Fold recognition by threading

Learning Observing Outputting Protein Patterns

T. Galor

Computational Molecular Biology Lab Ron Elber's group

Computer Science, Cornell

Motivation-general

The knowledge of molecular structure of a protein is prerequisite for understanding its function:

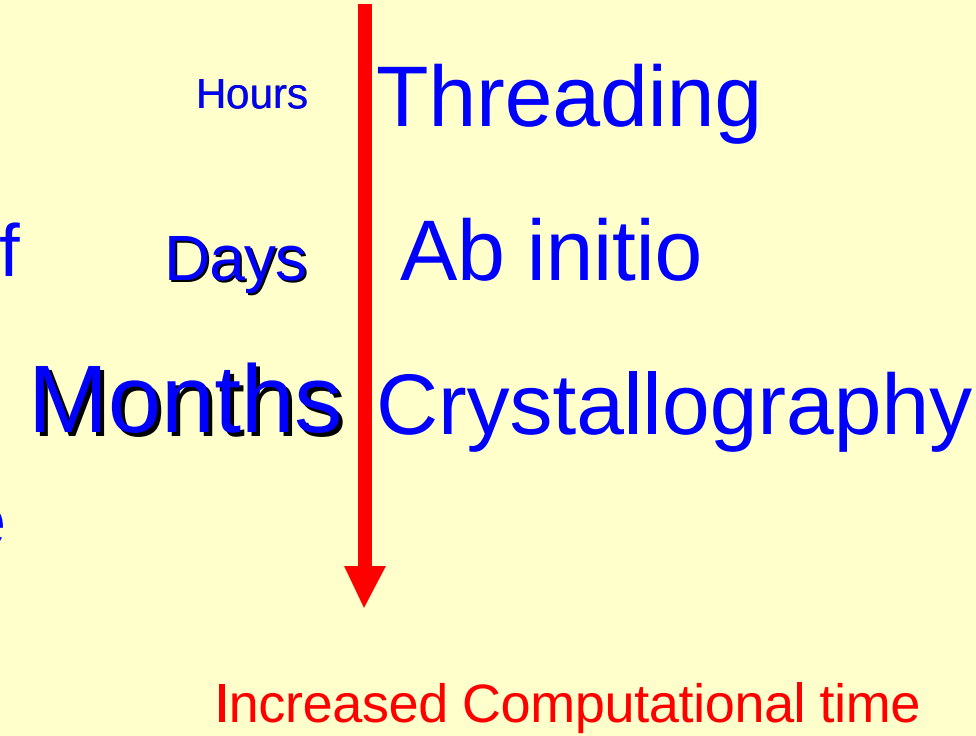
what are the active site residues?

Where are the protein-protein interaction sites?

There are many methods which address different aspects of protein fold and each one has its own limitation.

Methods for fold prediction

- Crystallography /NMR.
- Ab initio prediction of protein structure is not currently a general and reliable method.
- Threading



Accuracy of prediction

- Experiment: crystallography or NMR most accurate.
- Homology modeling quite accurate (needs templates)
- Threading reasonable accurate (needs templates)
- Ab-initio: least accurate (most general computational approach)

Why do we need computational methods for fold prediction?

- The rate of new protein sequences is growing exponentially relative to the rate of protein structures being solved by experimental methods.

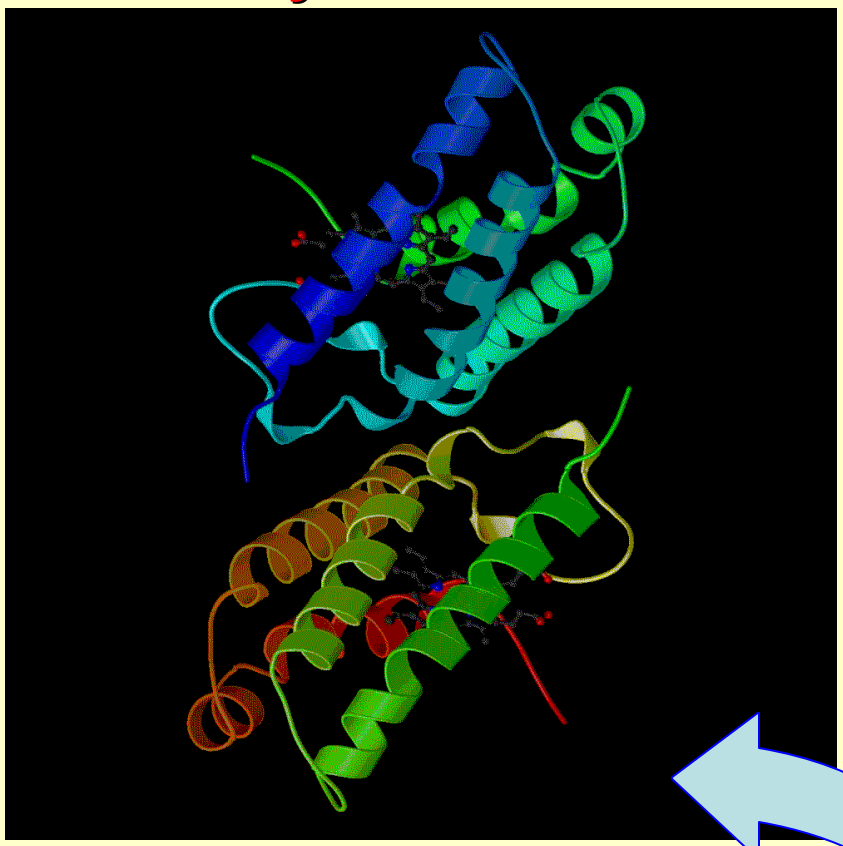
- An approximate model is a guiding tool in experiments.

Fold predication by threading when sequence identity is low

Evaluates how well amino acid sequence fits into known three-dimensional (3D) protein structure.

Test fitness of the probe sequence to structures from existing library.

Choose the best most significant fit of sequence/structure.



Protein name: 1I4Z_D
MGFPIPDPYV … KGKI

Why not use sequence alignment to detect remote homology?

- Protein structures can share similar folds despite no significant sequence similarity.

- Myoglobin 1mba and leghemoglobin 1bin:A.

http://cl.sdsc.edu/ce.html

Combinatorial Extension :

Developers :Ilya N. Shindyalov, Philip E. Bourne.

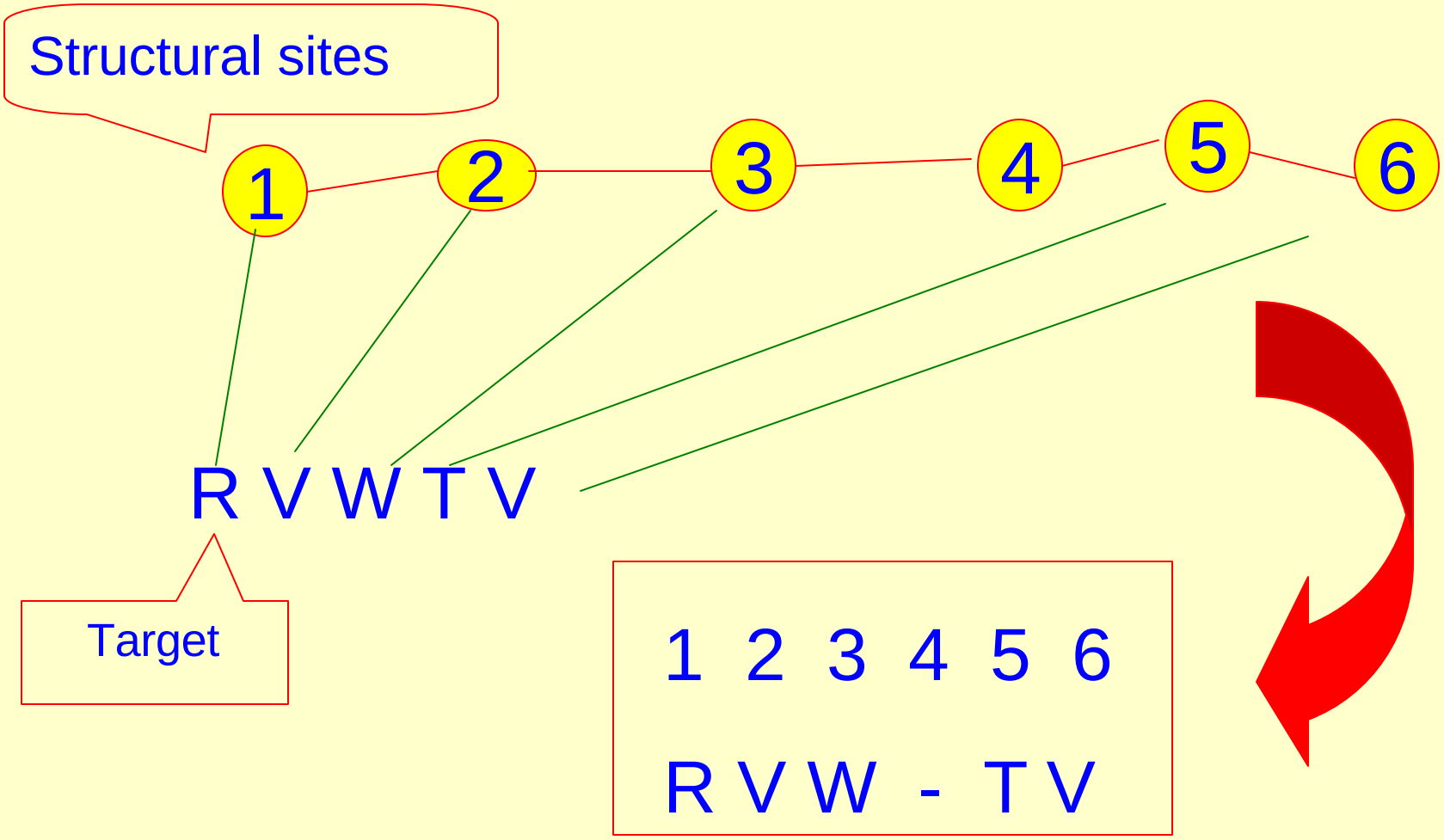
CE is a method for calculating pair wise structure alignments. CE aligns two polypeptide chains using characteristics of their local geometry as defined by vectors between C alpha positions.

