KTH/CSC

Aalto University
School of Science
and Technology

Inferring protein structures from many protein sequences

SF2935 Modern Methods of Statistical Learning

Guest lecture: Erik Aurell

December 6, 2017

Magnus Ekeberg, Yueheng Lan, Cecilia Lövkvist, **E.A.**, Martin Weigt, *Phys. Rev. E* **87**:012707 (2013)
Magnus Ekeberg, Tuomo Hartonen, **E.A.** , *Journal of Computational Physics* **276**:341-356 (2014)

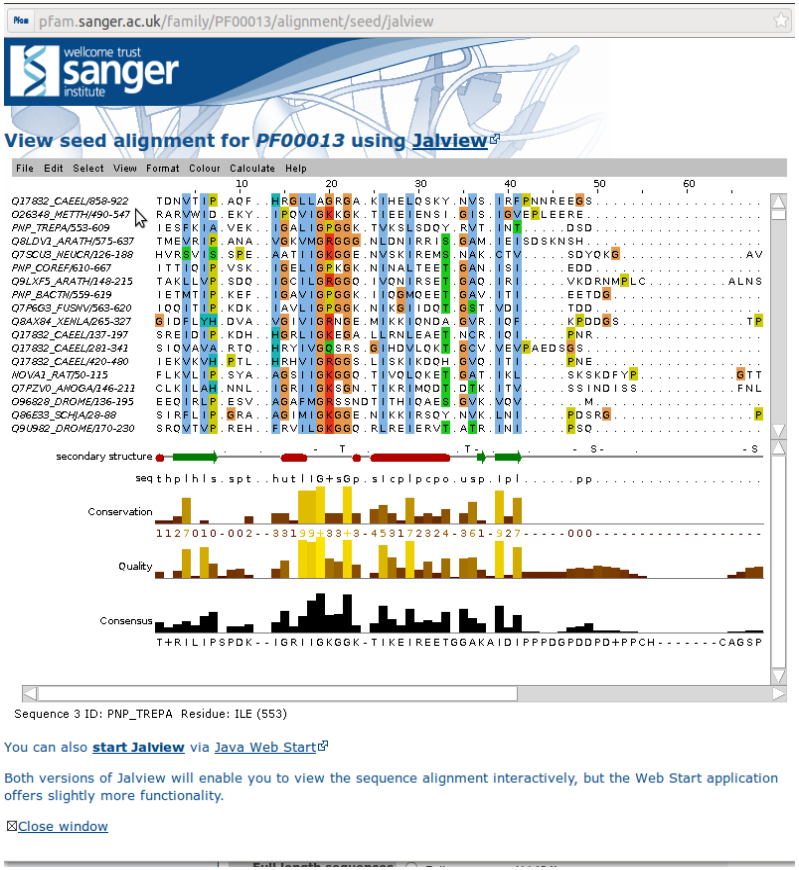
H. Chau Nguyen, Riccardo Zecchina, Johannes Berg, *Advances in Physics*, **66**: 197-261 (2017)

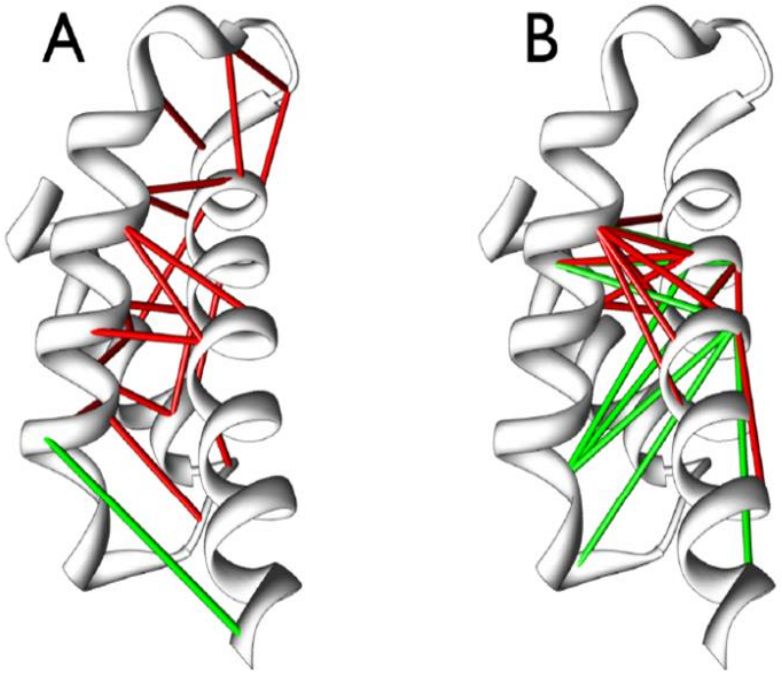
KTH/CSC

One-slide summary

Aalto University
School of Science
and Technology

From multiple sequence alignments of similar proteins one can learn their 3D structure (much) better than by correlations





direct coupling analysis (DCA)

ranking by correlations

Weigt *et al*, PNAS 2009; Morcos *et al* PNAS 2011; Hopf *et al* Cell 2012 + *others*

