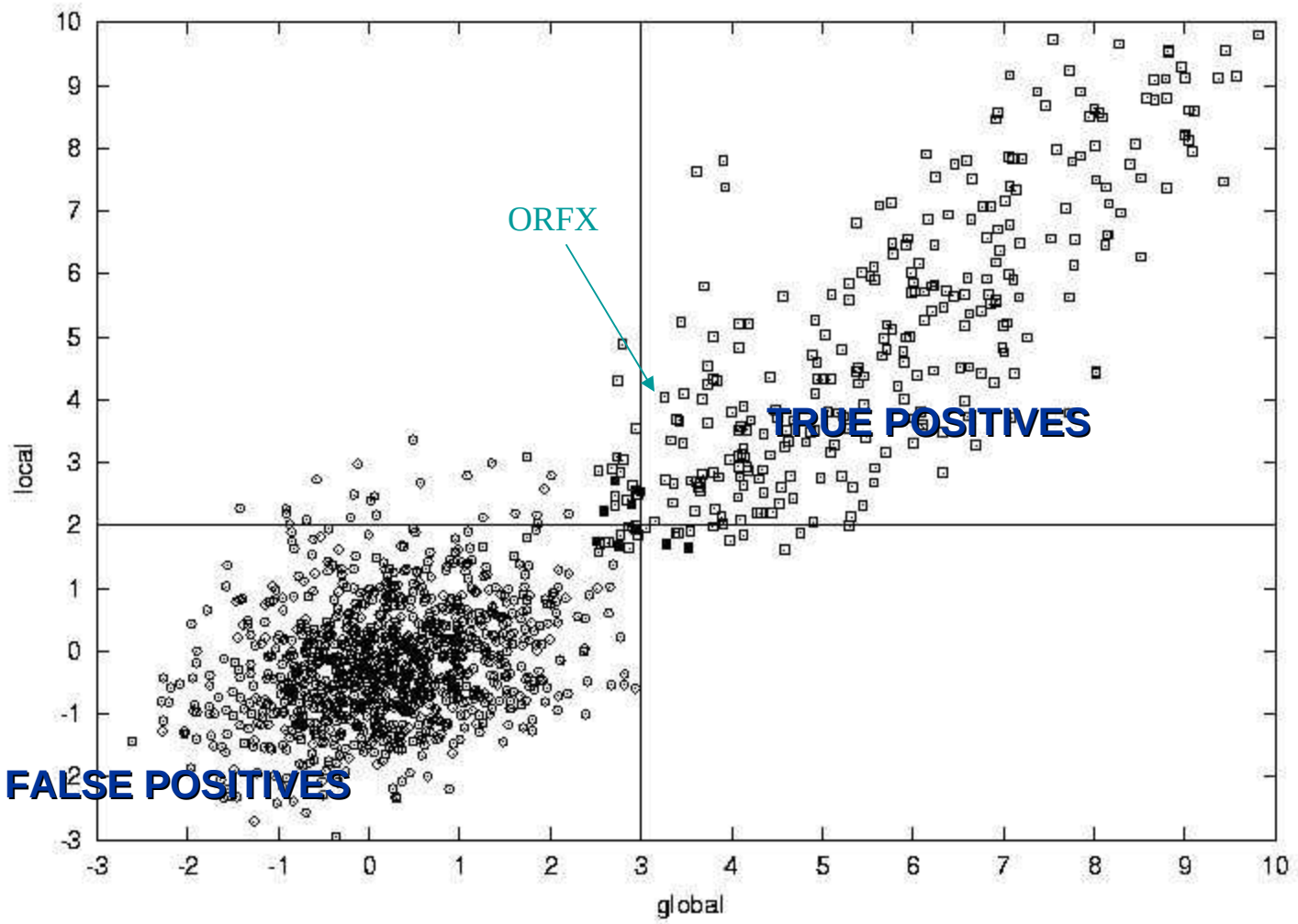
13 step potential one of the best around
(tested on the Baker's set, 65 sets –
Tamara Galor)

	aver. pos.	# correct	Z score
TE13	27	40	4.3
MJ	150	23	2.1
HL	163	15	2.0
SK	158	11	1.8
BT	148	15	2.0
THOM2	106	15	2.0