Threading

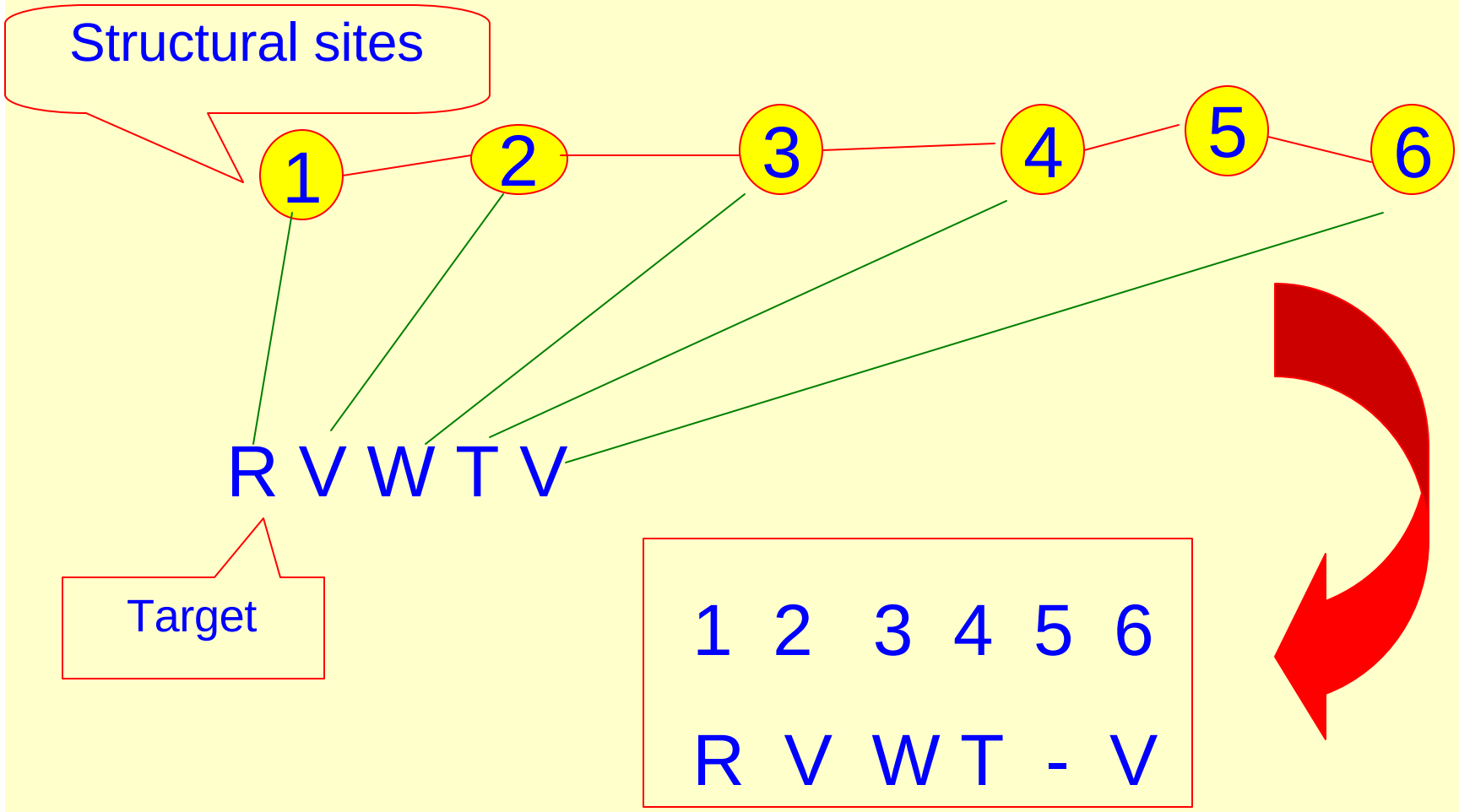
- We need to find a way to detect remote homologues (distant evolutionary relationship).
- Use structural information for each site in addition to sequence might help detect similarities unobserved by sequence alignment.

- Secondary structure (helix sheet loop, alanine like being in a helix)
- Contacts between sites (hydrophobic residues like site with many contacts)
- Surface exposure (buried exposed)