Outline

1. Background.
2. Examples. How well does DCA perform?
3. Methods. What is under the hood?
4. And then continued in next lecture on Dec 12

1A. Background

What is Direct Coupling Analysis?

DCA is a type of Model Learning

(1) learn models in exponential families from data; (2) use a small subset of largest inferred parameters to characterize the data

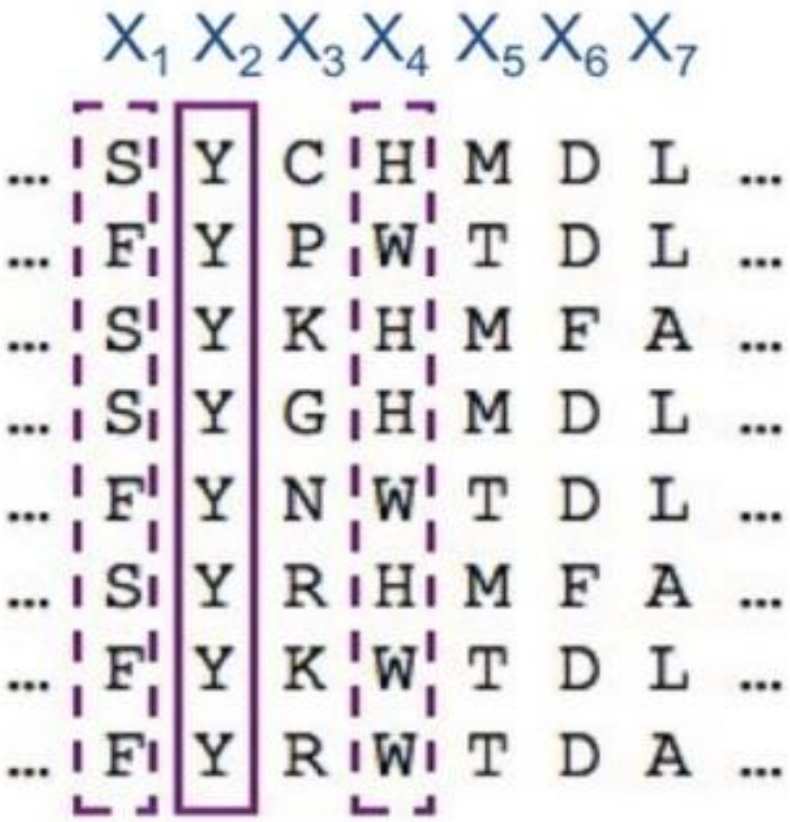
$$P(\mathbf{x}) = \frac{1}{Z(h,J)} \exp\left(\sum_i h_i(x_i) + \sum_{ij} J_{ij}(x_i, x_j) \right)$$

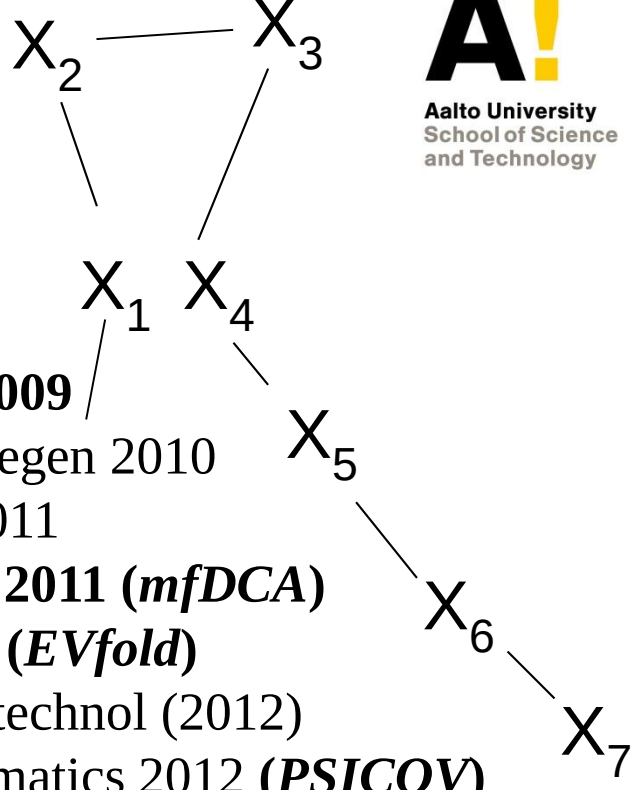
Executing (1) accurately and effectively on large data sets is a non-trivial task which has given rise to a fairly large methodological literature, *see* Nguyen, Berg & Zecchina [arXiv: 1702:01522].

