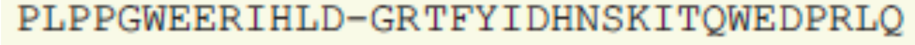
Statistical Physics, Inference and Applications to Biology

Physics Department, Ecole Normale Superieure, Paris, France.

Simona Cocco
Office:GH301 mail:cocco@lps.ens.fr

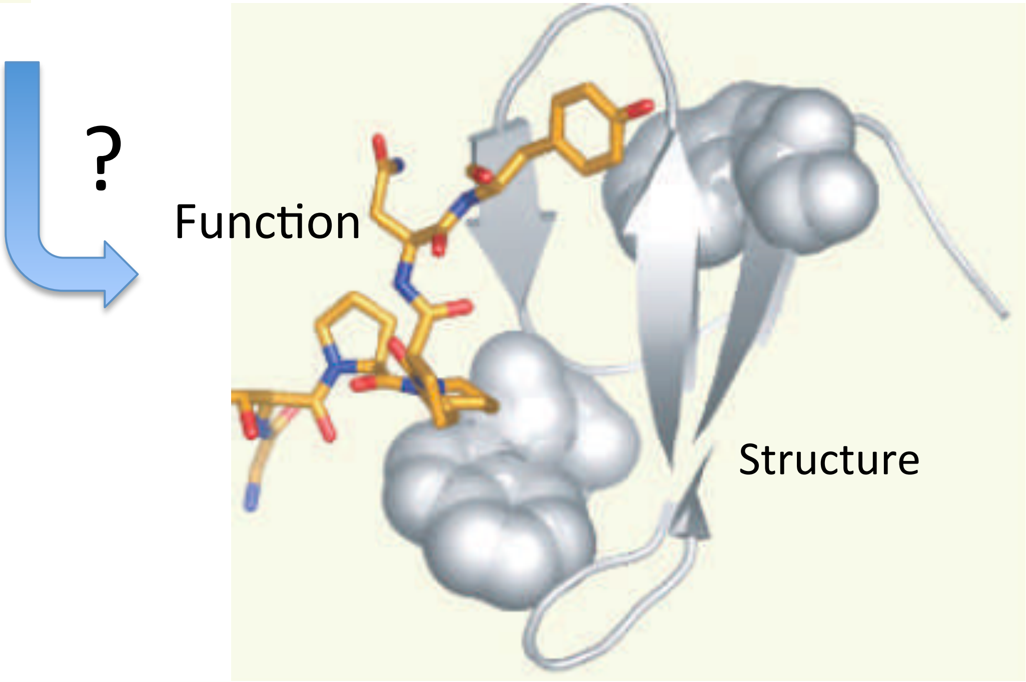
Deriving Protein Structure and Function from Sequence

Amino-Acid Sequence



PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ

WW protein



One of the major unsolved problems in biology.

Deriving Protein Structure and Function from Sequence

Amino-Acid Sequence

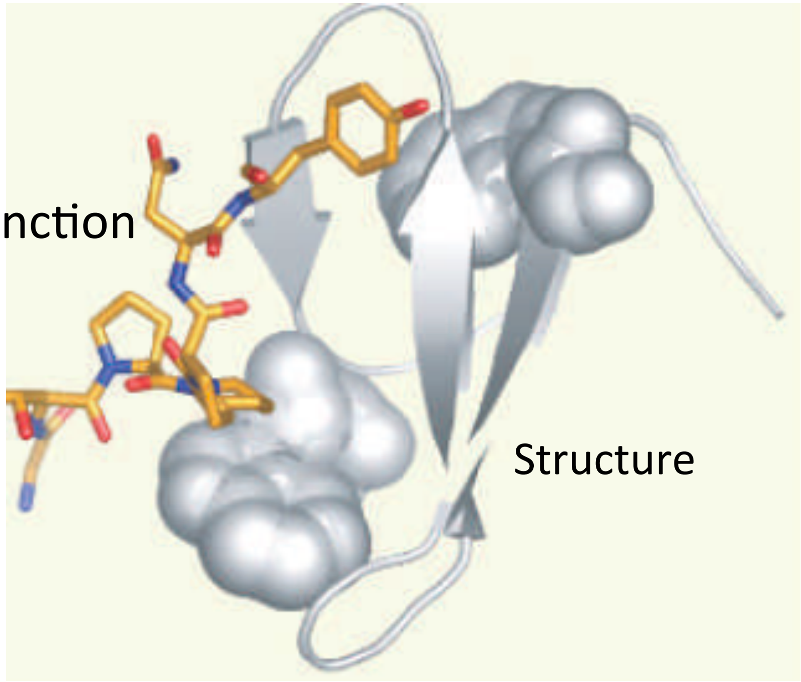
ALPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ

mutation

?

Function

WW protein

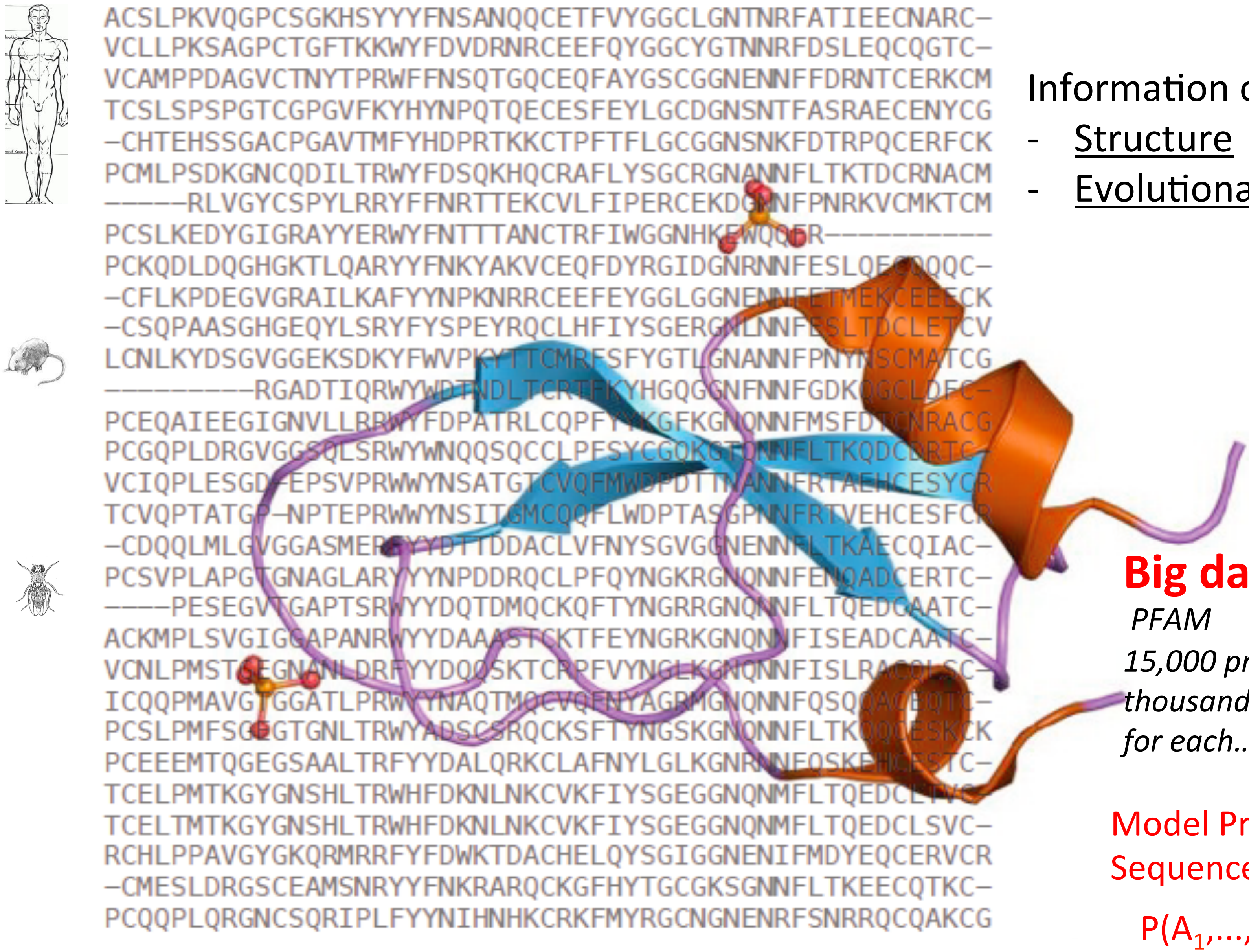


Structure

Mutations can be:

- **neutral**: no effect
- **detrimental**: **functionality decreases,** e.g. gives a cancer
- **beneficial: functionality increases**. Dangerous for us if the organism is a bacterium e.g. mutation confers resistance to antibiotics

Multi-sequence alignements (MSA)



```
ACSLPKVQGPCSGKHSYYYFNSANQQCETFVYGGCLGNTNRFATIEECNARC-
VCLLPKSAGPCTGFTKKWYFDVDRNRCEEFQYGGCYGTNNRFDSLEQCQGTC-
VCAMPPDAGVCTNYTPRWFFNSQTGQCEQFAYGSCGGNENNFFDRNTCERKCM
TCSLSPSPGTCGPGVFKYHYNPQTQECESFEYLGCDGNSNTFASRAECENYCG
-CHTEHSSGACPGAVTMFYHDPRTKKCTPFTFLGCGGNSNKFDTRPQCERFCK
PCMLPSDKGNCQDILTRWYFDSQKHQCRAFLYSGCRGNANNFLTKTDCRNACM
------RLVGYCSPYLRRYFFNRTTEKCVLFIPERCEKDGNNFPNRKVCMKTCM
PCSLKEDYGIGRAYYERWYFNTTTANCTRFIWGGNHKEWQQER----------
PCKQDLDQGHGKTLQARYYFNKYAKVCEQFDYRGIDGNRNNFESLQECQQQC-
-CFLKPDEGVGRAILKAFYYNPKNRRCEEFEYGGLGGNENNFETMEKCEEECK
-CSQPAASGHGEQYLSRYFYSPEYRQCLHFIYSGERGNLNNFESLTDCLETCV
LCNLKYDSGVGGEKSDKYFWVPKYTTCMRFSFYGTLGNANNFPNYNSCMATCG
----------RGADTIQRWYWDTNDLTCRTFKYHGQGGNFNNFGDKQGCLDFC-
PCEQAIEEGIGNVLLRRWYFDPATRLCQPFYYKGFKGNQNNFMSFDTCNRACG
PCGQPLDRGVGGSQLSRWYWNQQSQCCLPFSYCGQKGTQNNFLTKQDCDRTC-
VCIQPLESGD-EPSVPRWWYNSATGTCVQFMWDPDTTNANNFRTAEHCESYCR
TCVQPTATGP-NPTEPRWWYNSITGMCQQFLWDPTASGPNNFRTVEHCESFCR
-CDQQLMLGVGGASMERYYYDTTDDACLVFNYSGVGGNENNFLTKAECQIAC-
PCSVPLAPGTGNAGLARYYYNPDDRQCLPFQYNGKRGNQNNFENQADCERTC-
-----PESEGVTGAPTSRWYYDQTDMQCKQFTYNGRRGNQNNFLTQEDCAATC-
ACKMPLSVGIGGAPANRWYYDAAASTCKTFEYNGRKGNQNNFISEADCAATC-
VCNLPMSTGEGNANLDRFYYDQQSKTCRPFVYNGLKGNQNNFISLRACQLSC-
ICQQPMAVGTGGATLPRWYYNAQTMQCVQFNYAGRMGNQNNFQSQQACEQTC-
PCSLPMFSGEGTGNLTRWYADSCSRQCKSFTYNGSKGNQNNFLTKQQCESKCK
PCEEEMTQGEGSAALTRFYYDALQRKCLAFNYLGLKGNRNNFQSKEHCESTC-
TCELPMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLTVC-
TCELTMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLSVC-
RCHLPPAVGYGKQRMRRFYFDWKTDACHELQYSGIGGNENIFMDYEQCERVCR
-CMESLDRGSCEAMSNRYYFNKRARQCKGFHYTGCGKSGNNFLTKEECQTKC-
PCQQPLQRGNCSQRIPLFYYNIHNHKCRKFMYRGCNGNENRFSNRRQCQAKCG
```

Information on:

- <u>Structure</u>
- <u>Evolutionary Dynamics</u>

Big data bases!