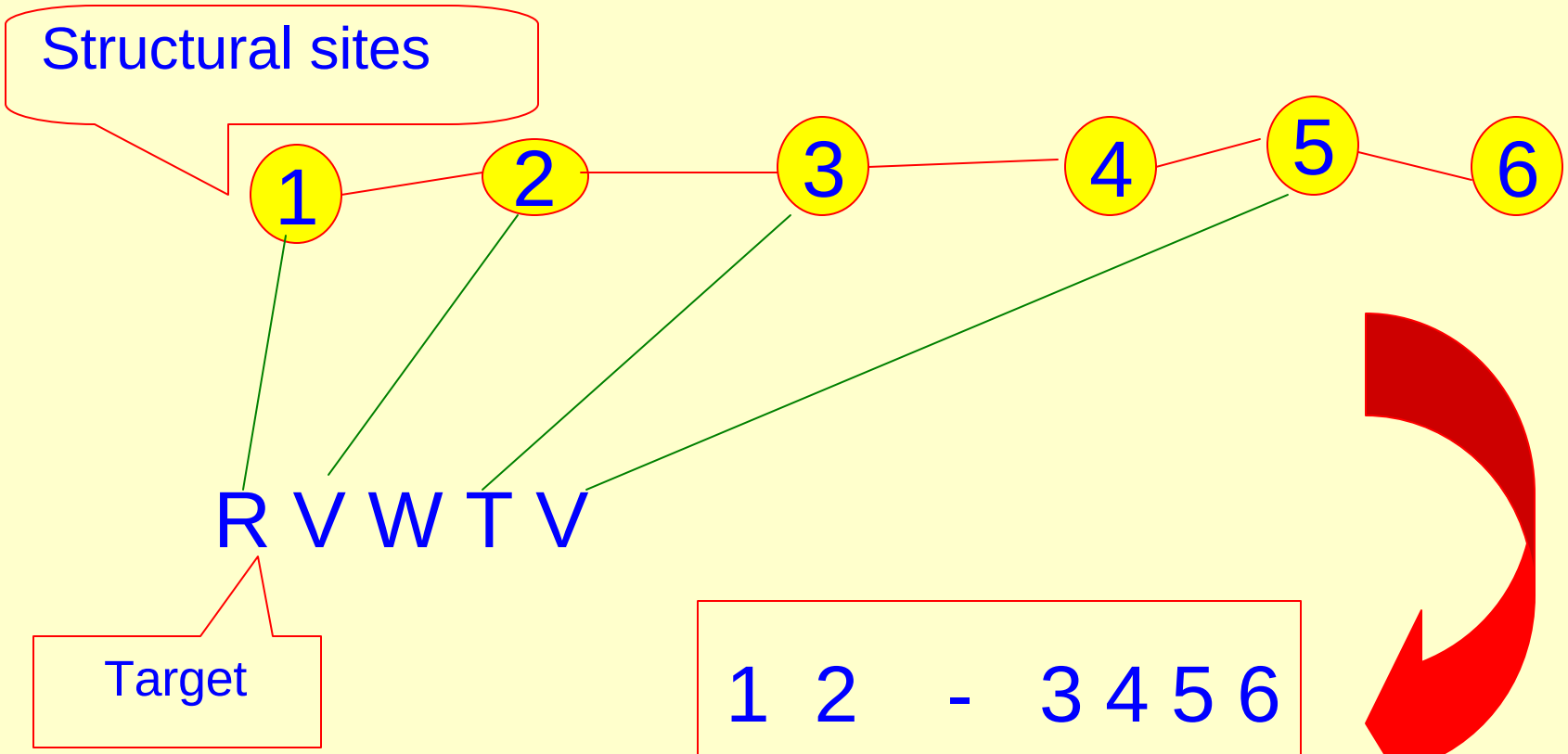
Example: alignment 1



Example: alignment 2



Example: alignment 3



The need to score an alignment
The Energy

1	2	3	4	5	6
R	V	W	-	T	V

$$E(R,1)+E(V,2)+E(W,3)+E(-,4)+E(T,5)+E(V,6)=E_1$$

1	2	3	4	5	6
R	V	W	T	-	V

$$E(R,1)+E(V,2)+E(W,3)+E(T,4)+E(-,5)+E(V,6)=E_2$$

How do we find the best alignments?

- The one with the lowest energy is chosen.
- Energy = 0 is the energy of the unfolded state.

How many alignment can we produce (m amino acids & n sites)?

In general : $\binom{m+n}{m} \approx \frac{2^n}{\sqrt{pn}}$. In the example $\binom{5+6}{5} = 420$

Scanning all possible alignments is time consuming (NP problem)

we need an efficient algorithm to find the optimal alignment without testing all possibilities.

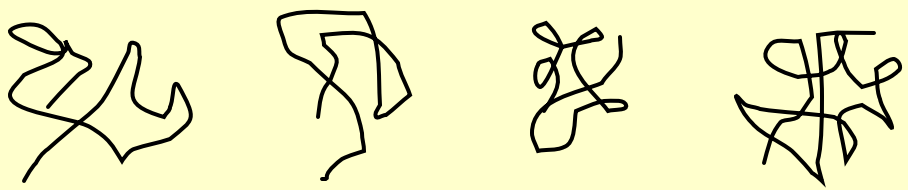
Threading general

query

A-R-W-S-S-H-

Define a potential

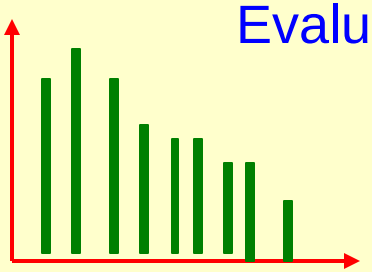
MODEL – define environment



Fold library

Align target

Using a potential



Simplifications

- Amino acids of different types have different preference for alternative structural environments. (alpha, beta, buried, loop).

- A single point representation of an amino acids determines overall protein fold and not atomic coordinates (need homology modeling to obtain atomic level structure).

Why are the simplifications sound?

- protein structures are known to be robust to most single point mutation.
- In addition, protein structures must be sufficiently stable to tolerate a few amino acid at unfavorable positions.

  (hydrophobic amino acids on the surface to bind hydrophobic substrate).

The number of folds in a fold library is restricted

- So far ~500 distinct folds are known in the protein data bank.
- In the protein databank there are 18359 structures: Many repeats
- We need to examine a small number of structures; can be done rapidly.

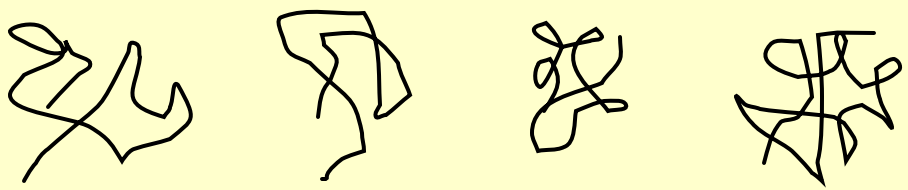
Threading

query

A-R-W-S-S-H-

Define a potential

MODEL – define environment

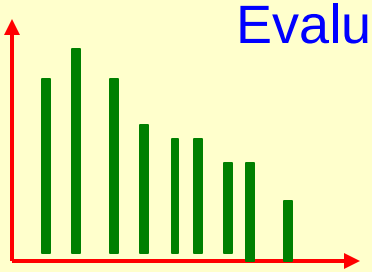


Fold library

Align target

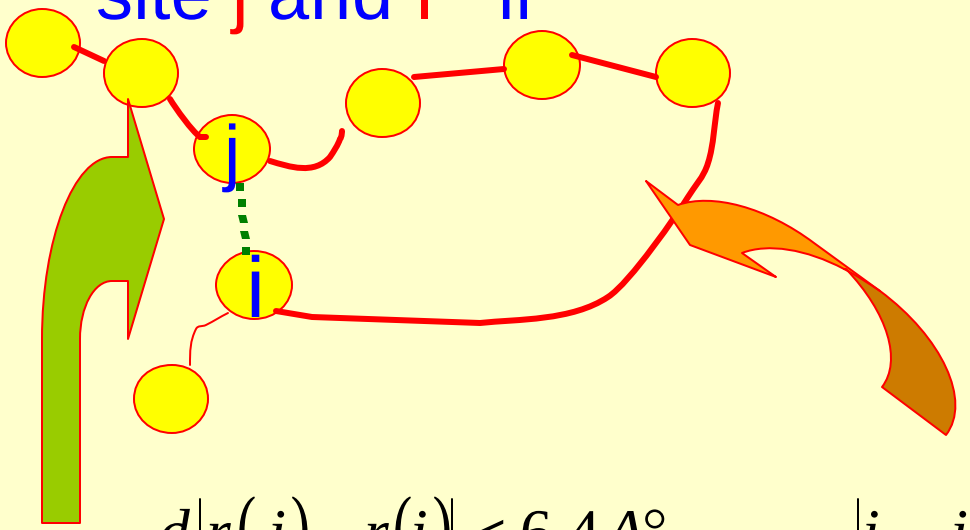
Using a potential

Evaluation



Energy is given as a sum of contact contributions
How should we define a contact?

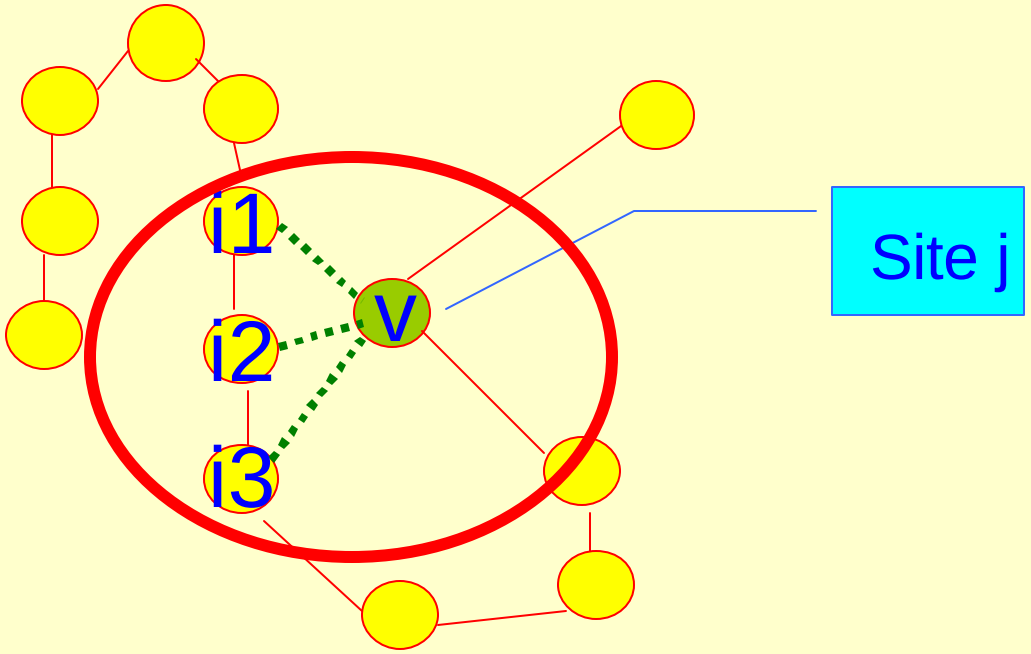
- We define a contact between site j and I if