More DCA background

Neher (1994)

Göbel, Sander, Schneider, Valencia (1994)





Lapedes et al 2001

Weigt et al PNAS 2009

Burger & van Nimwegen 2010

Balakrishnan et al 2011

Morcos et al PNAS 2011 (*mfDCA*)

Hopf et al Cell 2012 **(*EVfold*)**

Marks et al, Nat Biotechnol (2012)

Jones et al Bioinformatics 2012 **(*PSICOV*)**

Ekeberg et al Phys Rev E 2013 **(*plmDCA*)**

Skwark et al Bioinformatics 2013 **(*PconsC*)**

Kamisetty et al PNAS 2013 **(*GREMLIN*)**

Feinauer et al PLoS Comp Bio 2014 **(*gplmDCA*)**

Skwark et al PLoS Comp Bio 2014 **(*PconsC2*)**

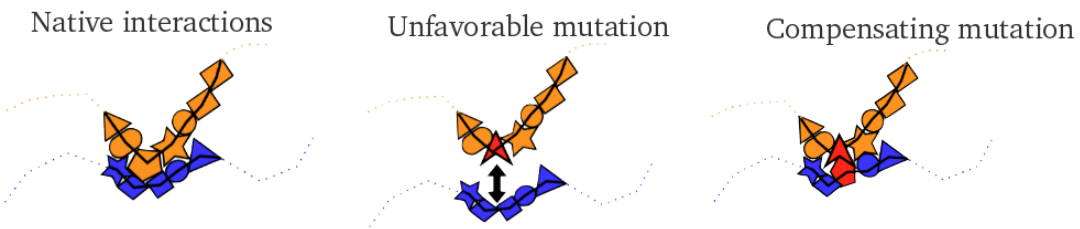
Jones et al Bioinformatics 2015 **(*MetaPSICOV*)**

…

KTH/CSC

Amino acids that are close together co-evolve

```
LLLDGSSSLPESYFDMMKSFAKAFISKANIGPHLTQVSVLQYGSINTID
LLLDGSSSLPASYFEEMKSFAKAFISKANIGPHHTQVSVLQYGSITTID
LLLDGSSGFPASEFDEMKSFAKAFISKANIGPQLTQVSVLQYGSITTID
FVLDGSSSVRASQFEEMKTFVKAFIKKVNIGVGATQVSVLQYGWRNILE
VLLDGSTNIMEPQFEEMKTFVKELIKKVDIGNNGTQISVVQYGKTNTLE
FILDTSSSVGKDNFEKIRKWVADLVDSFDVSPDKTRVAVVLYSDRPTIE
LAVDTSQSMEIQDLTVIKSVVDDFISHRK-N---DRIGLILFGTQAYLQ
FLVDTSGSLQKNGFDDEKVFVNSLLSHIRVSYKSTYVSVVLFGTSATID
LALDTSATTGETILDHITRGAQIGLAALS---DRSKVGVWLYGEDHRVV
YVIDTSGSMHGAKIEQTRESMVAILQDLH---EEDHFGILLFERKISYW
FLIDTSRSLGLRAYQKELQFVERVLEGYEIGTNRTKVAVITFSAGSRLE
ILLDTSSSIKINNFDLIRKFVANIINQFEVGRNGLMVGMATYS--RSVQ
FILDTSGSVGSYNFEKMKTFVKNVVDFFNIGPKGTHVAVITYSTWA--Q
FALDTSTSIGSQNFEREKQFVLAFVTDMDIGRSDVQVSVGTFSDNARRY
LLLDTSGSMQGAAIEALLSLKDEL-VKNSIAARRVEIAIVTFDSHINVV
LLLDTSGSMKGEPLDALRTFQQEL-DRDSLAKKRVEVAIVTFNSDVEIV
LSVDVSLSMLARRLSALRDIAIRFVQKRK----NDRVGLVTYSGEALAR
LAMDVSGSMQANRLEAAKDVAISFINNRNIG-------MVTFAGESFTQ
MSVDVSLSMLARRLTALKNIAKKFVDKRP----GDRIGLVTYSGEAFTK
VLADVSGSMQGEPIAA-AAFTRYL-QNEV-ASKRVEVAVVTFGTVATVL
```