PFAM

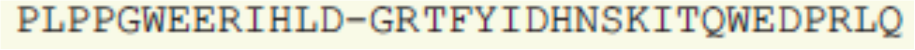
15,000 protein families

thousands of sequences for each…

Model Probability for Sequences in the MSA

$P(A_1,\ldots,A_N)$?

State of the art on Protein Sequences



PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ

Genetic code ←

P L P

CCUCUUCCGCCA....

State of the art on Protein Sequences

CCUCUUCCGCCA….

Theory

Classify sequences corresponding to different proteins (bioinformatics)

(Uniprot,PFAM,genbank)

Extract structural information from protein sequence data (statistical physics)

1995

~2000

~2010

2015

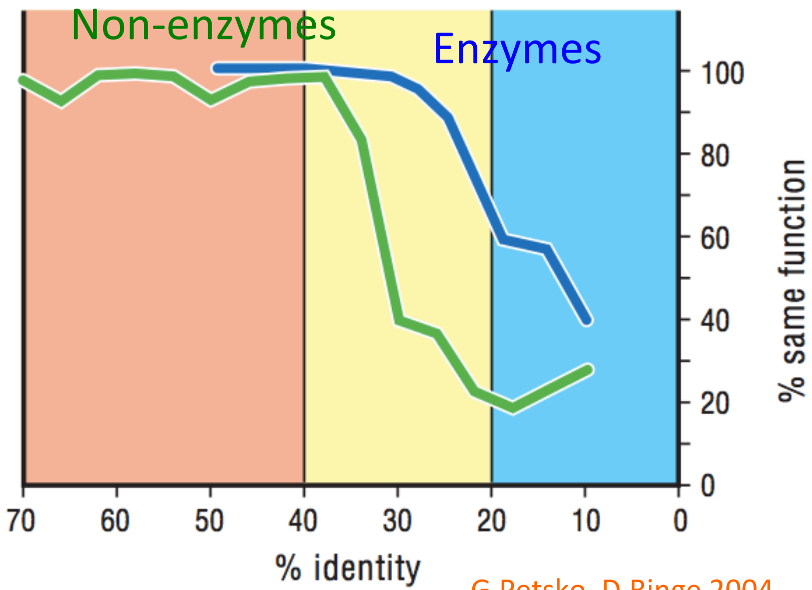
Sequencing

Cost of sequencing ($ per base): 1

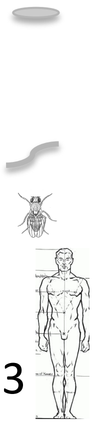
Cost of sequencing ($ per base): 10^{-6}

Nb. of protein sequences in databases ~ 10^8

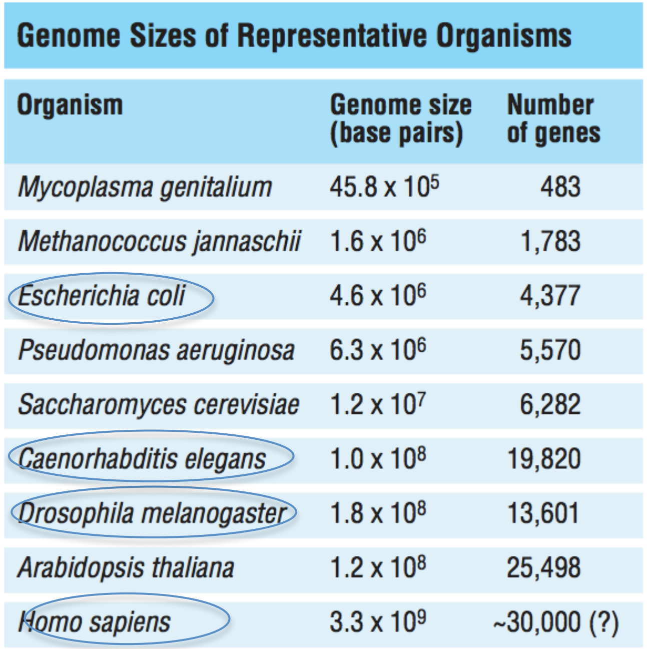
PLPPGWEERIHLD-GRTFYIDHNSKITQWEDPRLQ
GLPKPWIVKISRSRNRPYFFNTETHESLWEPPAAT



G.Petsko, D.Ringe 2004



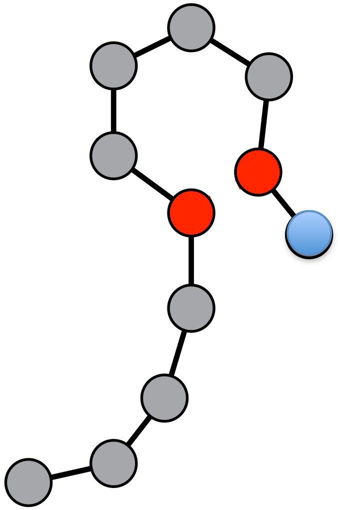
1998-2003

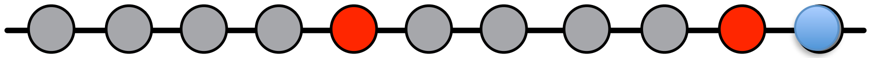


Genome Sizes of Representative Organisms

Organism	Genome size (base pairs)	Number of genes
Mycoplasma genitalium	45.8 x 10^5	483
Methanococcus jannaschii	1.6 x 10^6	1,783
Escherichia coli	4.6 x 10^6	4,377
Pseudomonas aeruginosa	6.3 x 10^6	5,570
Saccharomyces cerevisiae	1.2 x 10^7	6,282
Caenorhabditis elegans	1.0 x 10^8	19,820
Drosophila melanogaster	1.8 x 10^8	13,601
Arabidopsis thaliana	1.2 x 10^8	25,498
Homo sapiens	3.3 x 10^9	~30,000 (?)

Evolution reflects functional and structural constraints



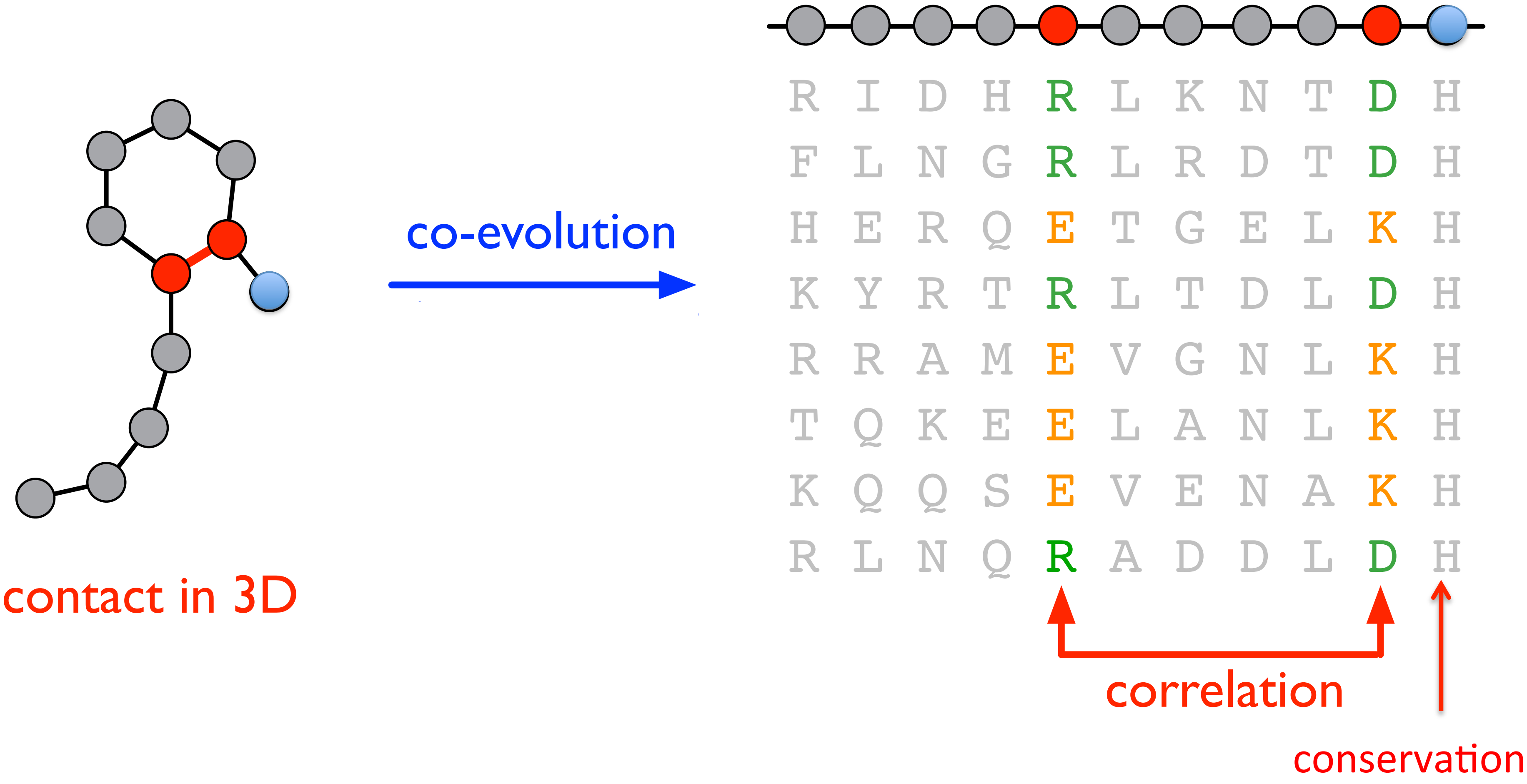


```
R I D H R L K N T D H
F L N G R L R D T D H
H E R Q E T G E L K H
K Y R T R L T D L D H
R R A M E V G N L K H
T Q K E E L A N L K H
K Q Q S E V E N A K H
R L N Q R A D D L D H
```