$$d\left|r(j)-r(i)\right|<6.4A^{\circ} \qquad \left|i-j\right|\geq 4$$

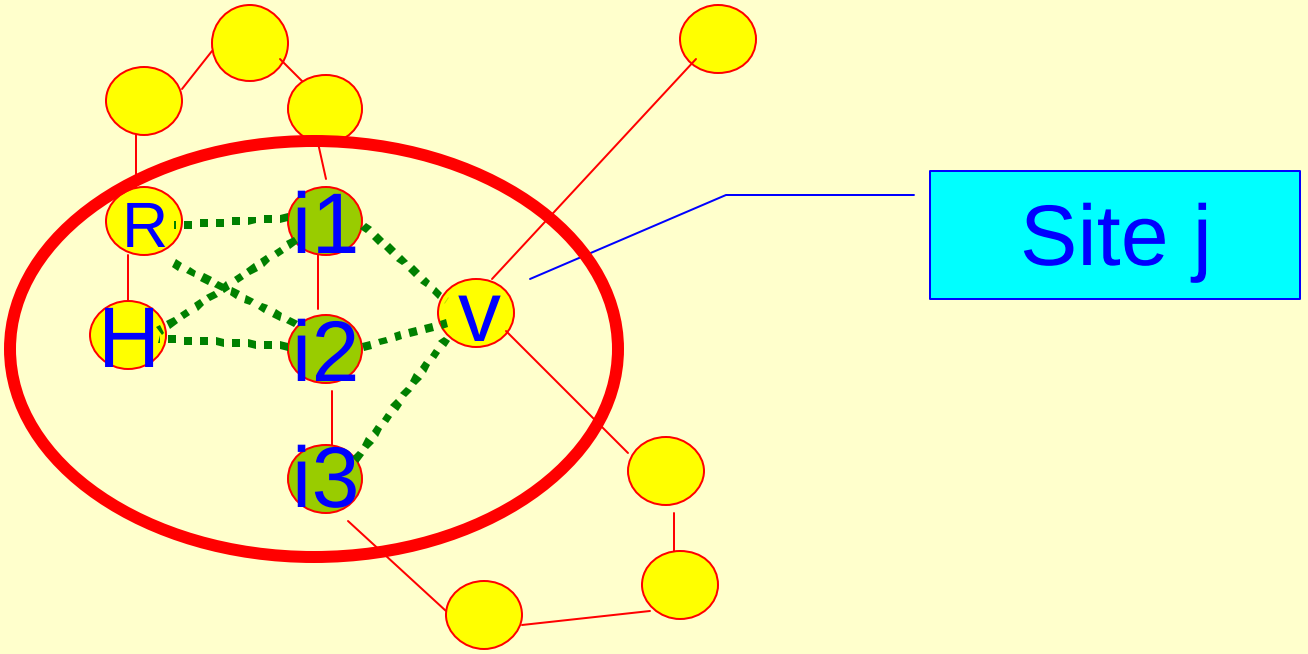
What is the structural environment of LOOPP?

- For each site we define a first contact shell



What is a structural environment in LOOPP?

- Then the second shell .



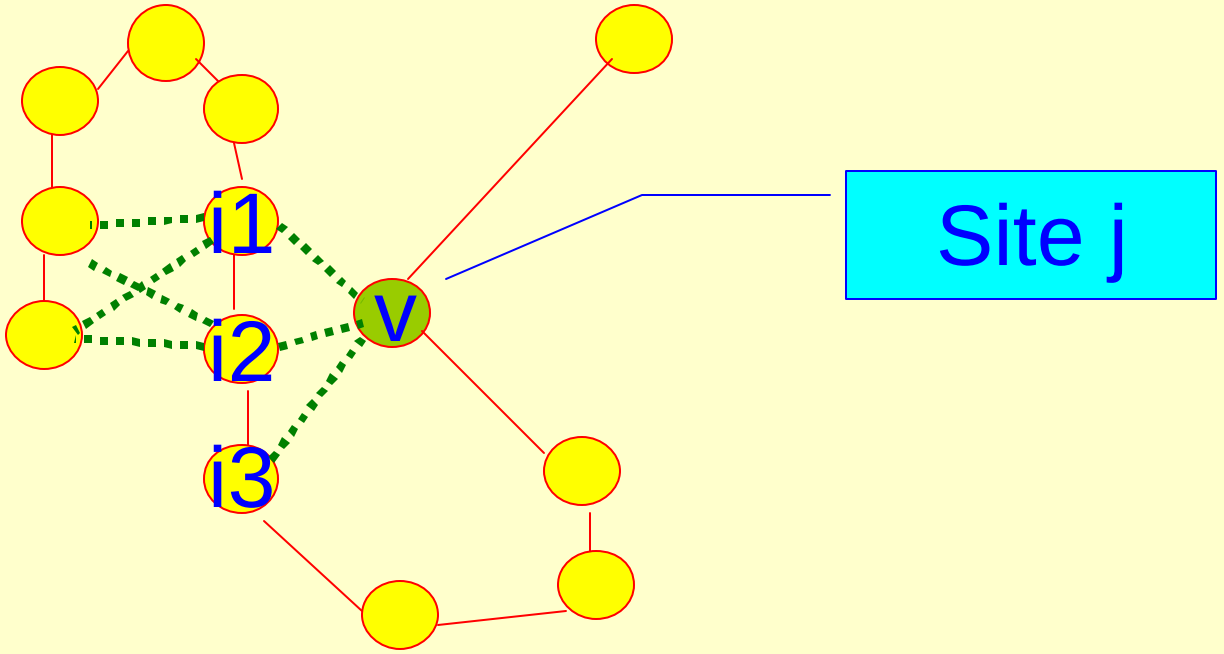
Threading Onion Model 2-has 16 different environments

	s1=1,2	s1=3,4, 5,6	s1>6
s2=1,2	x_1	x_2	x_3
s2=3,4	x_4	x_5	x_6
s2=5,6	x_7	x_8	x_9
s2=7,8	x_{10}	x_{11}	x_{12}
s2>9	x_{13}	x_{14}	x_{15}

- S1: First shell
- S2: Second shell
- Xi : course grained environment

$$l(1,1) = l(2,2) = l(1,2) = l(2,1) = x_1 \quad l(0,0) = x_{16}$$

Example:
look again at site j



The energy at site j is

$$E(j) = U(V, X_5) + U(V, X_5) + U(V, X_2)$$

$$E(j) = U(V, X_5) \cdot 2 + U(V, X_2) \cdot 1$$

This is a profile method it does not depend on the AA types of the first and second contact shell.