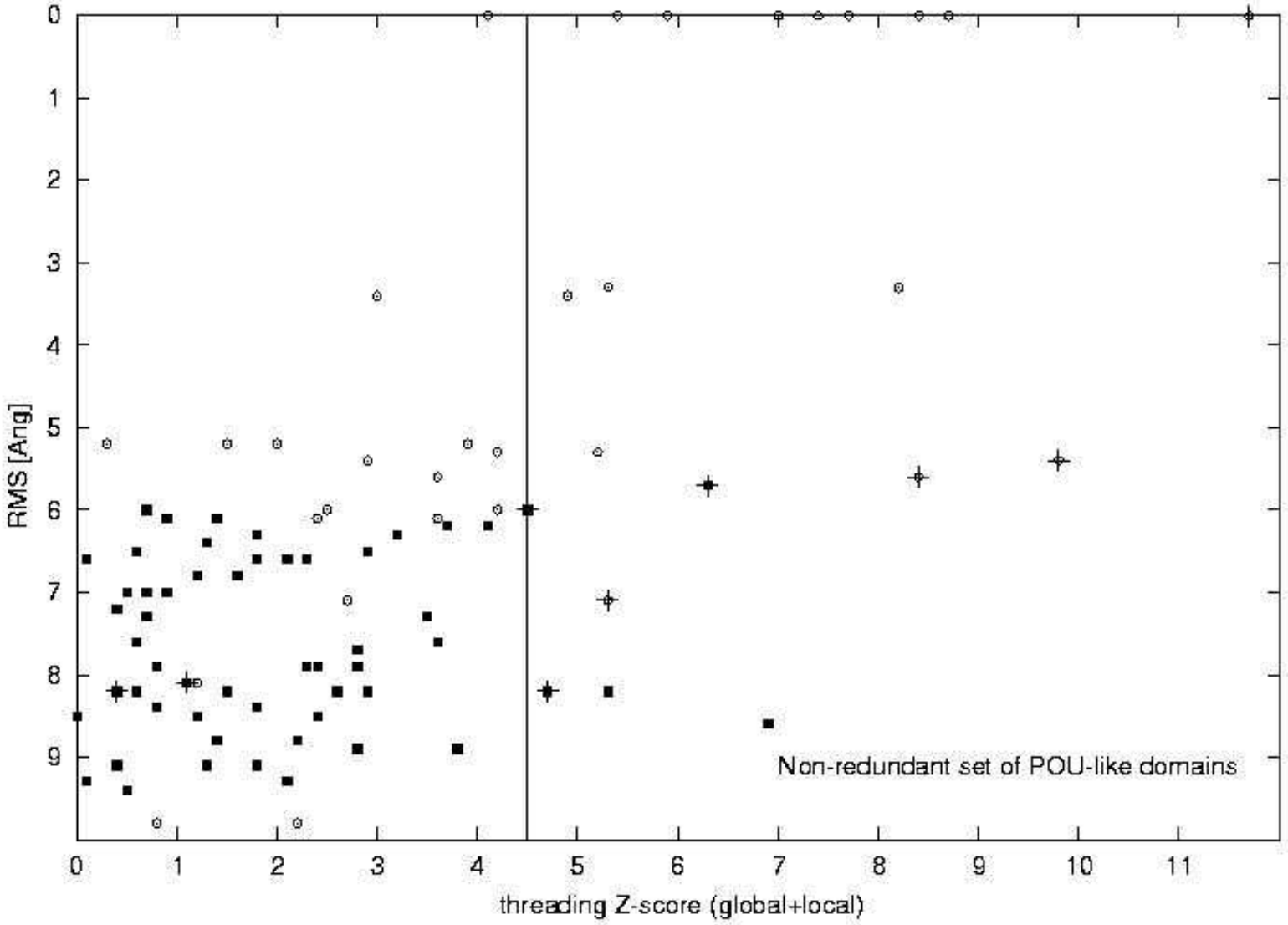
Need for statistical verification of predictions:

- Scoring according to an energy may be insufficient (good matches by similar length or composition)
- Z-score: a convenient measure of the strength of a match in terms of distribution of energies for random alignments