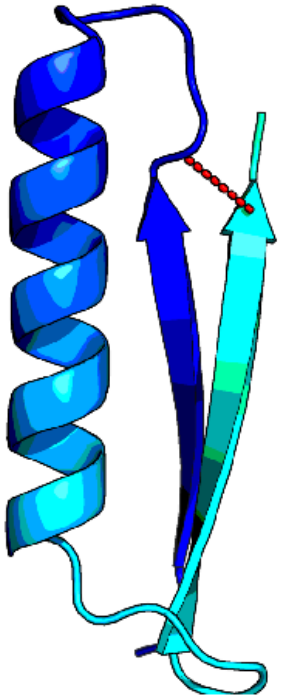


Image source: Andrea Pagnani, International School on Physics of Complex Systems 2012

1B. Background

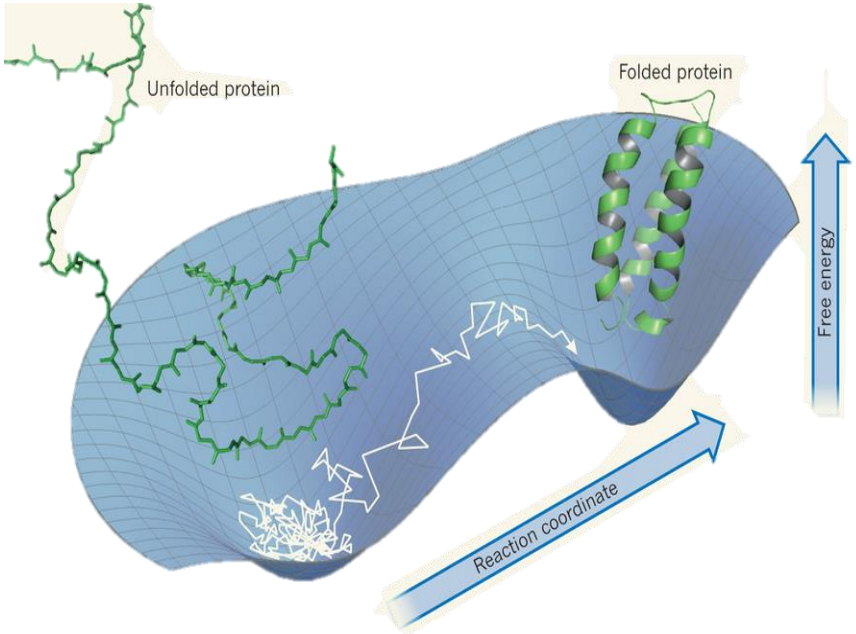
Why Direct Coupling Analysis?