In the general case

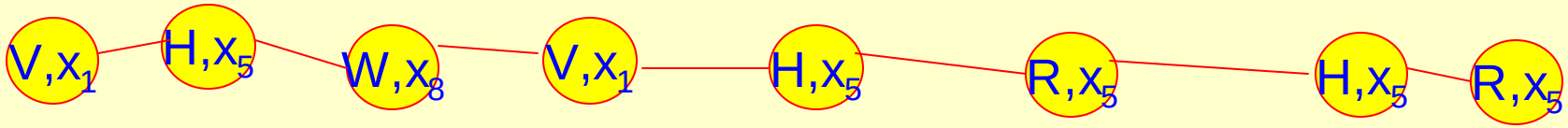
- The total energy of the sequence $S=(a_1,\ldots,a_n)$ into the structure $X=(x(1,_{i1}),\ldots,x(n,_{in}))$ is given by

$$E(S \rightarrow X, U) = \sum_{j=1}^{n} E(j)$$

$$E(S \rightarrow X, U) = \sum_{j=1}^{n} \sum_{i} U(a_j, X_{j_i})$$

The energy can be written as a scalar product

Example:



$$E = 2 \cdot U(V, x_1) + 3 \cdot U(H, x_5) + 2 \cdot U(R, x_5) + U(W, x_8)$$

General

$$E(S \to X) = \vec{U} \bullet \vec{n}$$

Explicitly

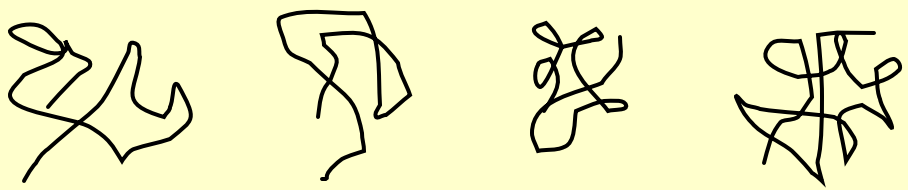
$$E(S \to X) = \sum_{j=1}^{20} \sum_{k=1}^{15} U(a_j, x_k) \cdot n(a_j x_k)$$

Threading

query A-R-W-S-S-H-

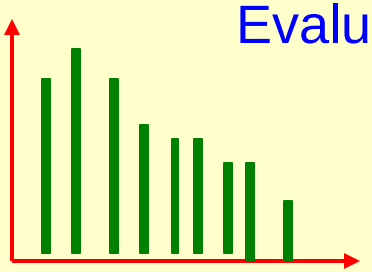
Define a potential

MODEL – define environment



Fold library

Align target, using the potential to score

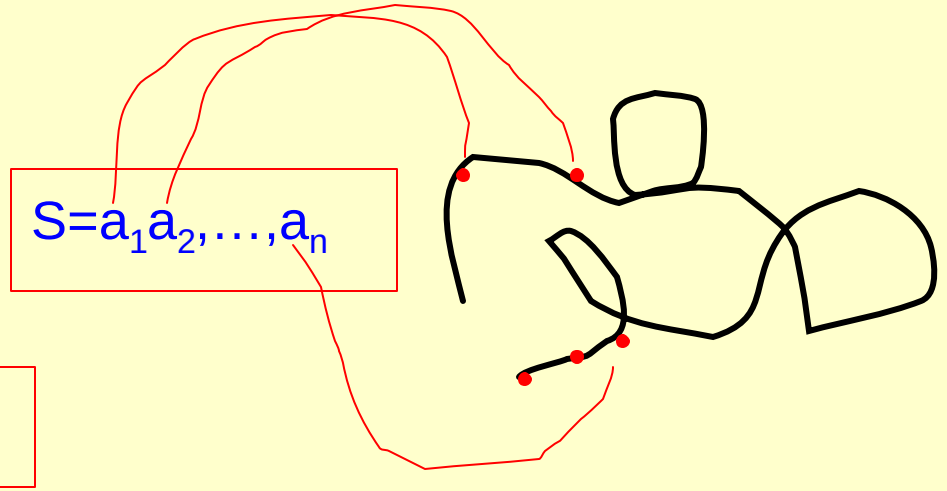


How can we find the potential U

threading Gapless

Assign a sequence to

A structure S->X



Native $S_n \rightarrow X_j$

Decoy $S_n \rightarrow X_n$

We try to find the potential U that satisfies the constrain:

$$E(S_n \rightarrow X_j, U) - E(S_n \rightarrow X_n, U) > 0 \quad , j \neq n$$

In what way decoys are different from native

frequencies of hydrophobic, polar.

Note for decoys the

hyd/pol~(1:1)

	Native hyd/pol	decoy hyd/pol
x1	1.59/3.5	5.48/5.85
x2	2.99/6.04	6.15/6.54
x3	0.15/0.26	0.17/0.18
x4	2.88/3.37	4.6/4.91
x5	13.01/11.08	12.9/13.7
x6	1.88/1.35	1.12/1.18
x7	1.81/0.96	1.54/1.64
x8	20.9/7.4	10.7/11.3

Deriving the potential

- The inequality can be simplified

$$\Delta E = U \bullet \Delta n > 0$$

a_j x

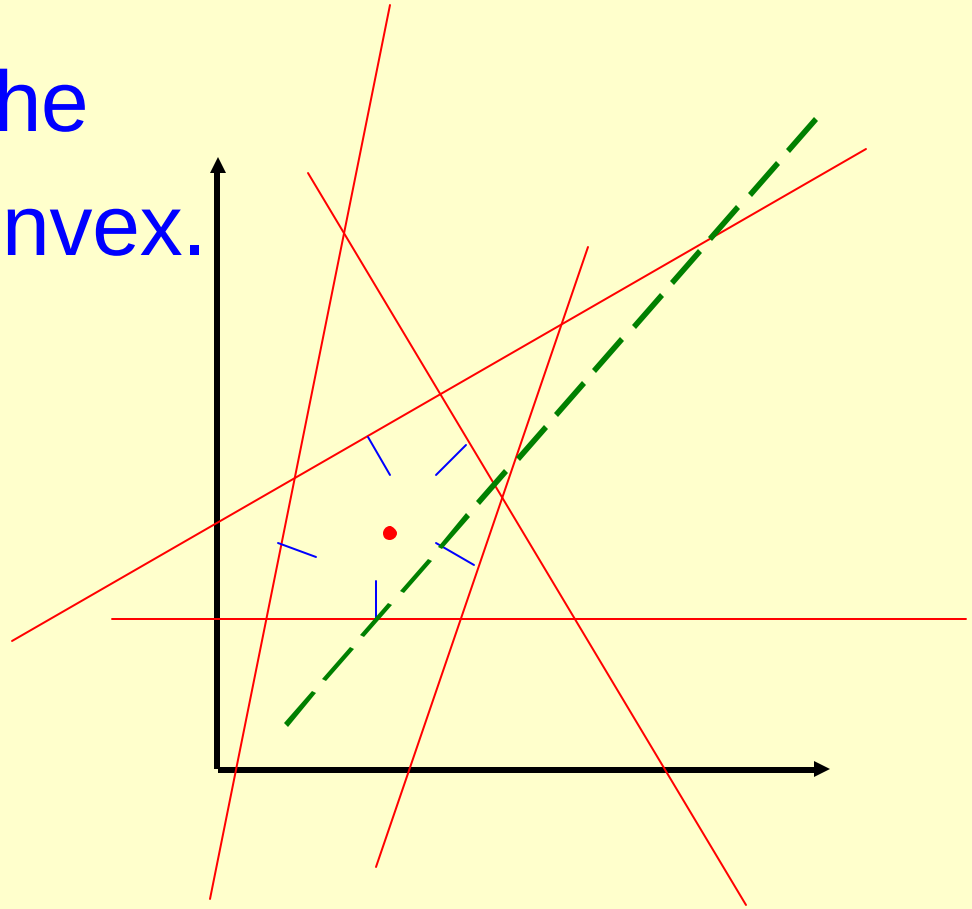
Where U is a set of 300=(20x15)

unknown potential parameters.

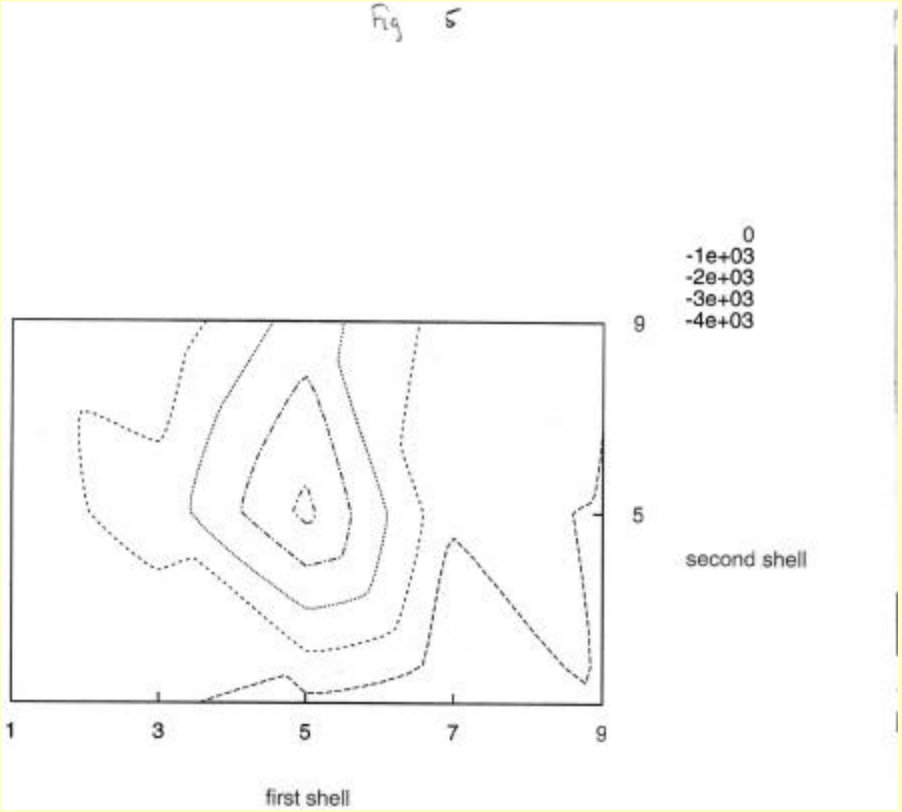
To find U, we solve a set of inequalities which Include millions of linear constrains. Standard Interior point algorithm finds the solution