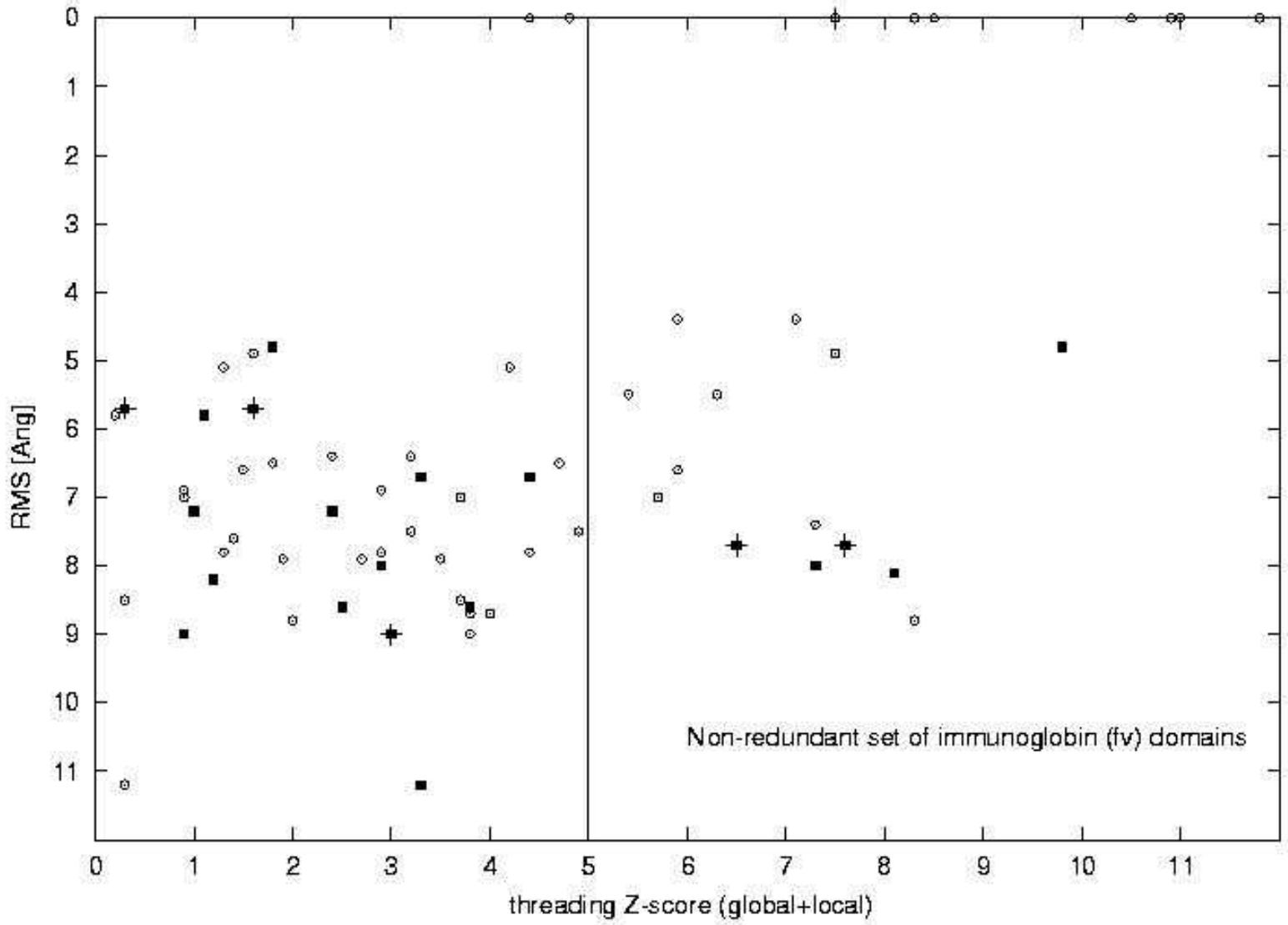
Joint Z-score (global and local threading) distribution:



Family recognition: POU-like domains



Family recognition: immunoglobins



Sample LOOPP Predictions

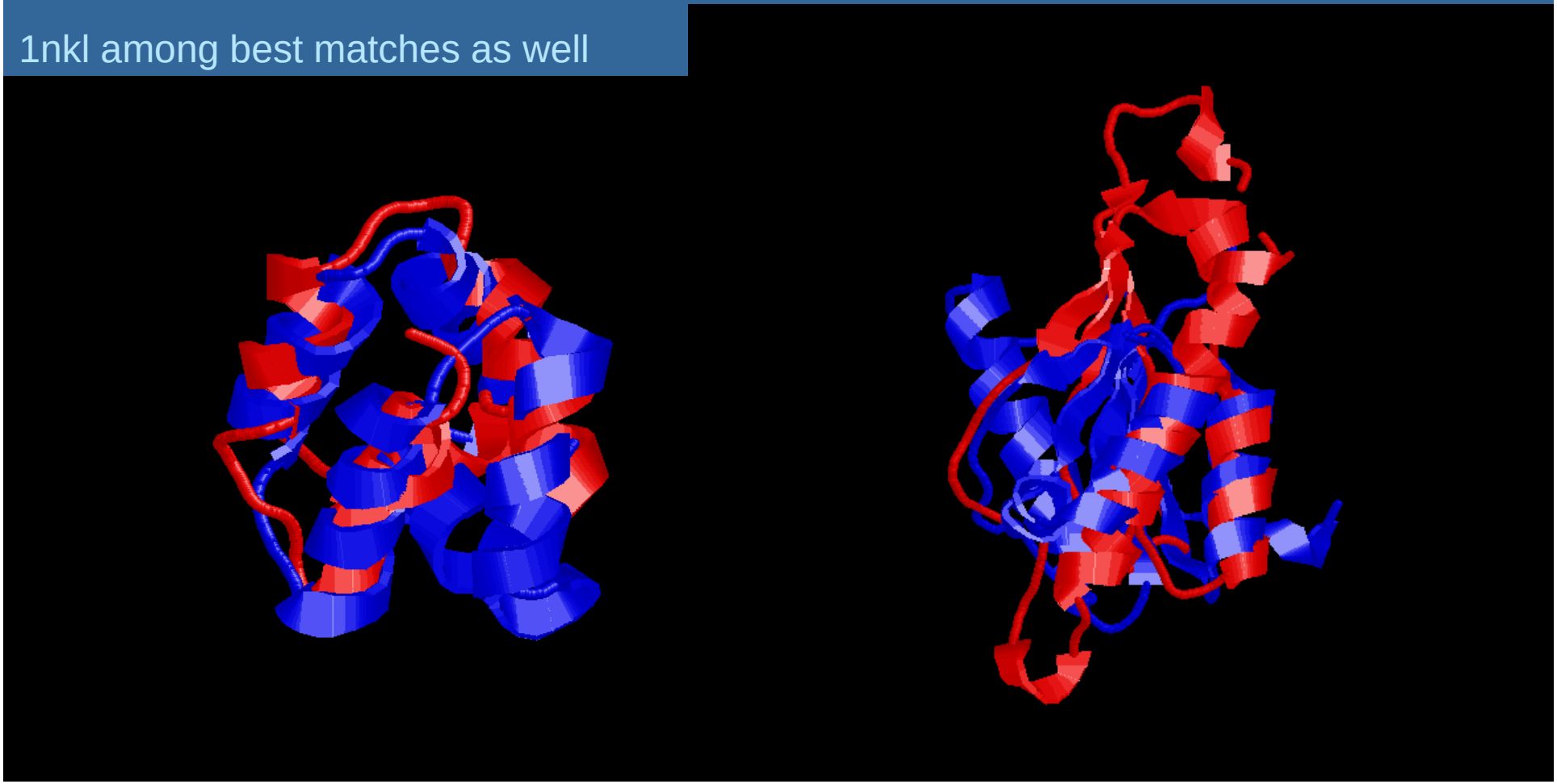
Predictions for difficult targets CAFASP & CASP:

T102 (70 res)

Model 1: 1bo9, 34 res with 2.5 A, 44 res with 3.1 A, 12th best (1st) model (M. Sippl), 1nkl among best matches as well

T116_2 (121 res)