↑

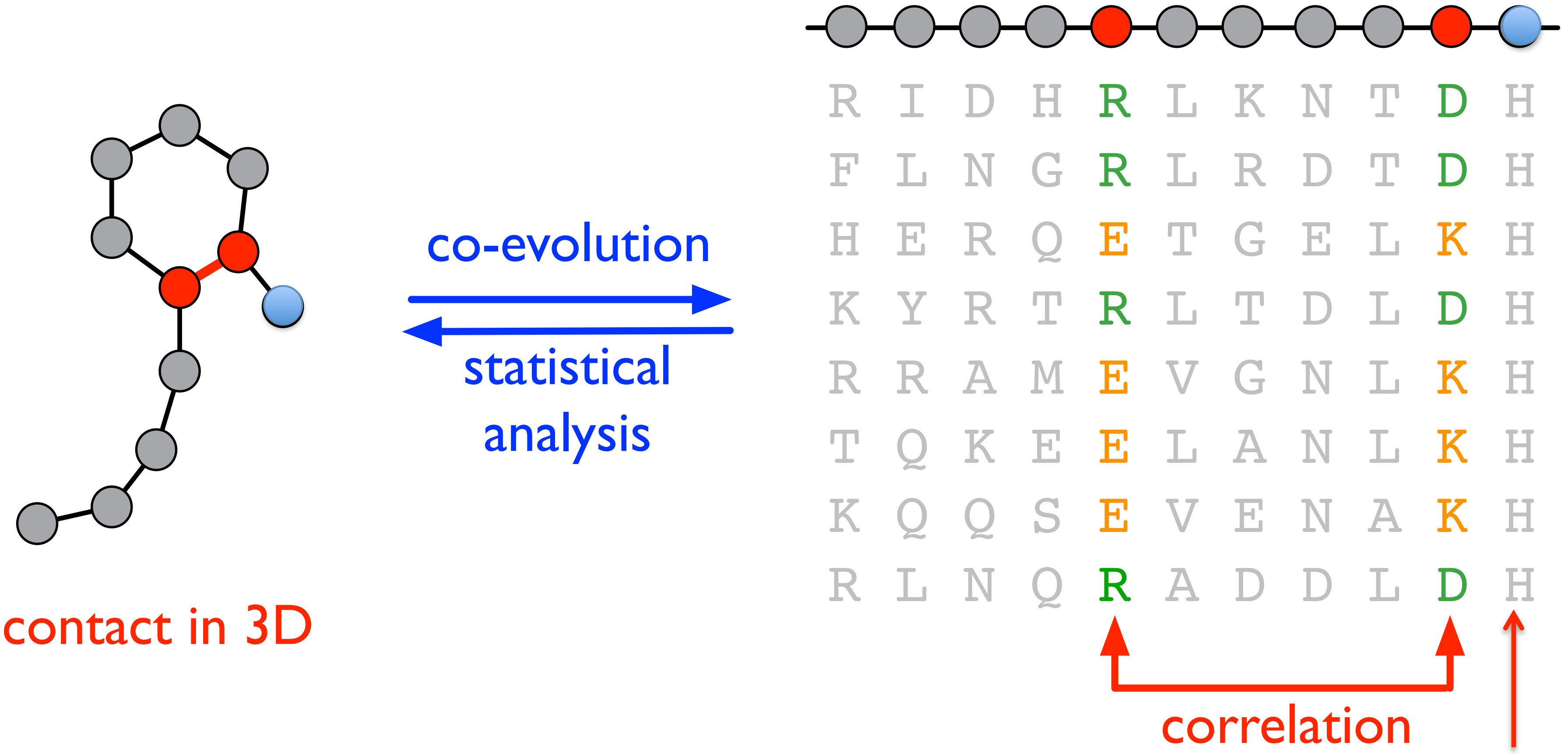
conservation

Evolution reflects functional and structural constraints



```
R I D H R L K N T D H
F L N G R L R D T D H
H E R Q E T G E L K H
K Y R T R L T D L D H
R R A M E V G N L K H
T Q K E E L A N L K H
K Q Q S E V E N A K H
R L N Q R A D D L D H
```

Evolution reflects functional and structural constraints



```
R I D H R L K N T D H
F L N G R L R D T D H
H E R Q E T G E L K H
K Y R T R L T D L D H
R R A M E V G N L K H
T Q K E E L A N L K H
K Q Q S E V E N A K H
R L N Q R A D D L D H
```

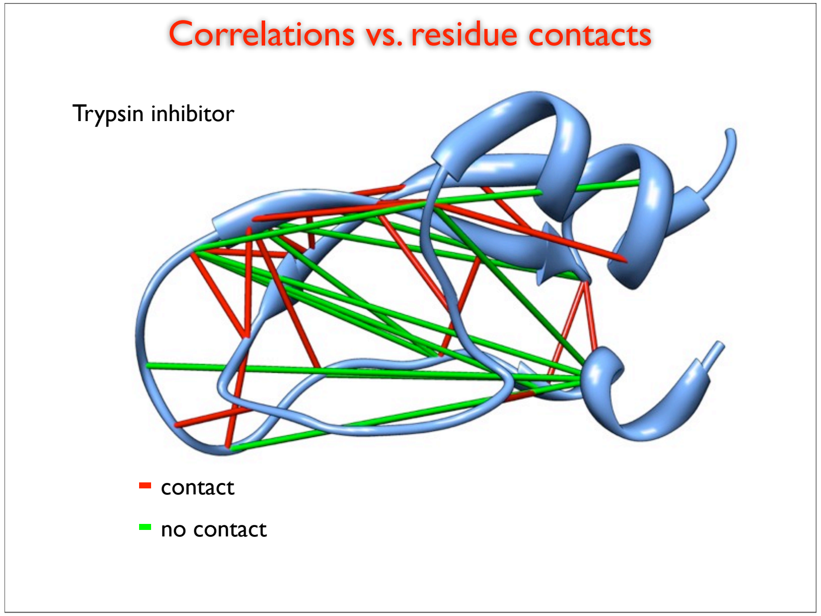
conservation

Inverse questions :

- Are sequence correlations indicative for residue-residue contacts?

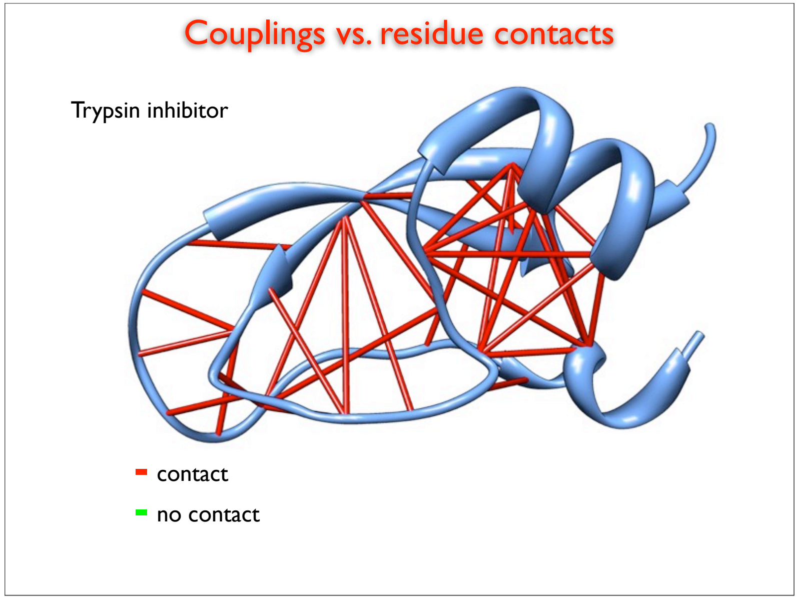
Statistical Couplings are better estimators of contacts than correlations …

Trypsin inhibitor (small protein, N=52 residues, contact if distance < 0.8 nm)



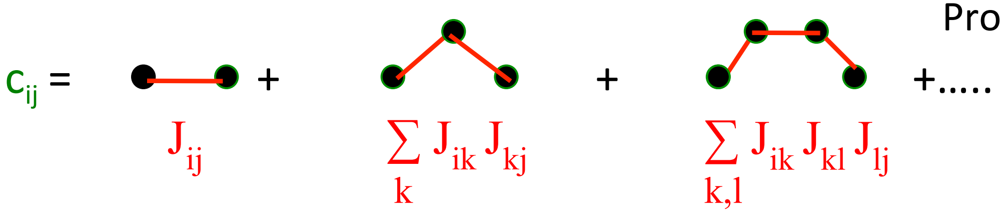
Correlations vs. residue contacts

[Gobel et al., Proteins 1994]



Couplings vs. residue contacts

[Weigt et al. PNAS 2009, Structural prediction:
Hopf et al. Cell 2012, Nugent, Jones 2012 (psicov)
Protein-protein interactions..]



$c_{ij} = J_{ij} + \sum_{k} J_{ik} J_{kj} + \sum_{k,l} J_{ik} J_{kl} J_{lj} + \ldots$

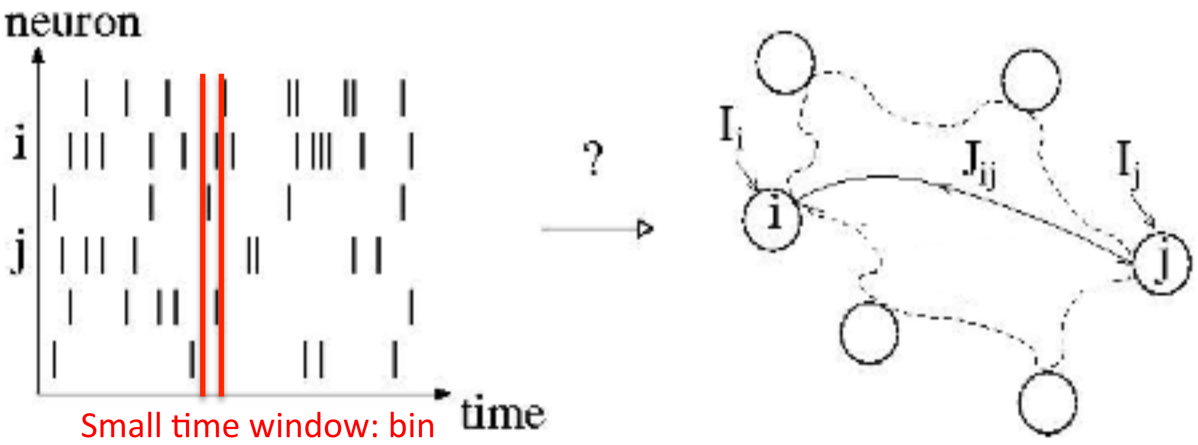
For Gaussian variables $J=-C^{-1}$

but C is affected by noise, need for regularization [S.C., Weigt, Monasson PloS Comp. Biol. 2013]