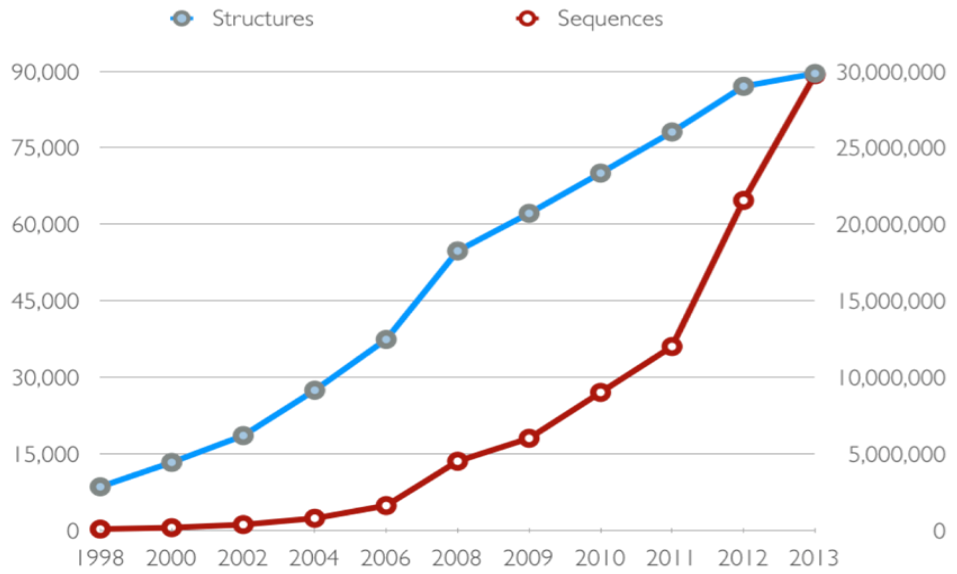
KTH/CSC

The sequencing revolution and sequence data explosion

Aalto University
School of Science
and Technology

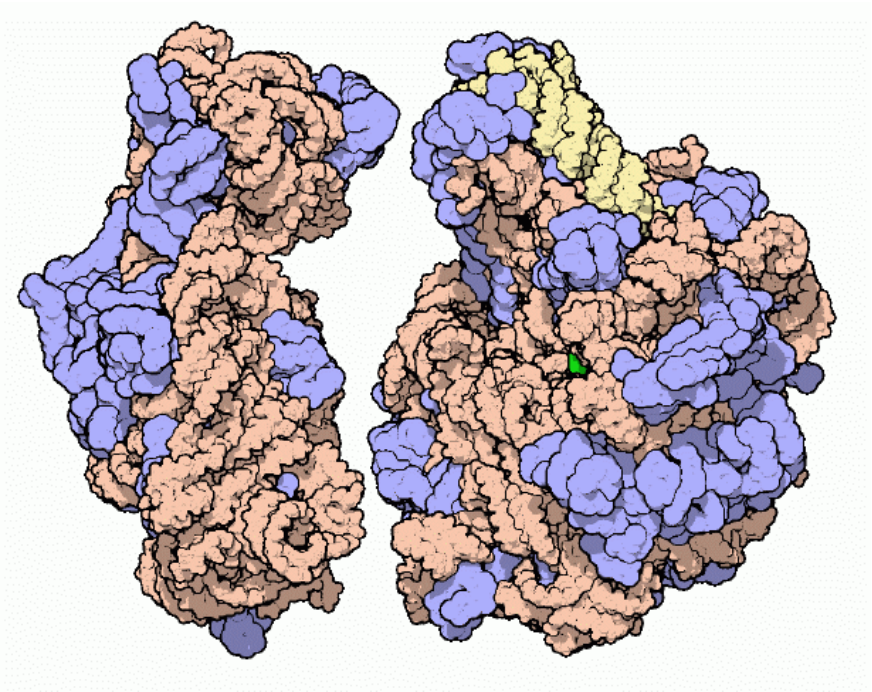
The number of protein sequences increases much faster than the number of solved protein structures. Folding proteins from one sequence *in silico* is hard – unless you have an already solved structure as template. Additional information from similar sequences with similar structures.

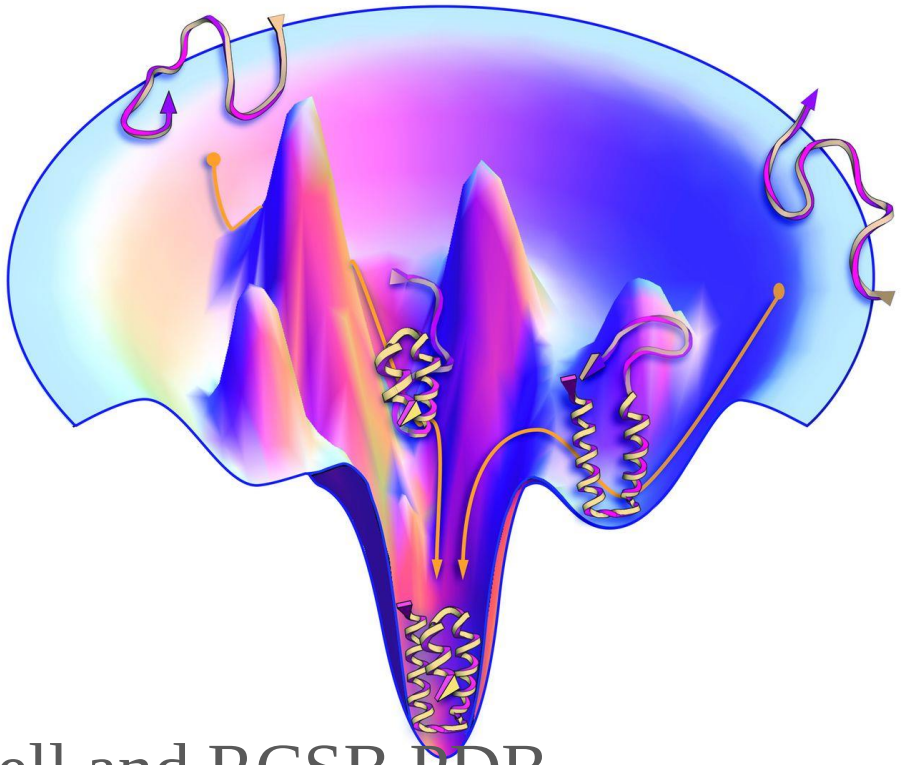




Can’t we just simulate ?