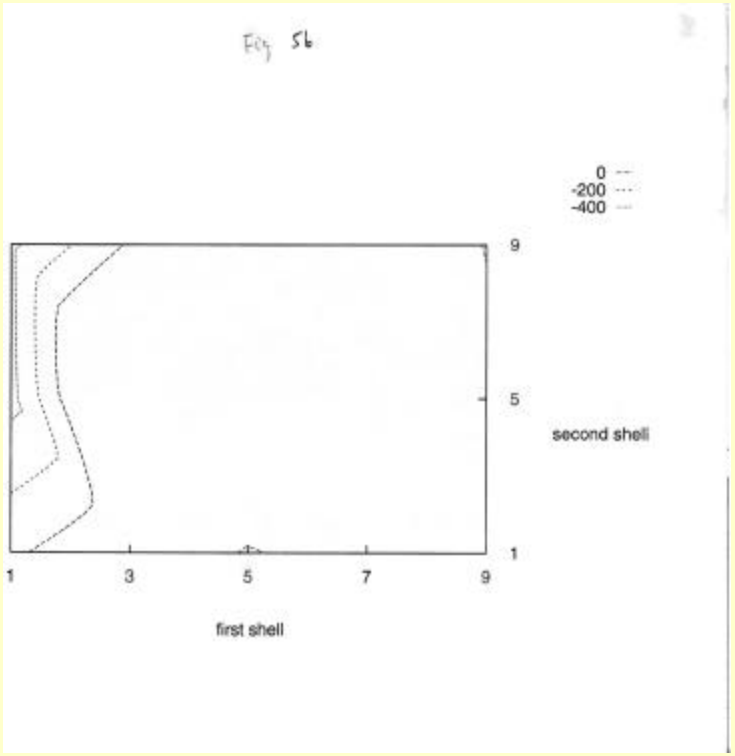
Which solution are we looking for?

The solution U is in the middle of the poly convex.



Potential values for val and lys as a function of the environment





Val, prefers 5 neighbors

lys

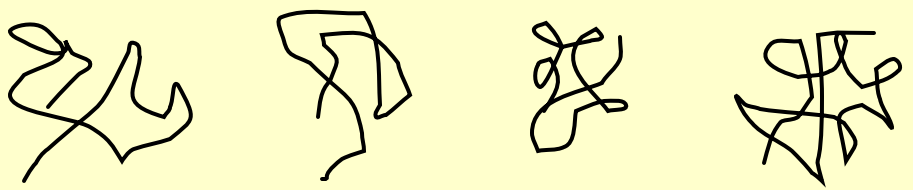
Threading

query

A-R-W-S-S-H-

Define a potential

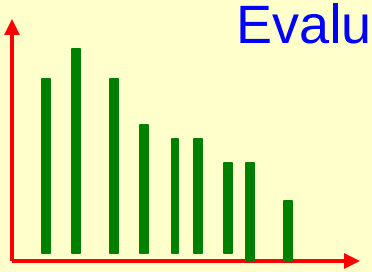
MODEL – define environment



Fold library

Align target, using the potential to score

Evaluation



LOCAL vs. GLOBAL ALIGNMENT

Local alignment - find similar domains

A V | R T H Y S A | V V T A R

k | R T - - S A | g w s

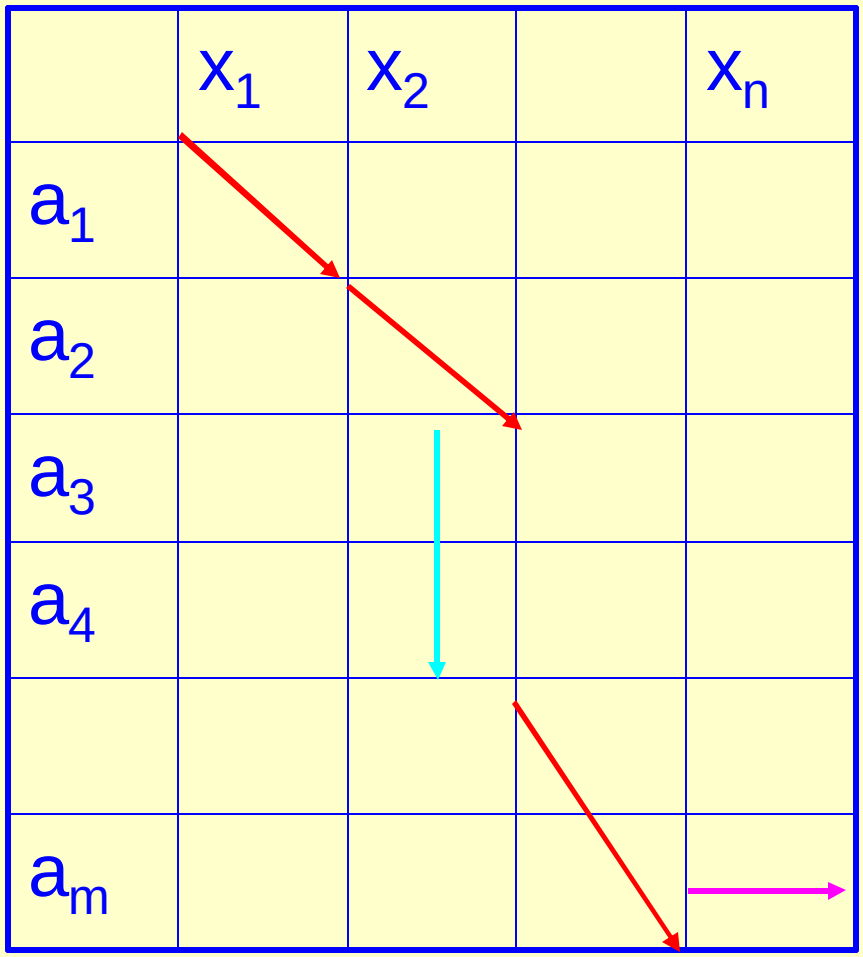
Global alignment – find similar proteins

A R R G I K L W N M S A V V

A V V - - K L W - M S - V -

How do we get the energy for an alignment

- Loopp uses Smith Waterman algorithm (DP) in order to align a sequence into a structure .
- We find the path with the minimal energy.
- Negative energy are good as they are less than the unfolded state=0.



Energy unfortunately is not is not a good measure (Baker: 65 sets)

potential	aver. pos. Ene.	# correct
TE13	27	40
MJ	150	23
HL	163	15
SK	158	11
BT	148	15
THOM2	106	15

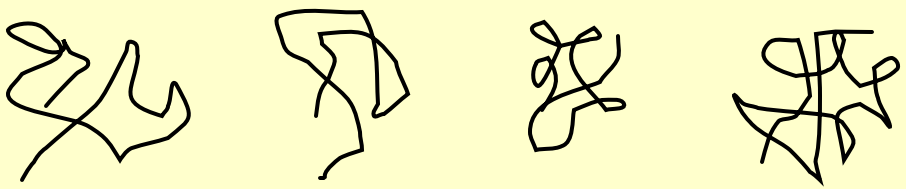
Threading

query

A-R-W-S-S-H-

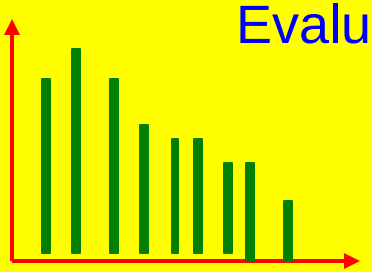
Define a potential

MODEL – define environment



Fold library

Align target, using the potential to score



A statistical measure for local fitness

We want to estimate how significant is our score (energy) compared to a random sequence that cannot be significant.