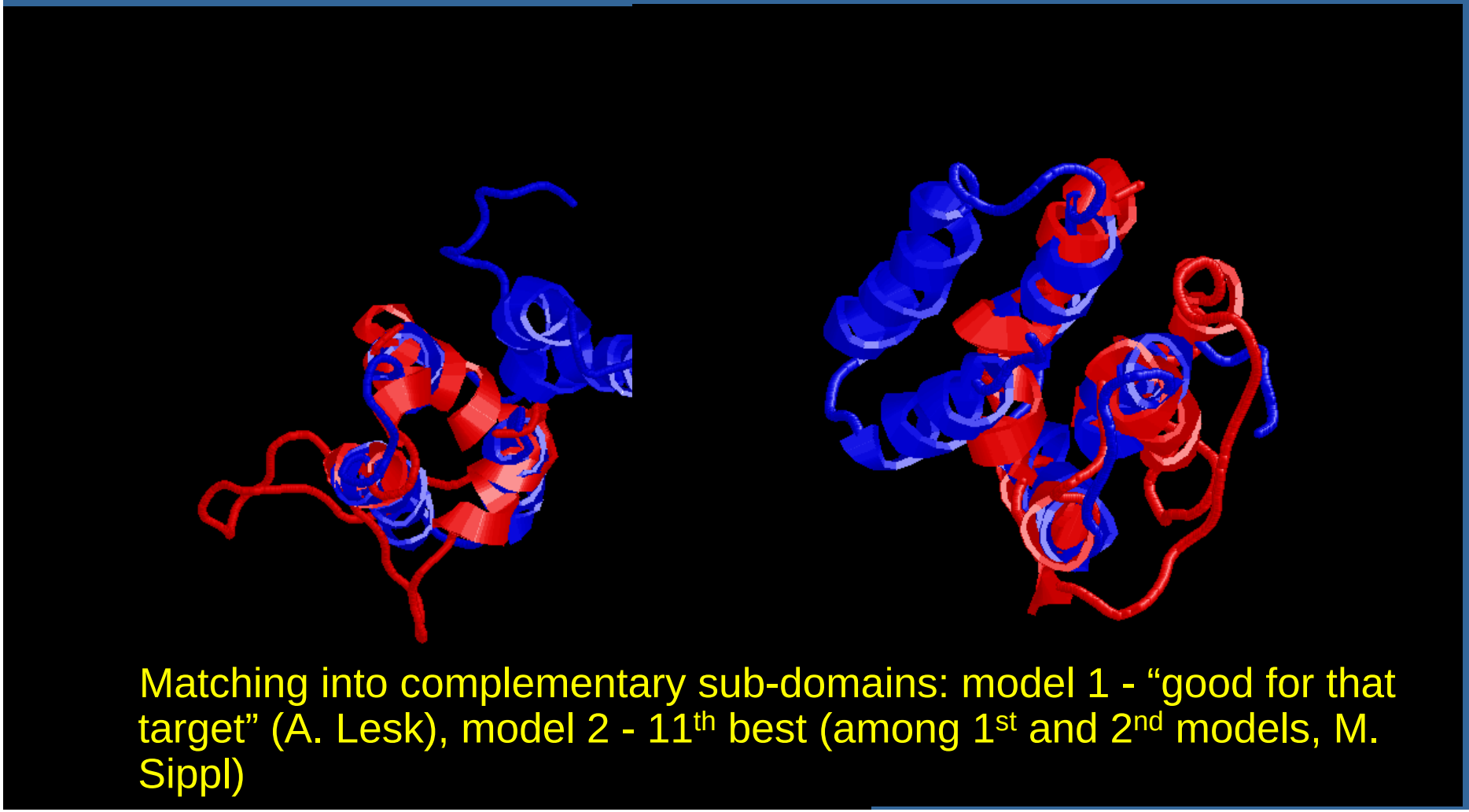
Model 1: 1a0cA, 50 res with 2.9 A, 2nd best (1st) model (M. Sippl)



predictions for difficult targets: T097 (104 res).