Inference of functional interactions from the spiking activity of a neural population

10-100 neurons

Raster Plot



Small time window: bin

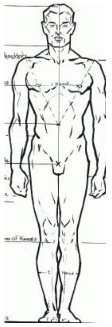
time: 1hour, 10^4-10^6 spikes

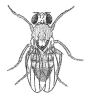
Retina, Cortex,…

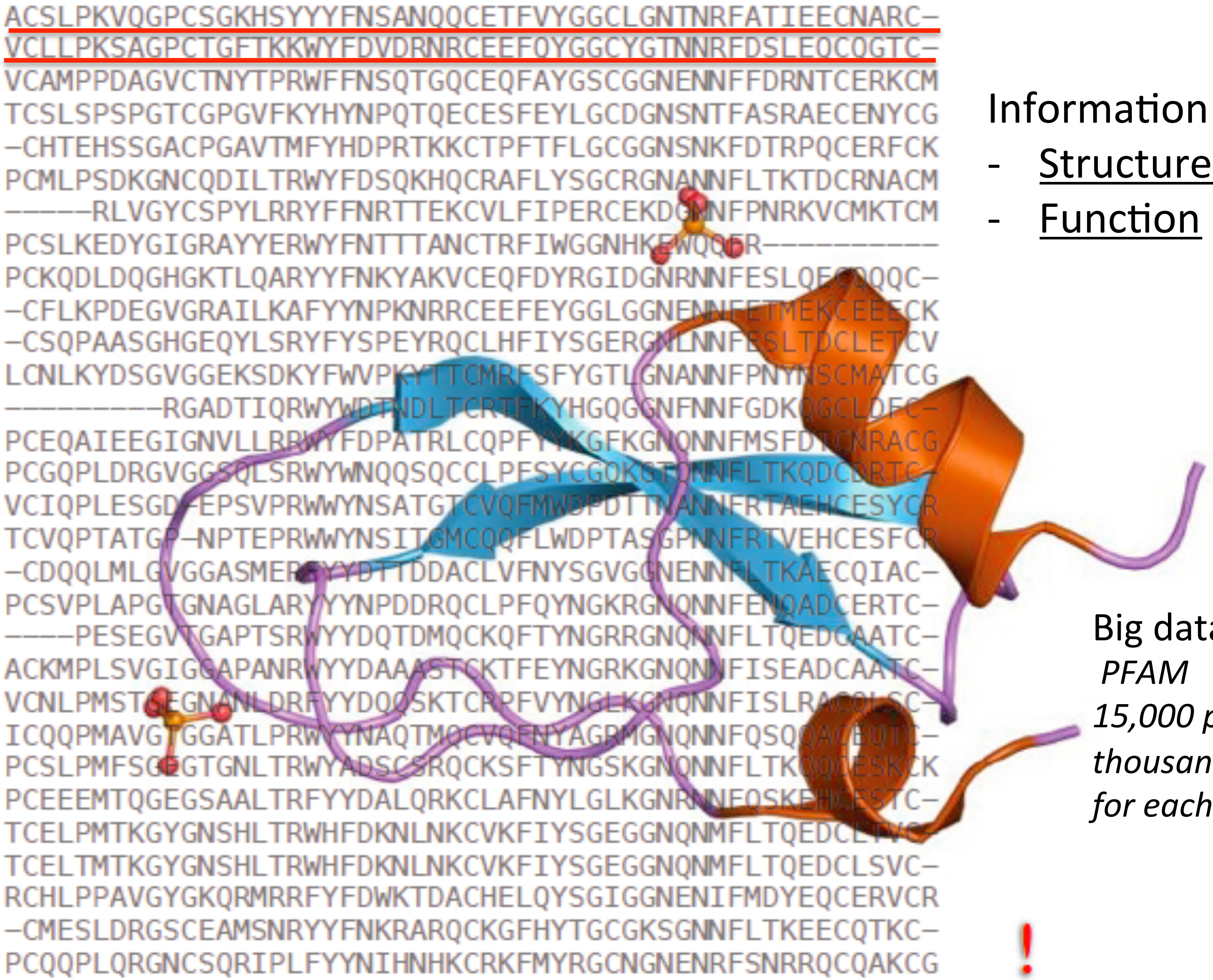
Correlations ⟶ *Interactions*

Aim: to find an Ising model $P(s_1 \ldots s_N)$ which reproduces and interprets the data

Multi-sequence alignements







```
ACSLPKVQGPCSGKHSYYYFNSANQQCETFVYGGCLGNTNRFATIEECNARC-
VCLLPKSAGPCTGFTKKWYFDVDRNRCEEFQYGGCYGTNNRFDSLEQCQGTC-
VCAMPPDAGVCTNYTPRWFFNSQTGQCEQFAYGSCGGNENNFFDRNTCERKCM
TCSLSPSPGTCGPGVFKYHYNPQTQECESFEYLGCDGNSNTFASRAECENYCG
-CHTEHSSGACPGAVTMFYHDPRTKKCTPFTFLGCGGNSNKFDTRPQCERFCK
PCMLPSDKGNCQDILTRWYFDSQKHQCRAFLYSGCRGNANNFLTKTDCRNACM
------RLVGYCSPYLRRYFFNRTTEKCVLFIPERCEKDGNNFPNRKVCMKTCM
PCSLKEDYGIGRAYYERWYFNTTTANCTRFIWGGNHKEWQQFR----------
PCKQDLDQGHGKTLQARYYFNKYAKVCEQFDYRGIDGNRNNFESLQECQQQC-
-CFLKPDEGVGRAILKAFYYNPKNRRCEEFEYGGLGGNENNFETMEKCEEECK
-CSQPAASGHGEQYLSRYFYSPEYRQCLHFIYSGERGNLNNFESLTDCLETCV
LCNLKYDSGVGGEKSDKYFWVPKYTTCMRFSFYGTLGNANNFPNYNSCMATCG
----------RGADTIQRWYWDTNDLTCRTFKYHGQGGNFNNFGDKQGCLDFC-
PCEQAIEEGIGNVLLRRWYFDPATRLCQPFTYKGFKGNQNNFMSFDTCNRACG
PCGQPLDRGVGGSQLSRWYWNQQSQCCLPFSYCGQKGTQNNFLTKQDCDRTC-
VCIQPLESGD-EPSVPRWWYNSATGTCVQFMWDPDTTNANNFRTAEHCESYCR
TCVQPTATGP-NPTEPRWWYNSITGMCQQFLWDPTASGPNNFRTVEHCESFCR
-CDQQLMLGVGGASMERYYYDTTDDACLVFNYSGVGGNENNFLTKAECQIAC-
PCSVPLAPGTGNAGLARYYYNPDDRQCLPFQYNGKRGNQNNFENQADCERTC-
-----PESEGVTGAPTSRWYYDQTDMQCKQFTYNGRRGNQNNFLTQEDCAATC-
ACKMPLSVGIGGAPANRWYYDAAASTCKTFEYNGRKGNQNNFISEADCAATC-
VCNLPMSTGEGNANLDRFYYDQQSKTCRPFVYNGLKGNQNNFISLRACQLSC-
ICQQPMAVGTGGATLPRWYYNAQTMQCVQFNYAGRMGNQNNFQSQQACEQTC-
PCSLPMFSGEGTGNLTRWYADSCSRQCKSFTYNGSKGNQNNFLTKQQCESKCK
PCEEEMTQGEGSAALTRFYYDALQRKCLAFNYLGLKGNRNNFQSKEHCESTC-
TCELPMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLTVC-
TCELTMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLSVC-
RCHLPPAVGYGKQRMRRFYFDWKTDACHELQYSGIGGNENIFMDYEQCERVCR
-CMESLDRGSCEAMSNRYYFNKRARQCKGFHYTGCGKSGNNFLTKEECQTKC-
PCQQPLQRGNCSQRIPLFYYNIHNHKCRKFMYRGCNGNENRFSNRRQCQAKCG
```

Information on:

- Structure
- Function

Big data bases!

PFAM

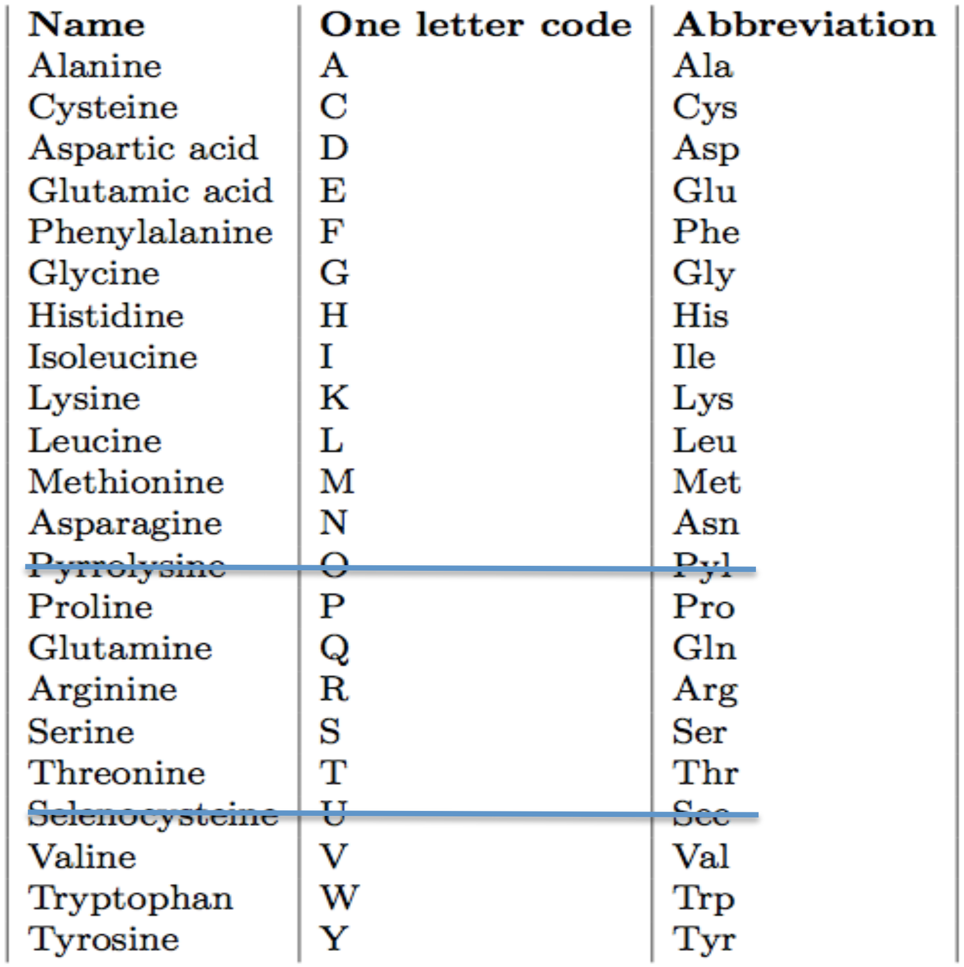
15,000 protein families

thousands of sequences for each…

!