The Z scores

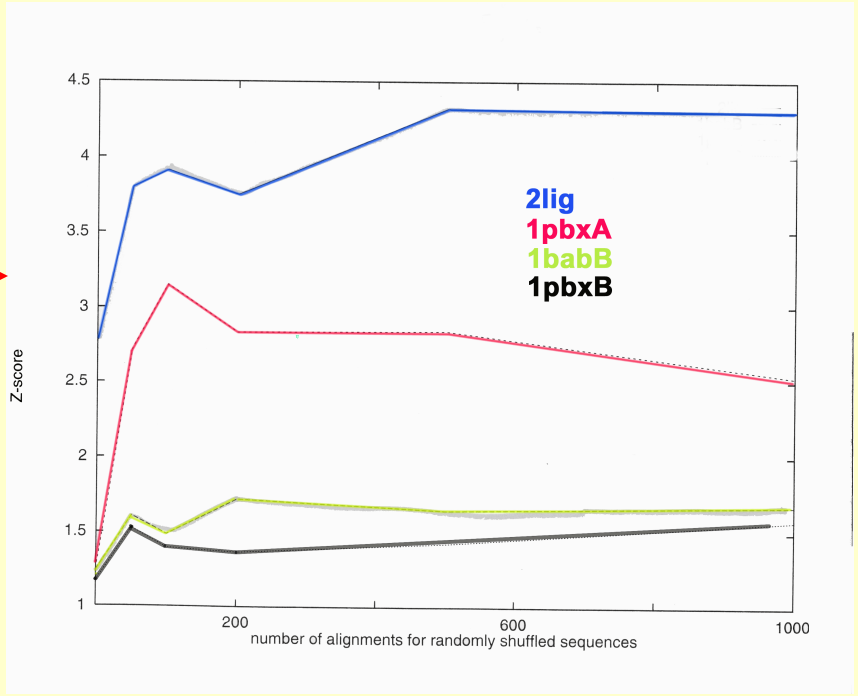
- A random sequence is defined by shuffling amino acids. The shuffled sequence is aligned to the structure to compute the (random) energy .

$$\left\langle E\left(S \to X, U\right)\right\rangle = \frac{1}{M}\sum_{i=1}^{M} E\left(\text{Permute}(S) \to X, U\right)$$

$$Z = \frac{\left\langle E\right\rangle - E}{\sqrt{\left\langle E^2\right\rangle - \left\langle E\right\rangle^2}}$$

Calculating Z score is expensive

- How many shuffles do we need to get convergence ?
- Four examples
- When the signal is not strong, it is important to fully converge the value of z score.



The false positive Z score

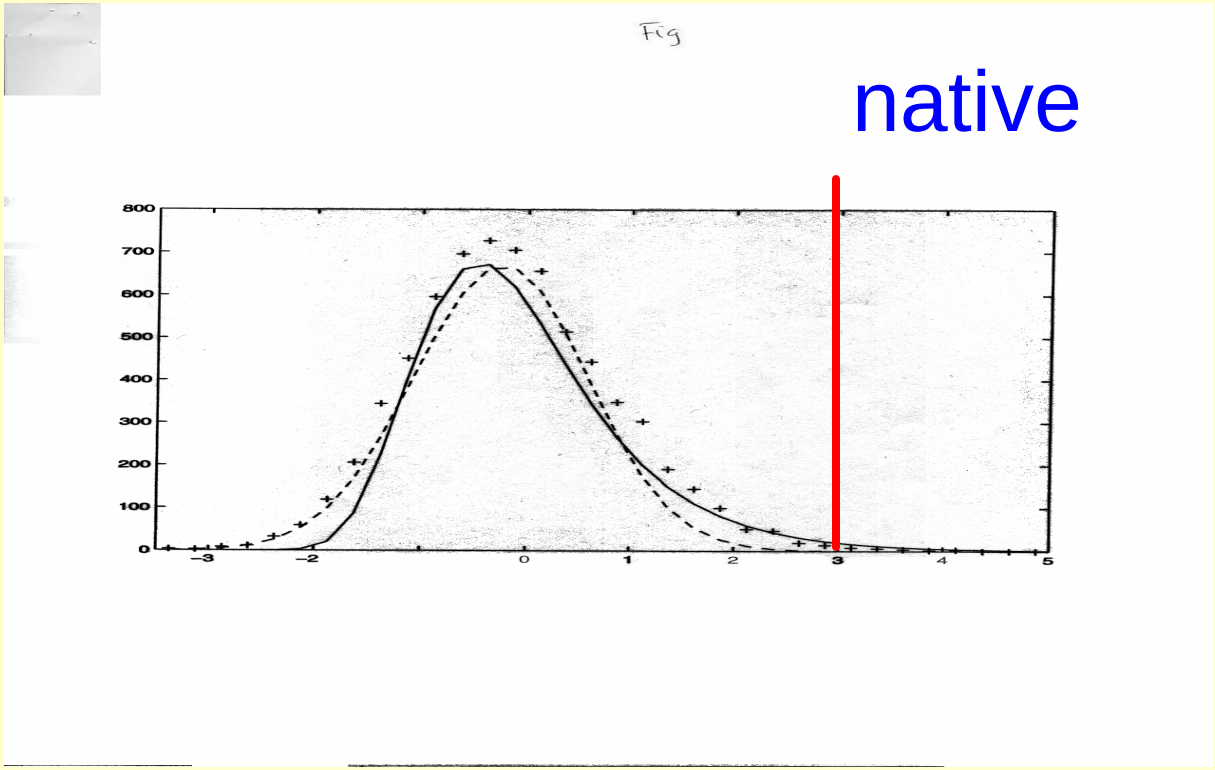
Local alignment has high number of false positive. What is a good threshold?

Let us look on the probability of false positive.

$$\Pr\{z_{fp} > z\}$$

The probability of a false positive

- An upper bound for the probability tail of false positive is the extreme value distribution



How many proteins are annotated incorrectly ?

For an alignment it is

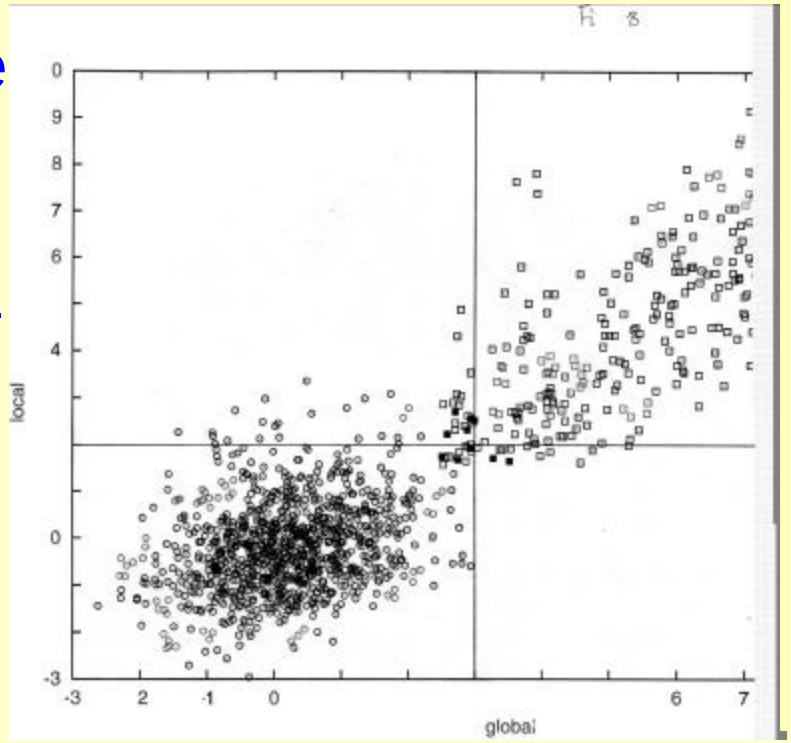
$$\Pr\{z_{fp} > 3\} < 0.003$$

- When searching a large data base the expected number of false positive can be large.

$$N \Pr\{z_{fp} > z\}$$

How can we decrease the probability for false positive?