It appears not (except for small proteins)





Ribosome subunits: David Goodsell and RCSB PDB
Energy funnel: Ken Dill & Justin MacCallum

What is the state-of-the-art?

KTH/CSC

Comparative modeling

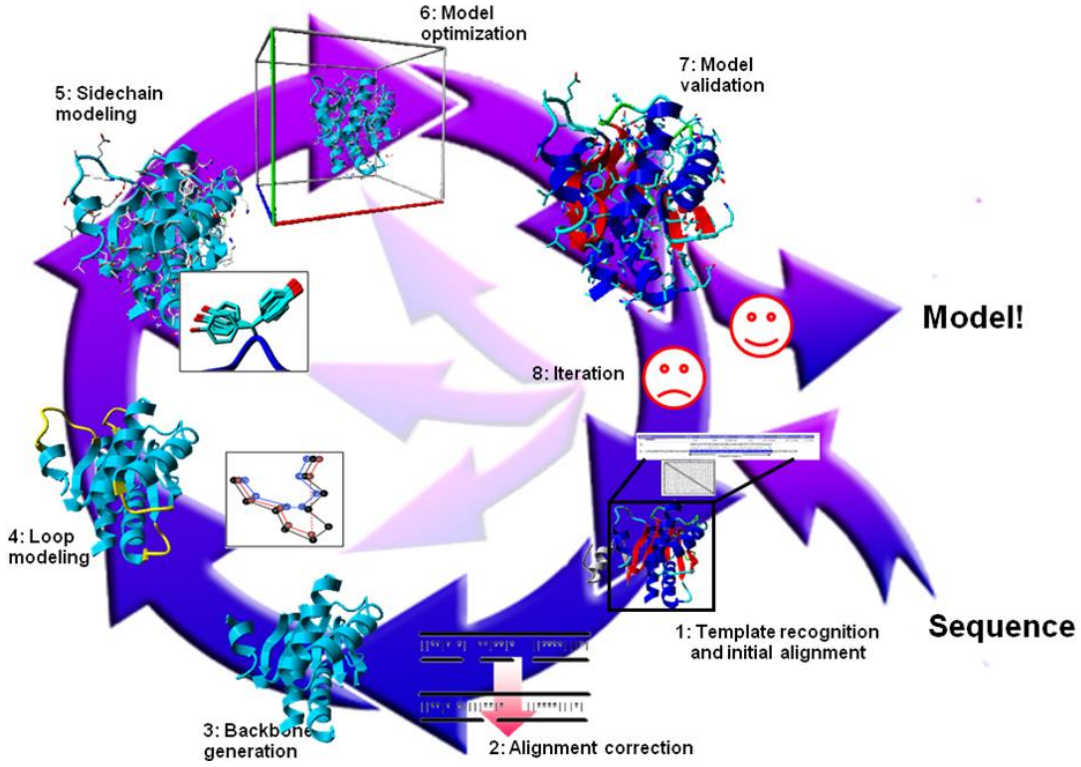


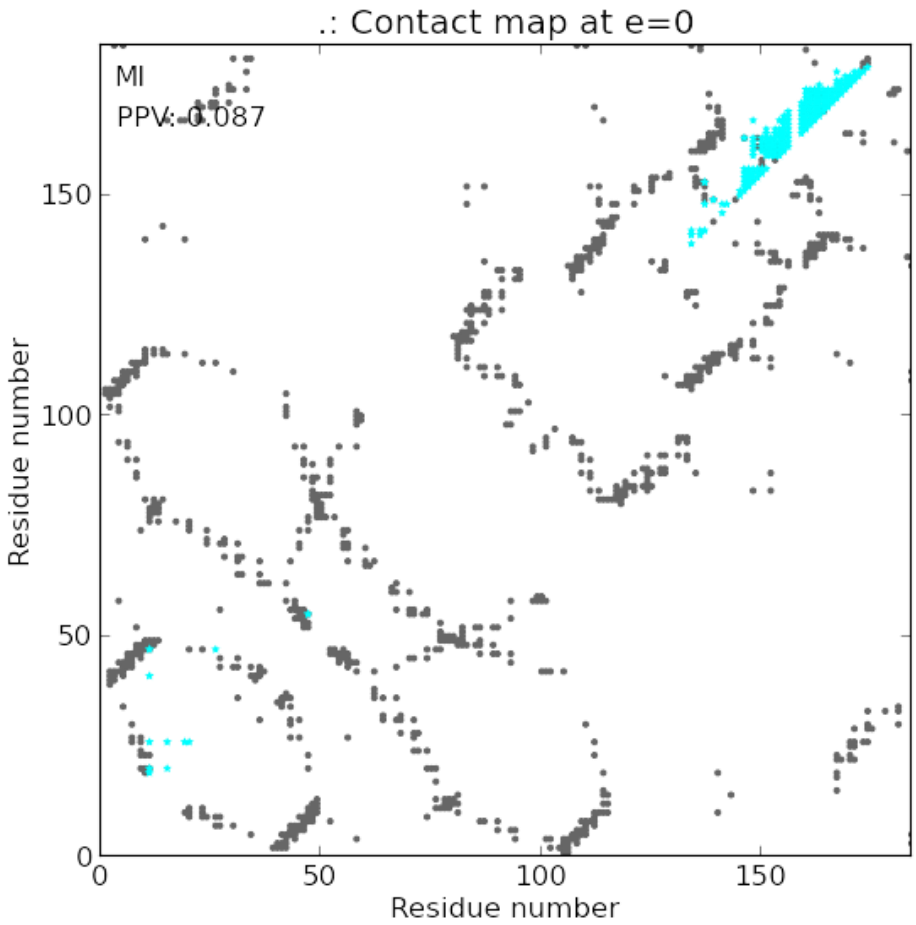
Image courtesy E. Krieger and G. Vriend

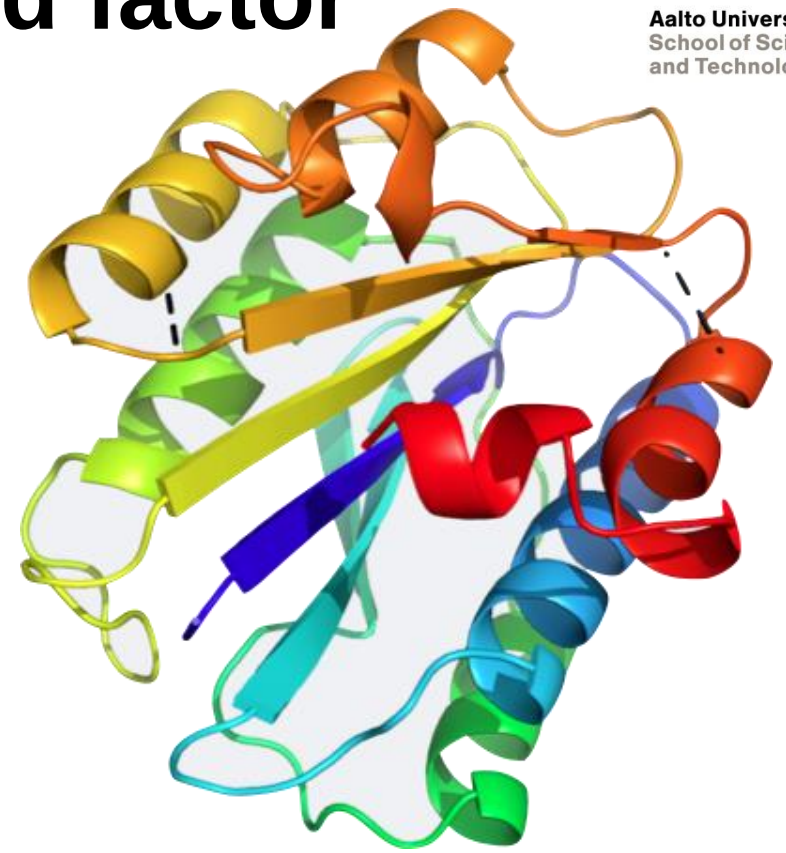
2. Examples. How well does DCA perform?