Aim: to find a Potts model P(A_1…A_N) which reproduces and interpret the data

A 20+1 possible state model: Potts model



Name	One letter code	Abbreviation
Alanine	A	Ala
Cysteine	C	Cys
Aspartic acid	D	Asp
Glutamic acid	E	Glu
Phenylalanine	F	Phe
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Lysine	K	Lys
Leucine	L	Leu
Methionine	M	Met
Asparagine	N	Asn
~~Pyrrolysine~~	~~O~~	~~Pyl~~
Proline	P	Pro
Glutamine	Q	Gln
Arginine	R	Arg
Serine	S	Ser
Threonine	T	Thr
~~Selenocysteine~~	~~U~~	~~Sec~~
Valine	V	Val
Tryptophan	W	Trp
Tyrosine	Y	Tyr

Broad Outline: Applications

- Potts model to understand structure of proteins
- Potts model to decode relationship between protein's sequence and function (genotype-fenotype mapping):

- Design new proteins with the same structure and function of the   natural ones .

-Forecasting Viral Evolution

-Predicting antibiotic resistance

Plan

Theory :

- The coupled 20-letter model for Proteins: The Potts Model
- Pseudo-Likelihood Method

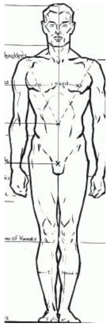
Applications:

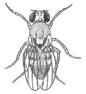
Analysis of Multiple Sequence Alignment in Proteins for:

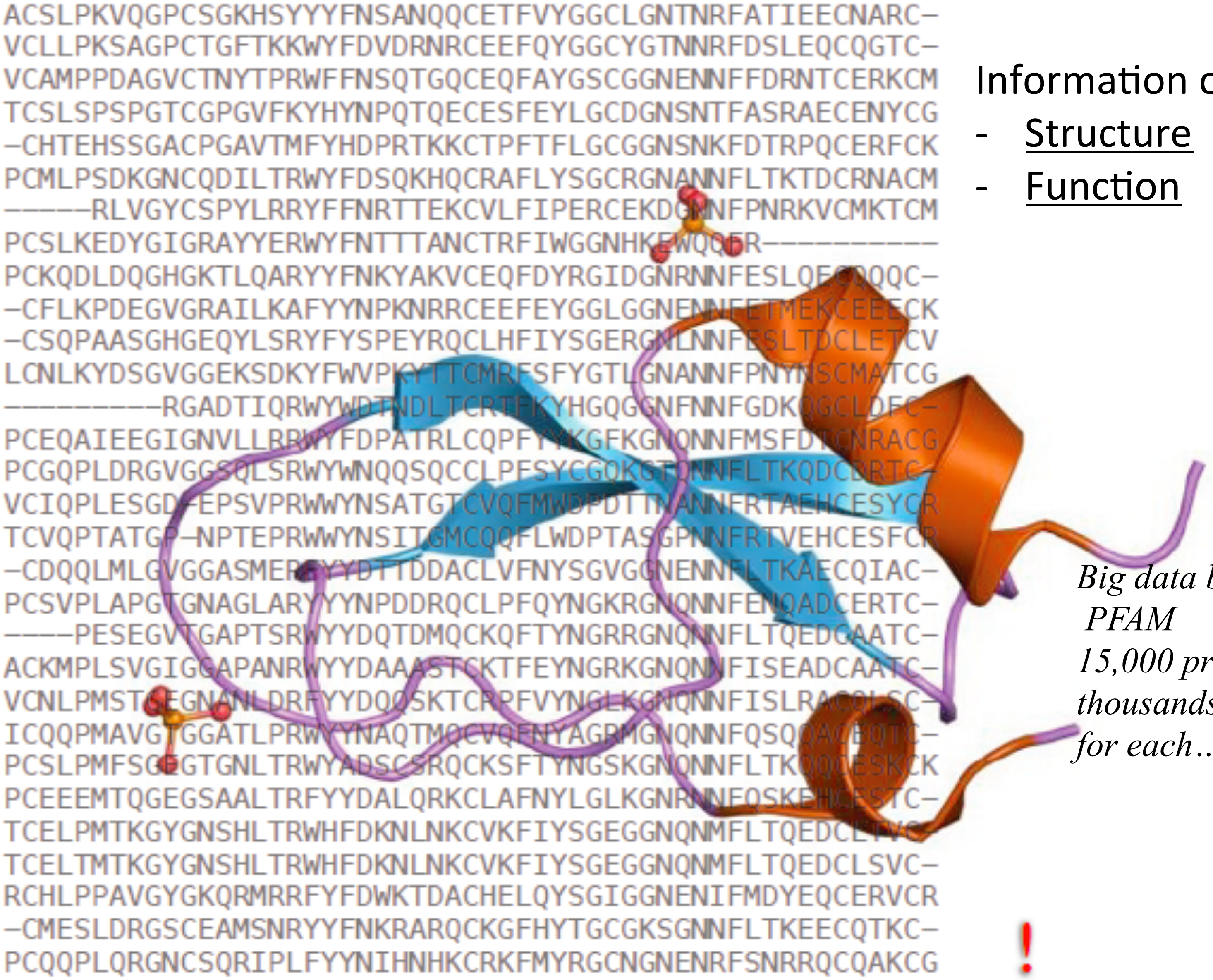
- Contacts predictions
- Structural prediction
- Design new functional proteins the case of a small protein domain WW
- Predict viral evolution for HIV virus
- Predict antibiotic resistance of TEM 1

Example 4:

Multi-sequence alignements







```
ACSLPKVQGPCSGKHSYYYFNSANQQCETFVYGGCLGNTNRFATIEECNARC-
VCLLPKSAGPCTGFTKKWYFDVDRNRCEEFQYGGCYGTNNRFDSLEQCQGTC-
VCAMPPDAGVCTNYTPRWFFNSQTGQCEQFAYGSCGGNENNFFDRNTCERKCM
TCSLSPSPGTCGPGVFKYHYNPQTQECESFEYLGCDGNSNTFASRAECENYCG
-CHTEHSSGACPGAVTMFYHDPRTKKCTPFTFLGCGGNSNKFDTRPQCERFCK
PCMLPSDKGNCQDILTRWYFDSQKHQCRAFLYSGCRGNANNFLTKTDCRNACM
-----RLVGYCSPYLRRYFFNRTTEKCVLFIPERCEKDGNNFPNRKVCMKTCM
PCSLKEDYGIGRAYYERWYFNTTTANCTRFIWGGNHKEWQQER----------
PCKQDLDQGHGKTLQARYYFNKYAKVCEQFDYRGIDGNRNNFESLQECQQQC-
-CFLKPDEGVGRAILKAFYYNPKNRRCEEFEYGGLGGNENNFETMEKCEEECK
-CSQPAASGHGEQYLSRYFYSPEYRQCLHFIYSGERGNLNNFESLTDCLETCV
LCNLKYDSGVGGEKSDKYFWVPKYTTCMRFSFYGTLGNANNFPNYNSCMATCG
----------RGADTIQRWYWDTNDLTCRTFKYHGQGGNFNNFGDKQGCLDFC-
PCEQAIEEGIGNVLLRRWYFDPATRLCQPFYYKGFKGNQNNFMSFDTCNRACG
PCGQPLDRGVGGSQLSRWYWNQQSQCCLPFSYCGQKGTQNNFLTKQDCDRTC-
VCIQPLESGD-EPSVPRWWYNSATGTCVQFMWDPDTTNANNFRTAEHCESYCR
TCVQPTATGP-NPTEPRWWYNSITGMCQQFLWDPTASGPNNFRTVEHCESFCR
-CDQQLMLGVGGASMERYYYDTTDDACLVFNYSGVGGNENNFLTKAECQIAC-
PCSVPLAPGTGNAGLARYYYNPDDRQCLPFQYNGKRGNQNNFENQADCERTC-
-----PESEGVTGAPTSRWYYDQTDMQCKQFTYNGRRGNQNNFLTQEDCAATC-
ACKMPLSVGIGGAPANRWYYDAAASTCKTFEYNGRKGNQNNFISEADCAATC-
VCNLPMSTGEGNANLDRFYYDQQSKTCRPFVYNGLKGNQNNFISLRACQLSC-
ICQQPMAVGTGGATLPRWYYNAQTMQCVQFNYAGRMGNQNNFQSQQACEQTC-
PCSLPMFSGEGTGNLTRWYADSCSRQCKSFTYNGSKGNQNNFLTKQQCESKCK
PCEEEMTQGEGSAALTRFYYDALQRKCLAFNYLGLKGNRNNFQSKEHCESTC-
TCELPMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLTVC-
TCELTMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLSVC-
RCHLPPAVGYGKQRMRRFYFDWKTDACHELQYSGIGGNENIFMDYEQCERVCR
-CMESLDRGSCEAMSNRYYFNKRARQCKGFHYTGCGKSGNNFLTKEECQTKC-
PCQQPLQRGNCSQRIPLFYYNIHNHKCRKFMYRGCNGNENRFSNRRQCQAKCG
```

Information on:

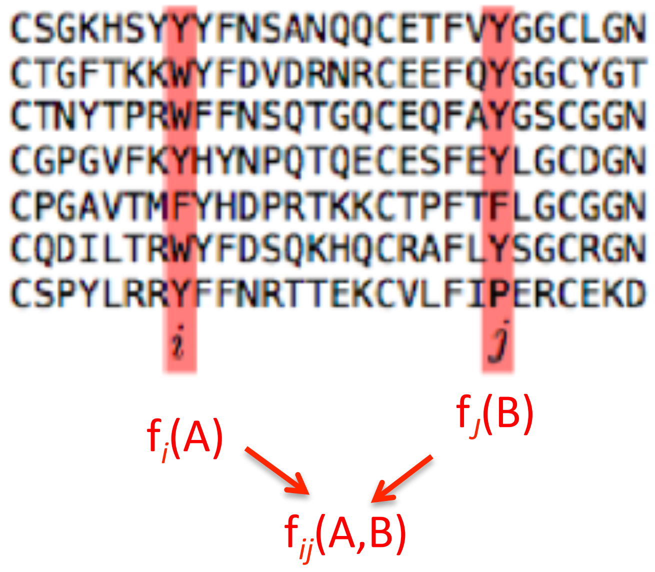
- <u>Structure</u>
- <u>Function</u>

Big data bases!
PFAM
15,000 protein families
thousands of sequences
for each…

!