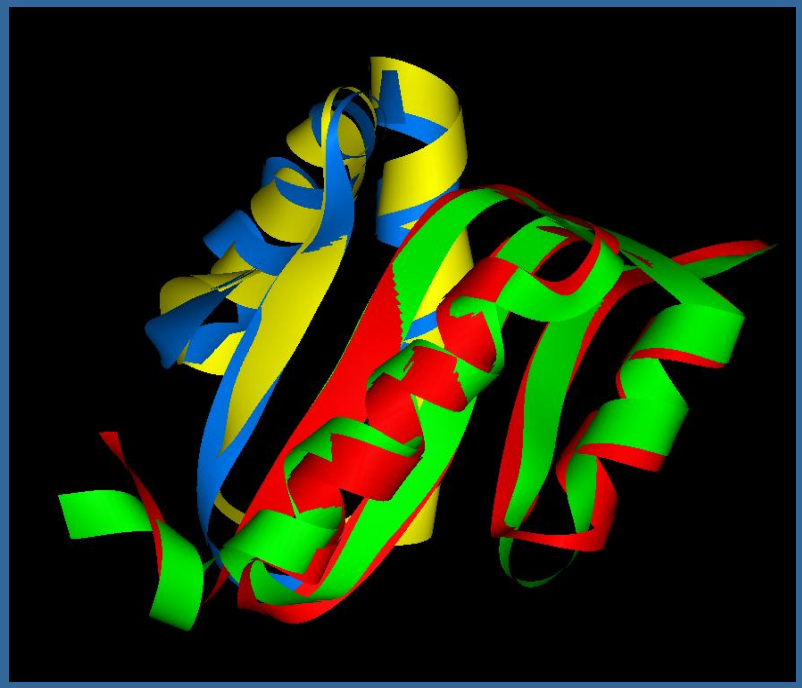
Model 1: 2hfh, 39 res with 3.3 A

Model 2: 3itr, 54 res with 3.2 A



Matching into complementary sub-domains: model 1 - “good for that target” (A. Lesk), model 2 - 11th best (among 1st and 2nd models, M. Sippl)

CASP prediction: Target T0280



Target	Best Loopp RMSD	Best Other RMSD	Is the best hit chosen?	RMSD of best hit	If best hit is not chosen, is one of the chosen hits true hits?
288	2.0	1.3	Best hit not chosen	1.01	Yes
290	0.53	0.48	Best hit chosen	0.47	
291	1.6	0.7	Best hit chosen	0.86	
292	3.1	2.9	Best hit not chosen	2.68	Yes
293	5.4	4.6	Best hit not chosen	2.42	Yes
294	2.7	2.1	Best hit chosen	1.55	
295	2.4	1.8	Best hit chosen	2.12	
297	4.3	2.8	Best hit chosen	2.37	
298	5.7	2.3	No good hit[#]	--	
302	1.7	1.5	No good hit	--	
303	2.8	2.2	Best hit chosen	2.16	
305	1.3	1.0	Best hit not chosen	1.17	Yes
308	2.0	1.4	Best hit not chosen	1.18	Yes

Sometimes we do really bad…

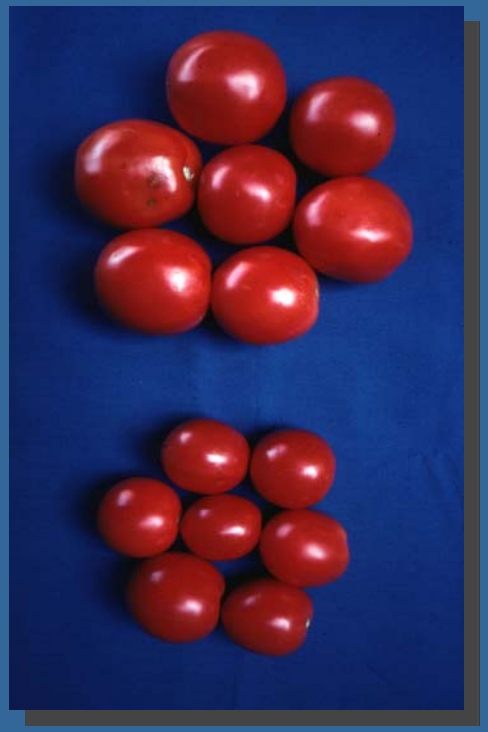
CASP7 Target	Best Loopp RMSD	Best other RMSD	Reason for Loopp going wrong
283	8.4	5.8	Hit present in DB but wrong parent Chosen
289	7.4	6.2	Hit present in DB but wrong parent Chosen
296	22.3	5.1	No true hit in database
299	15.7	5.0	Hit present in DB but wrong parent Chosen
300	11.0	1.2	Hit present in DB but wrong parent Chosen
301	22.4	8.0	No true hit in database
304	10.0	4.9	Hit present in DB but wrong parent Chosen
306	12.9	5.3	No true hit in database
307	13.0	6.6	Hit present in DB but wrong parent Chosen
309	12.0	7.0	No true hit in database