- If we increase the threshold for local Z score the number of false positive will decrease but we may miss hits of low Z score signal.
- A cutoff of $Z_{tg}>3$ $Z_{tl}>2$ are sufficient to eliminate most false prediction.



■ ○ False positives

How we define a hit in loopp?

- Local sequence alignment (Blosom50 + structural gap penalty) Z_s>8.

OR

- Local alignment with Thom2 potential Z_{lt}>2.0.
- Global alignment with Thom2 potential Z_{gt}>3.0.

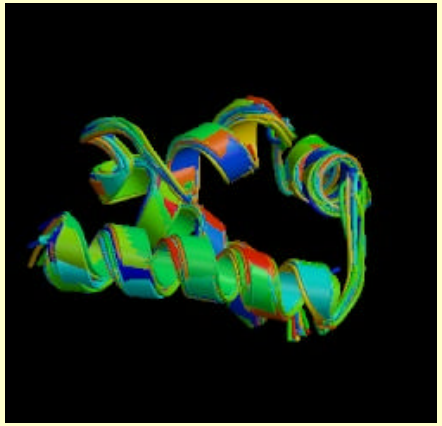
Results

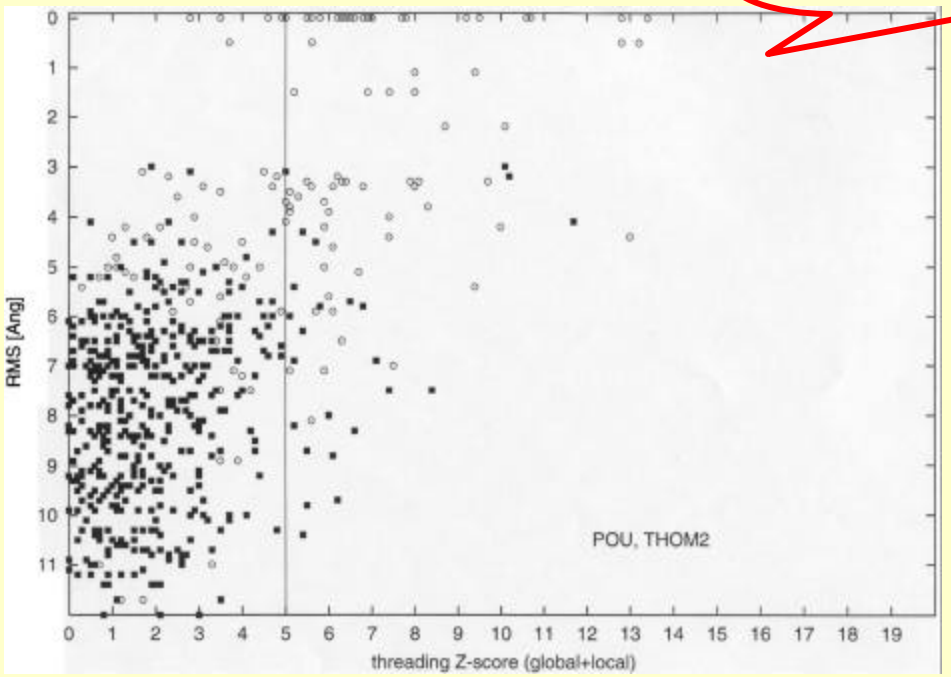
- Native structure are found with high probability.
- Thom2 identifies related structure with very high confidence only if their RMS distance is not larger then 3A°.
- Thom2 identifies most of the related structure of 3<RMS<5A°

Prediction for families

Name native	native pos.	thread	esq.	length	family
121p	1	9/10	6/10	166	ras
2blg	1	6/8	2/8	162	lactoglobin
2gsq	1	6/9	3/9	202	glutathione
1fynA	3	3/9	7/9	62	phosphotransferase
2cxbA	1	3/7	2/7	124	cytochrome
1meyC	1	3/5	4/5	86	Zinc finger

POU-DNA binding family

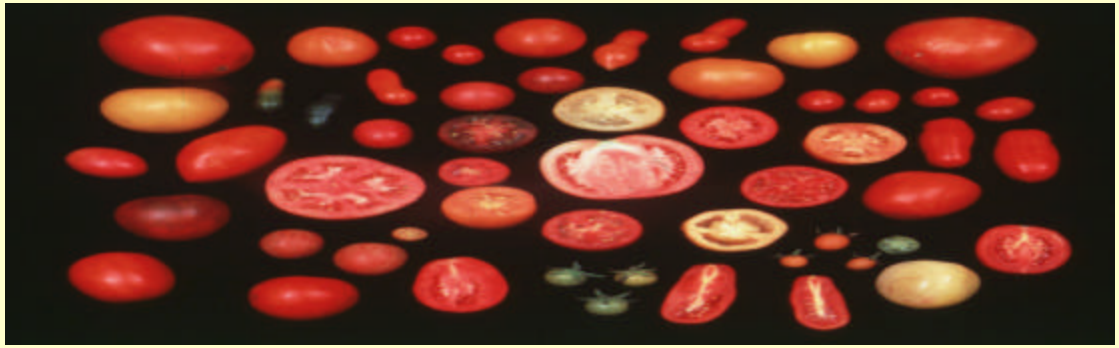




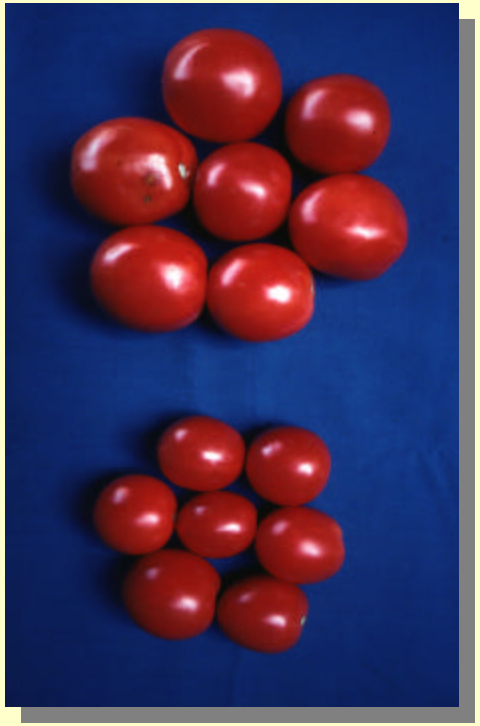
A Quantitative Trait Locus Key to the Evolution of Tomato Fruit

Size

- Domestication of many plants is correlated with dramatic increases in fruit size.



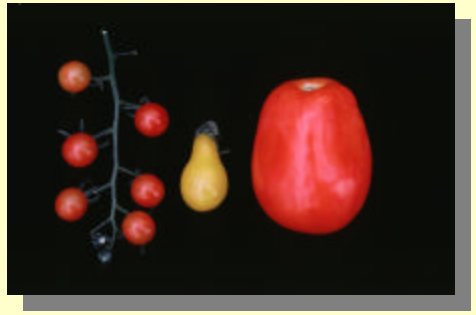
Chromosome 2



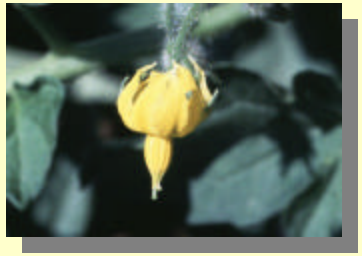
fw2.1, 2.2, 2.3

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

stuffer



ovate



se2.1

Loopp found a similar structure where Blast fails

- A gene that regulates the size of the tomato was found In Steven Tanksley lab. (Cornell) .
- The protein expressed by the gene was unknown. Blast failed to find a similar protein sequence.
- Loopp found structural similarity to the human oncogene c-H-ras .

A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size Anne Frary, T. Clint Nesbitt, Amy Frary, Silvana Grandillo, Esther van der Knaap, Bin Cong, Jiping Liu, Jaroslaw Meller, Ron Elber, Kevin B. Alpert, and Steven D. Tanksley *Science* 2000 July 7; 289: 85-88.

Literature

- Linear programming optimization and a double statistical filter for protein threading protocol. Protein : structure, function and genetics 45 241-261 (2001) Ron Elber Jarek Meller