Inference of couplings and fields for Potts model

Multiple sequence alignment (MSA): $\{A_i^m \mid i=1,..,N\ ;m=1,..M\}$



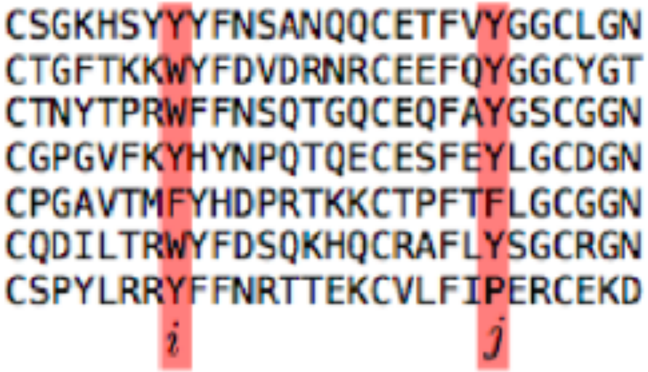
CSGKHSYYYFNSANQQCETFVYGGCLGN
CTGFTKKWYFDVDRNRCEEFQYGGCYGT
CTNYTPRWFFNSQTGQCEQFAYGSCGGN
CGPGVFKYHYNPQTQECESFEYLGCDGN
CPGAVTMFYHDPRTKKCTPFTFLGCGGN
CQDILTRWYFDSQKHQCRAFLYSGCRGN
CSPYLRRYFFNRTTEKCVLFIPERCEKD

$$\frac{1}{M}\sum_{m=1}^{M}\delta_{A,A_i} = f_i(A) \qquad \Leftarrow$$

$$\frac{1}{M}\sum_{m=1}^{M}\delta_{A,A_i}\delta_{B,A_j} = f_{ij}(A,B)$$

Inference of couplings and fields for Potts model

Multiple sequence alignment (MSA): $\{A_i^m \mid i=1,..,N\ ; m=1,..M\}$



CSGKHSYYYFNSANQQCETFVYGGCLGN
CTGFTKKWYFDVDRNRCEEFQYGGCYGT
CTNYTPRWFFNSQTGQCEQFAYGSCGGN
CGPGVFKYHYNPQTQECESFEYLGCDGN
CPGAVTMFYHDPRTKKCTPFTFLGCGGN
CQDILTRWYFDSQKHQCRAFLYSGCRGN
CSPYLRRYFFNRTTEKCVLFIPERCEKD

$$\begin{tikzcd} f_i(A) \arrow[dr] & & f_j(B) \arrow[dl] \\ & f_{ij}(A,B) & \end{tikzcd}$$

$$\sum_{\{A\}} P(A_1,\ldots,A_N)\,\delta_{A,A_i} = f_i(A) \qquad \Leftarrow$$

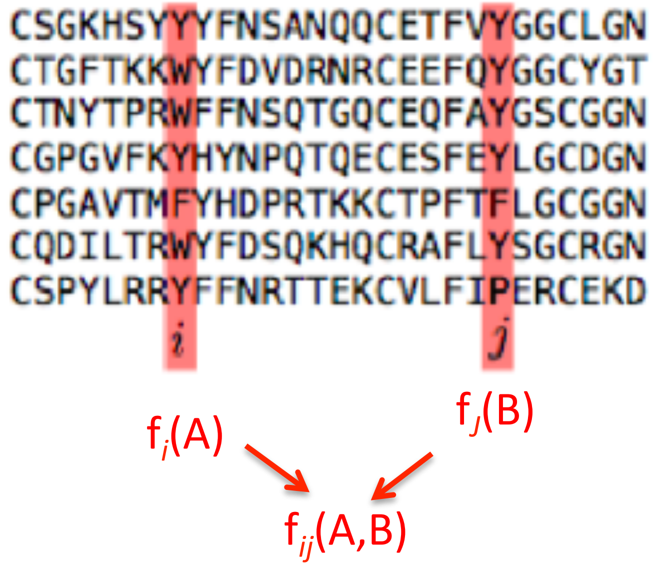
$$\sum_{\{A\}} P(A_1,\ldots,A_N)\,\delta_{A,A_i}\delta_{B,A_j} = f_{ij}(A,B)$$

Potts Model : $P(A_1,\ldots,A_N) = \dfrac{e^{\sum_{i<j} J_{ij}(A_i,A_j) + \sum_i h_i(A_i)}}{Z[\{J_{ij}(A,B),h_i(A)\}]}$

$\Rightarrow$ 21 N+21^2 N (N-1)/2 Coupled equations to solve!!

Inference of couplings and fields for Potts model

Multiple sequence alignment (MSA): $\{A_i^m \mid i=1,..,N\ ;m=1,..M\}$



```
CSGKHSYYYFNSANQQCETFVYGGCLGN
CTGFTKKWYFDVDRNRCEEFQYGGCYGT
CTNYTPRWFFNSQTGQCEQFAYGSCGGN
CGPGVFKYHYNPQTQECESFEYLGCDGN
CPGAVTMFYHDPRTKKCTPFTFLGCGGN
CQDILTRWYFDSQKHQCRAFLYSGCRGN
CSPYLRRYFFNRTTEKCVLFIPERCEKD
```

$f_i(A)$ $f_j(B)$

$f_{ij}(A,B)$

$$\sum_{\{A\}} P(A_1,...,A_N)\ \delta_{A,A_i} = f_i(A) \qquad \Leftarrow$$

$$\sum_{\{A\}} P(A_1,...,A_N)\ \delta_{A,A_i}\delta_{B,A_i} = f_{ij}(A,B)$$

Potts Model : $P(A_1,...,A_N) = \dfrac{e^{\sum_{i<j} J_{ij}\ (A_i,A_j)\ + \sum_i h_i(A_i)}}{Z[\{J_{ij}(A_i,A_j),h_i(A_i)\}]}$

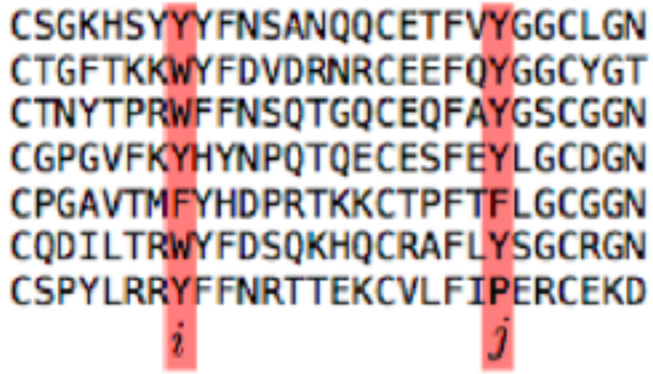
$\Rightarrow$ 21 N+21^2 N (N-1)/2 Coupled equations to solve!!

Problem rewritten as the Minimization of the Cross Entropy

$$S\ =\ \log Z[\{J_{ij}(A,B),h_i(A)\}] - \sum_{i<j,A,B} J_{ij}\,(A,B)\ \ f_{ij}(A,B) - \sum_{i,A} h_i(A)\ \ f_i(A)$$

Inference of couplings and fields for Potts model

Multiple sequence alignment (MSA): $\{A_i^m \mid i=1,..,N\ ;m=1,..M\}$



```
CSGKHSYYYFNSANQQCETFVYGGCLGN
CTGFTKKWYFDVDRNRCEEFQYGGCYGT
CTNYTPRWFFNSQTGQCEQFAYGSCGGN
CGPGVFKYHYNPQTQECESFEYLGCDGN
CPGAVTMFYHDPRTKKCTPFTFLGCGGN
CQDILTRWYFDSQKHQCRAFLYSGCRGN
CSPYLRRYFFNRTTEKCVLFIPERCEKD
```

$$\begin{tikzcd} f_i(A) \arrow[r] & f_{ij}(A,B) & f_j(B) \arrow[l] \end{tikzcd}$$

$$\sum_{\{A\}} P(A_1,\dots,A_N)\,\delta_{A,A_i} = f_i(A) \qquad \Leftarrow$$

$$\sum_{\{A\}} P(A_1,\dots,A_N)\,\delta_{A,A_i}\delta_{B,A_j} = f_{ij}(A,B)$$

Potts Model : $$P(A_1,\dots,A_N) = \frac{e^{\sum_{i<j} J_{ij}(A_i,A_j) + \sum_i h_i(A_i)}}{Z[\{J_{ij}(A_i,A_j),h_i(A_i)\}]}$$

$\Rightarrow$ 21 N+21^2 N (N-1)/2 Coupled equations to solve!!

Problem rewritten as the Minimization of the Cross Entropy

$$S = \log Z[\{J_{ij}(A,B),h_i(A)\}] - \sum_{i<j,A,B} J_{ij}(A,B)\ f_{ij}(A,B) - \sum_{i,A} h_i(A)\ f_i(A)$$

Regolarization is needed !

1. The Potts model is a generalization of the Ising model (binary variables) for any given number of category or Potts states in each site (21-a.a.)

2. Gauge invariance and overparametrization

It is possible to arbitrarly fix $h_i(c_i)$ =0, J_{ij} (c_i,b)=0

or other gauges without changing the model probability distribution

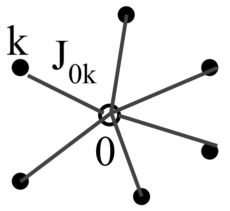
3. Minimising the Cross Entropy allows to find the parameters to reproduce frequencies and correlations

4. Equivalence between minimising the Cross Entropy and maximising the Log-Likelihood of the data given the model

5. Pseudo-Likelihood model

Pseudo-likelihood method

Pseudo log-likelihood of the node 0: l_0
=- pseudo cross entropy: S_0



$$S_0 = \frac{1}{B} \sum_{\tau=1}^{B} \log \left[\sum_{A} \exp\left(\sum_{,j} J_{0j}(A, A_j^{\tau}) + h_0(A) \right) \right] - \sum_{A} \left[h_0(A)\, p_0(A) - \sum_{j,Aj} J_{0j}(A, A_j)\, p_{ij}(A, A_j) \right]$$