Structure prediction for a tomato fruit-weight protein

- ORFX gene, controlling the size of a tomato fruit, has been predicted to share structural similarity with human Ras p21 (work in collaboration with Tanksley's group, Cornell, Science 289,85-89(2000))

Phylogeny of Lycopersicon

Chromosome 2

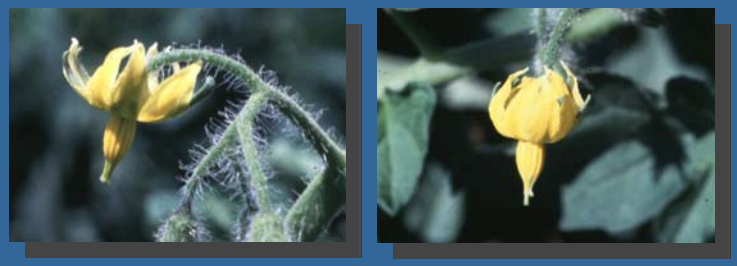


fw2.1, 2.2, 2.3

TG608
TG189
CT205
TG554
TG493
TG266
TG469
TG463
CT9
TG337
TG48
TG34
TG91
TG167
TG151
CT59
TG154

stuffer

ovate



se2.1

Human Ras p21

- Molecular switch based on GTP hydrolysis
- Cellular growth control and cancer
- Ras oncogene: single point mutations at positions Gly12 or Gln61

LOOPP prediction for tomato ORFX

ORFX is predicted to have a structure similar to G–protein:

- Global and local alignments of ORFX sequence to ras 6q21A structure are consistent and indicate very good matching.
  Other good local alignments are to domains of similar topology.

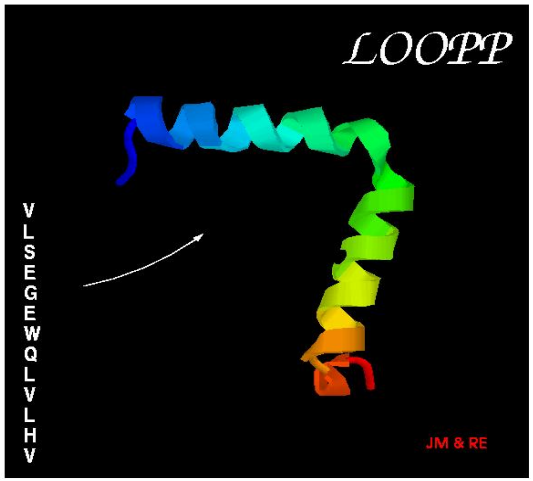
- Statistical significance of both global and local alignments is high – Z–score of 3.2 and 4.0, respectively. We never observed false positives with such Z–scores.

- Hydrophobicity profile indicates that ORFX is a soluble protein.

- Independent secondary structure predictions indicate alpha/beta type with positions of loops consistent with that of ras (PsiPred, PHD, Predator).

- Plausible counterparts of the crucial Switch I and Switch II loops are conserved in the multiple alignments to ORFX homologs.

- Ras active site fingerprint (TFGQ instead of TAGQ) is found in Switch II loop. Ras metal coordination sites and nucleotide binding sites are found in the predicted ORFX counterparts of P–loop, Switch I and Loop 5.



Yet bigger tomatoes …

Some references to LOOPP

- Dror Tobi, Gil Shafran, Nathan Linial and Ron Elber, "On the design and analysis of protein folding potentials", "Proteins, Structure Function and Genetics", 40, 71-85 (2000).
- Dror Tobi and Ron Elber, "Distance dependent, pair potential for protein folding: Results from linear optimization", Proteins, Structure Function and Genetics, 41, 40-16 (2000).
- Jaroslaw Meller and Ron Elber, "Linear Optimization and a double Statistical Filter for protein threading protocols", Proteins, Structure, Function and Genetics, 45,241-261(2001)
- Jian Qiu and Ron Elber, "Atomically detailed potentials to recognize native and approximate protein structures", Proteins, Structure, Function, and Bioinformatics, 61:44-55,2005