Idea: avoid calculation of partition function using the sampled configurations

Total Pseudo-log likelihood: Sum of contributions of each node

Prior: increase signal/noise ratio by exploiting the sparsity of J_{ij}

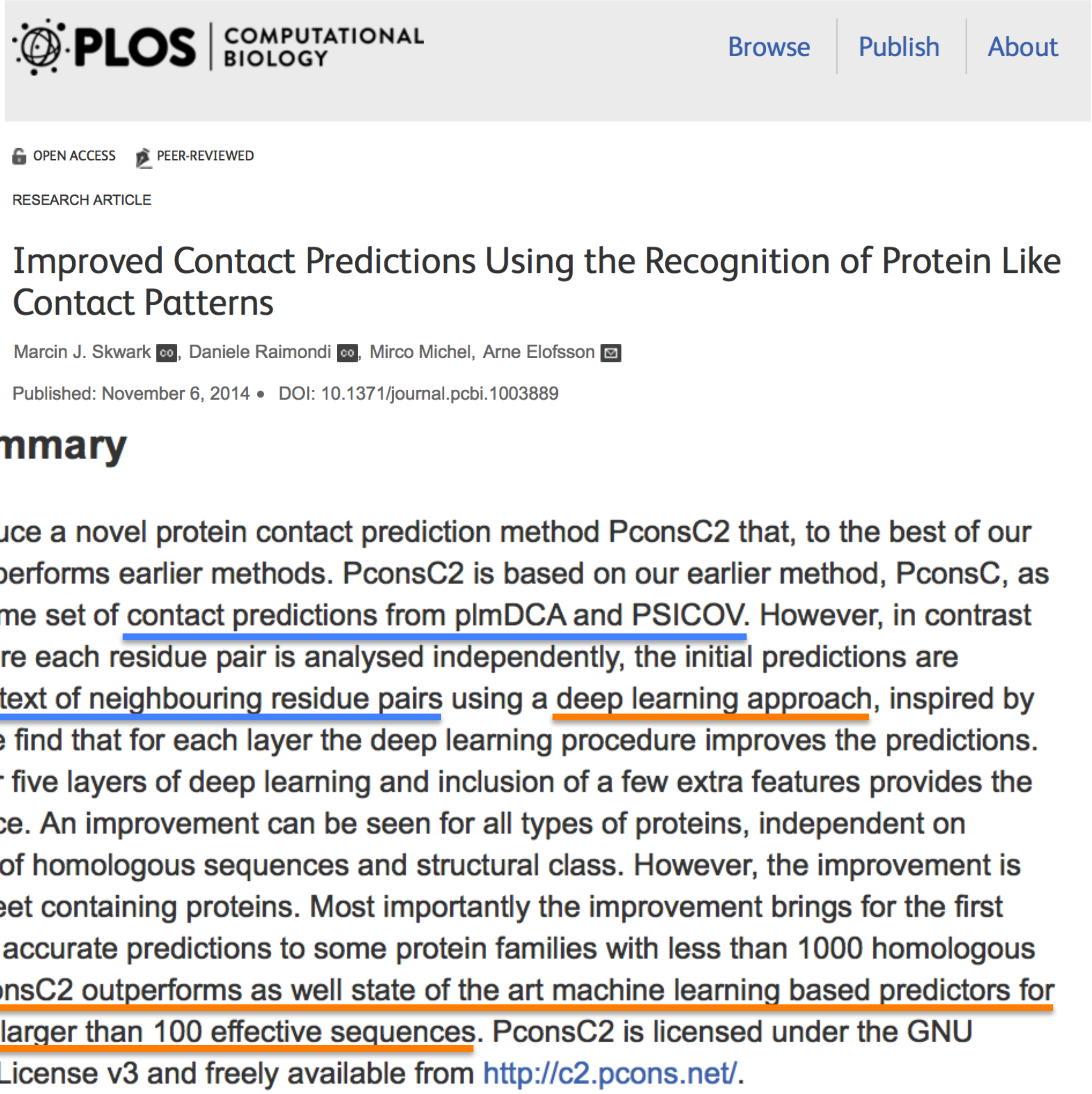
$$\text{cost function } (\{J\}) = \text{pseudo-cross-entropy} + \left\{ \begin{array}{l} \Gamma \sum_{ij} |J_{ij}(A,B)| \\ \Gamma \sum_{ij} J_{ij}^{2}(A,B) \end{array} \right.$$

Biblyography

- P. Ravikumar, M. J. Wainwright, and J. D. Lafferty (2010). *High-dimensional Ising model selection using ℓ1-regularized logistic regression*. J. Ann Stat. 38, 1287.

- M. Ekeberg, C. Lovkvist, Y. Lan, M. Weigt, E. Aurell (2013)*. Improved contact prediction in proteins: Using pseudolikelihood to infer Potts models*. Phys. Rev. E 87, 012707.

State of the art in protein structure prediction



PLOS | COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns

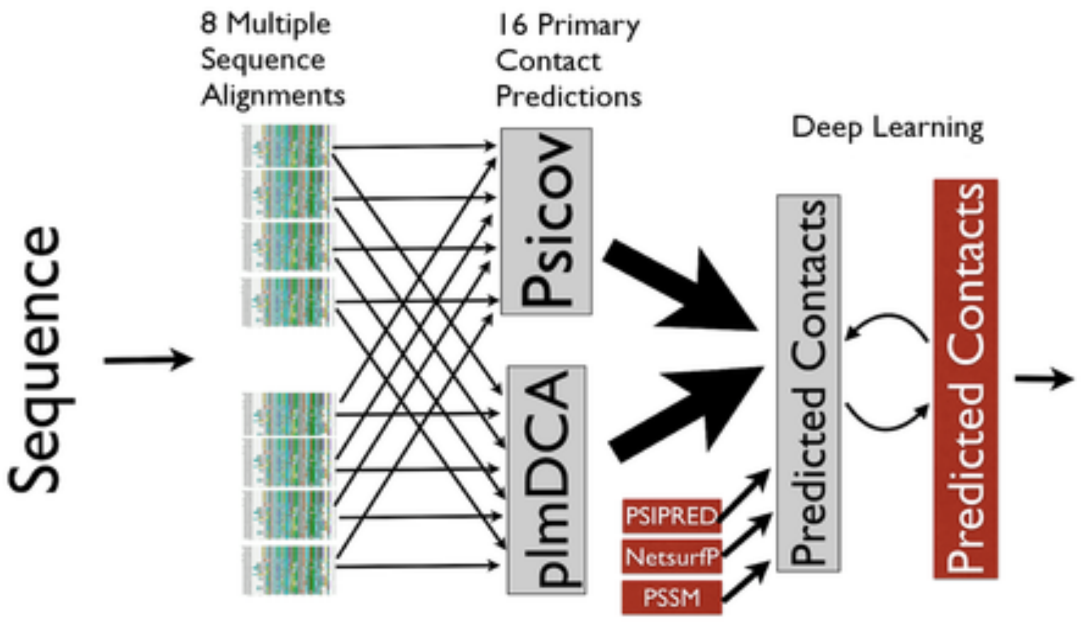
Marcin J. Skwark, Daniele Raimondi, Mirco Michel, Arne Elofsson

Published: November 6, 2014 • DOI: 10.1371/journal.pcbi.1003889

Author Summary

Here, we introduce a novel protein contact prediction method PconsC2 that, to the best of our knowledge, outperforms earlier methods. PconsC2 is based on our earlier method, PconsC, as it utilizes the same set of contact predictions from plmDCA and PSICOV. However, in contrast to PconsC, where each residue pair is analysed independently, the initial predictions are analysed in context of neighbouring residue pairs using a deep learning approach, inspired by earlier work. We find that for each layer the deep learning procedure improves the predictions. At the end, after five layers of deep learning and inclusion of a few extra features provides the best performance. An improvement can be seen for all types of proteins, independent on length, number of homologous sequences and structural class. However, the improvement is largest for β-sheet containing proteins. Most importantly the improvement brings for the first time sufficiently accurate predictions to some protein families with less than 1000 homologous sequences. PconsC2 outperforms as well state of the art machine learning based predictors for protein families larger than 100 effective sequences. PconsC2 is licensed under the GNU General Public License v3 and freely available from http://c2.pcons.net/.

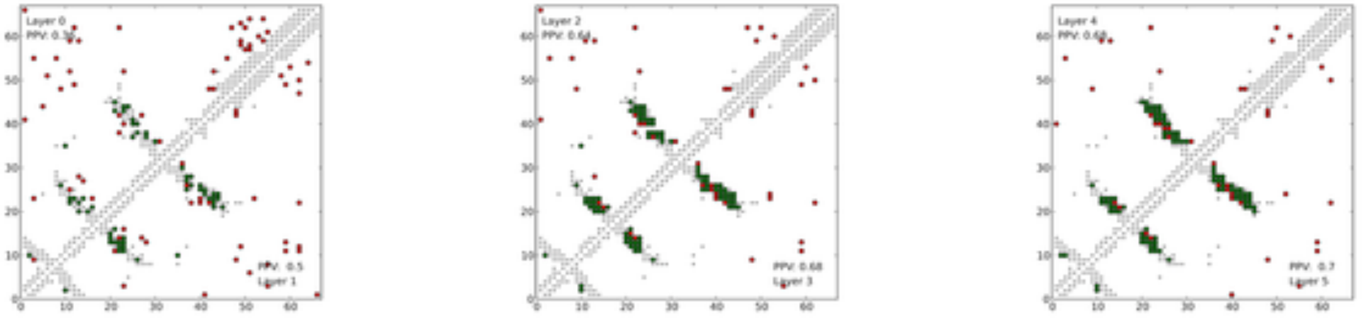
State of the art in protein structure prediction



True Positive rates:
(L predictions, 8 Å)

PconsC2	0.56
MaxEnt alone	0.37
Correlations	0.09

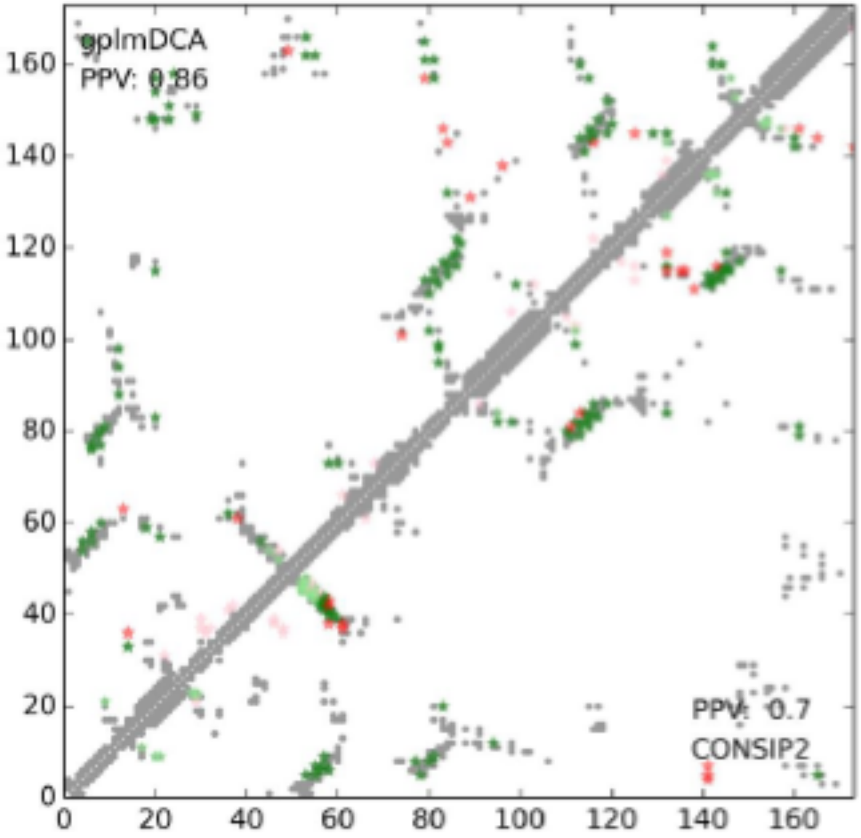
(a) Pipeline

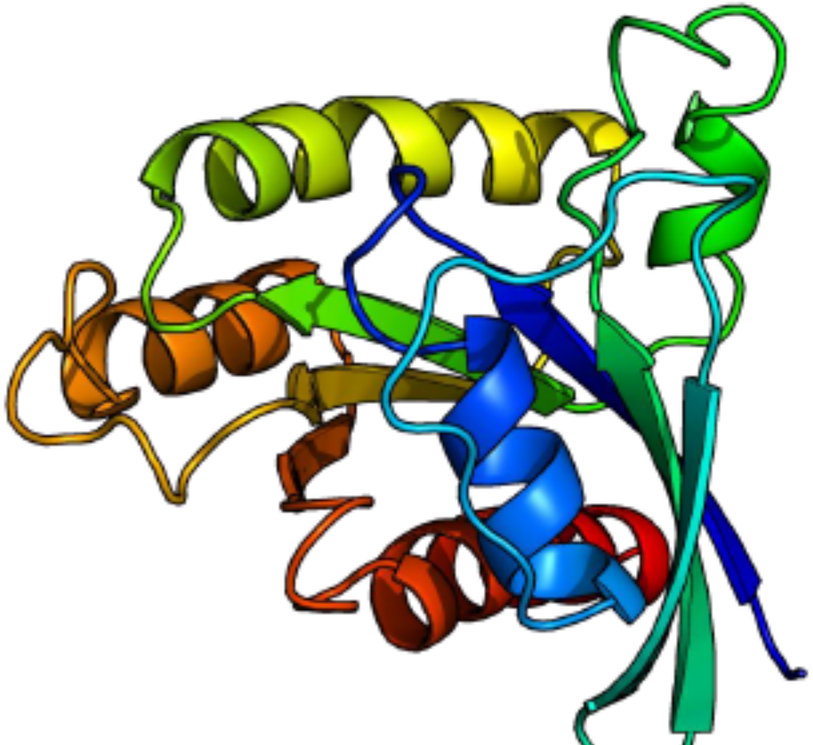


(b) 1pcf:A Layer 0-1 (c) 1pcf:A Layer 2-3 (d) 1pcf:A Layer 4-5

plmDCA: Ekeberg, Aurell (2014)

Prediction of Contacts by Pseudo-Likelihood in CASP 2015



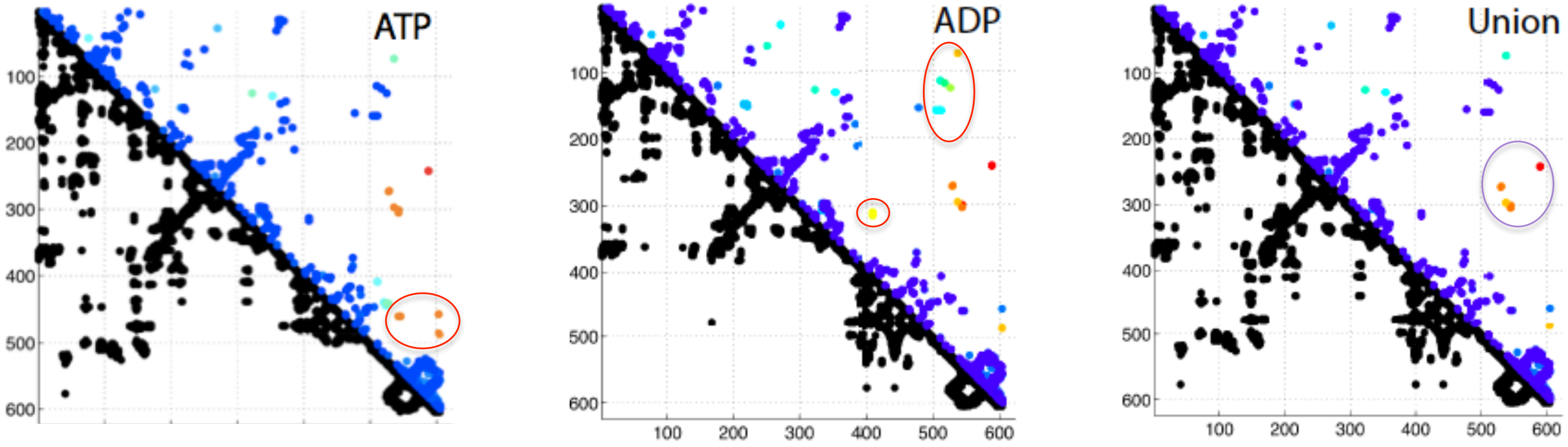


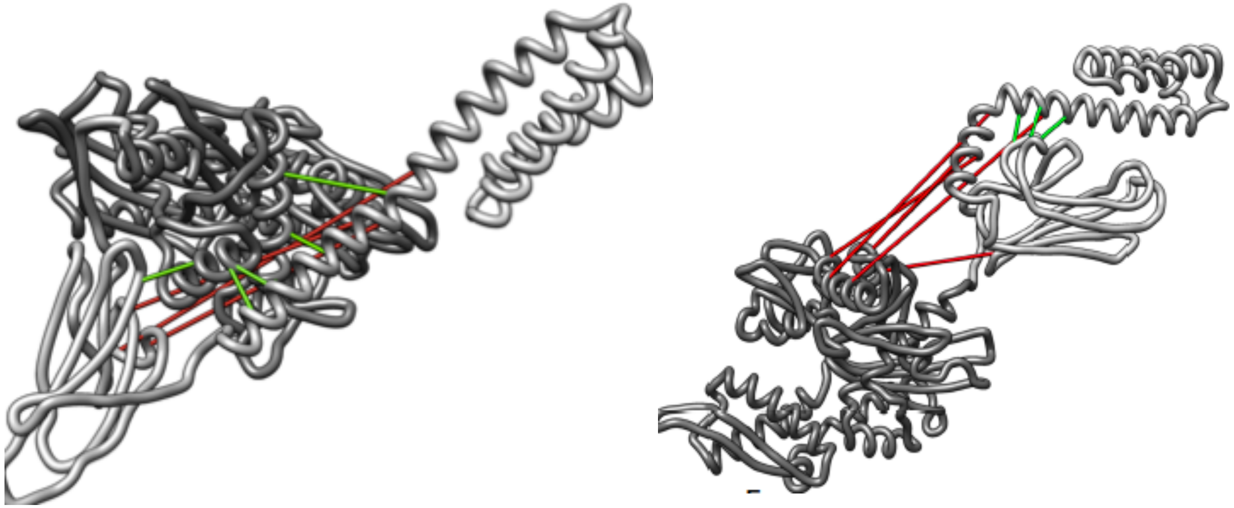
Marcin. Skwark (4th position on contact predictio)

T0798: RAS11B a protein involved in membrane trafficking, 88253 sequences at 90%

Ekeberg et al Phys Rev E 2013,Monastryskyy et al CASP 2015

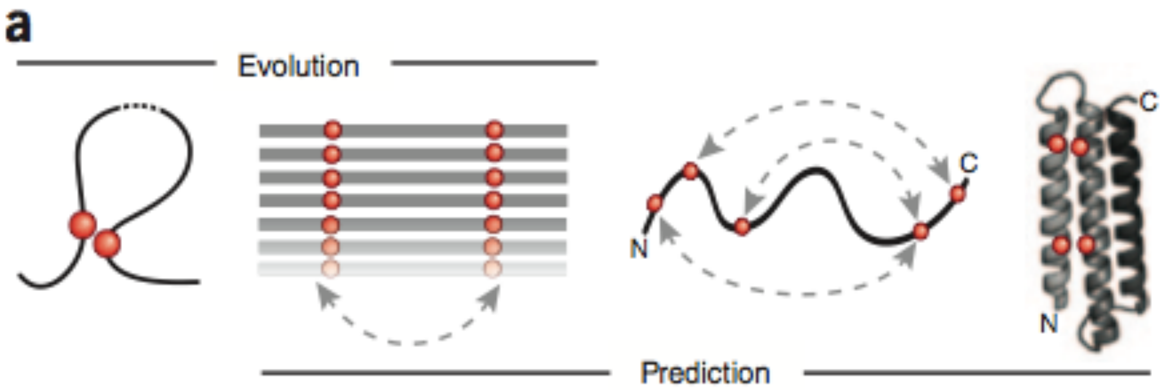
Prediction of Contacts in the Allosteric HSP70 protein





Maliverni D,.., De Los Rios P. HSP70 dimerization predicted by Sequence Variation Plos Comp Bio 100426 (2015).

Contact Prediction and Protein Folding



D. Marks T.Hopf, C. Sanders
Protein Structure Prediction Via Sequence variation
Nat. Biotechnology 2012.

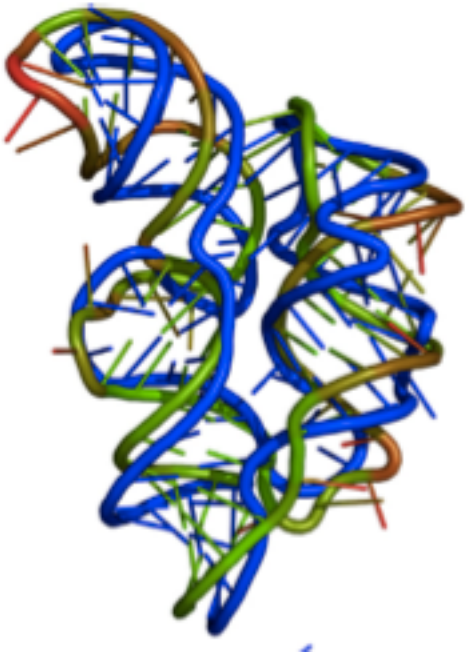
Classical Methods:
-All atom Potential:
Computationally very expensive!
-Effective Potential
difficult: find right level of
description, take into account
context dependence..

-Combine Contact Prediction and classical
Molecular Dynamic simulations or
Monte Carlo Methods with effective Potentials:
structures with very high resolution in the range of 2-6 Å

-Blind prediction for transmembranar proteins that are hard to crystallize

Contact Prediction and RNA Folding

RNA + Rosetta



Rosetta (Baker group): Monte Carlo Minimization
Of Effective Interaction potential between amino acids.
The bond angles are chosen among a library of known
and possible structures.
+ Lennard Jones attractive Potentials for pairs in contact

Rosetta alone RMSD ≈16 Å
Rosetta + First 25 Coplings:7.5 Å
Rosetta+ All contacts: 6.3 Å

Riboswitch:
2gdi (RF00059)

De Leonardis et al. NAR (2015)