KTH/CSC

Aalto University
School of Science
and Technology

von Willebrand factor





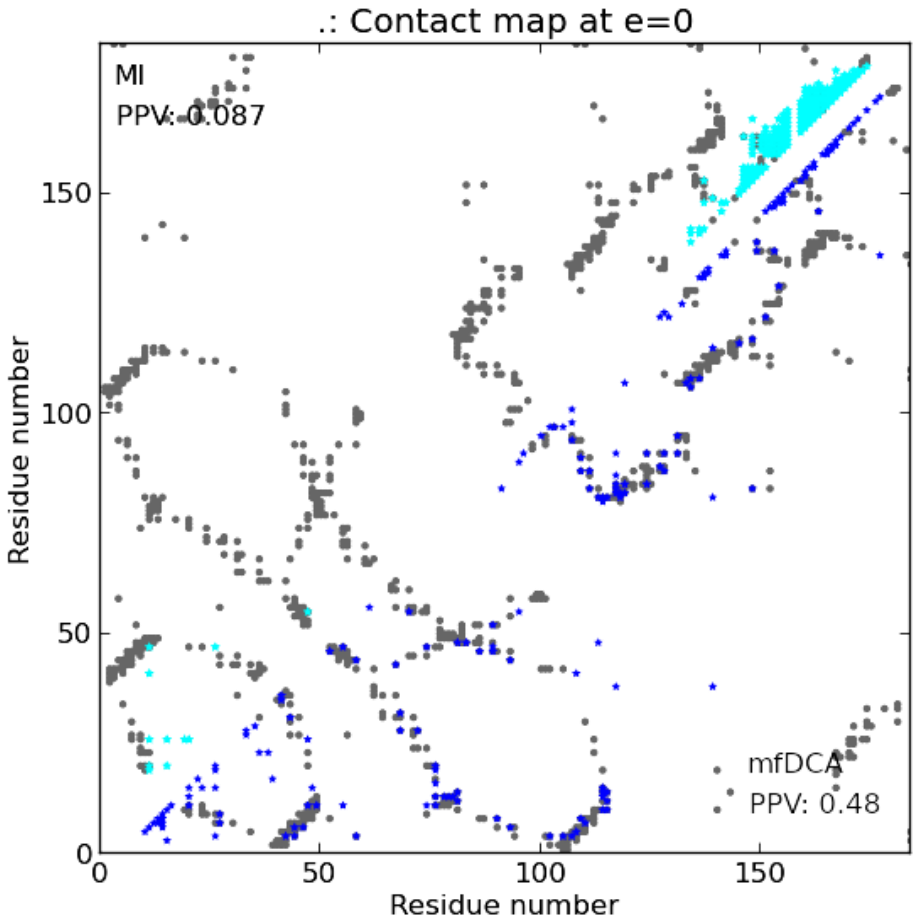
A3 domain of human von Willebrand factor (1atz), *courtesy* M Skwark 2013

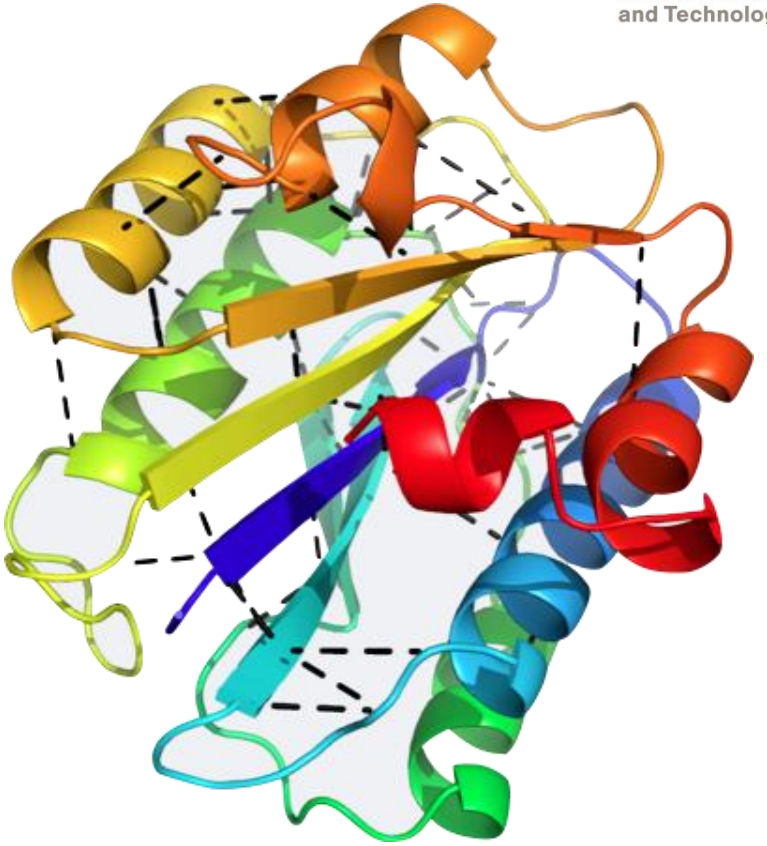
Fodor, A.A. and Aldrich, R.W. “Influence of conservation on calculations of amino acid covariance in multiple sequence alignments” (2004)

Correlations *vs* mfDCA

KTH/CSC

Aalto University
School of Science
and Technology





"Naïve mean-field" inverse Potts computation, *courtesy* M Skwark 2013

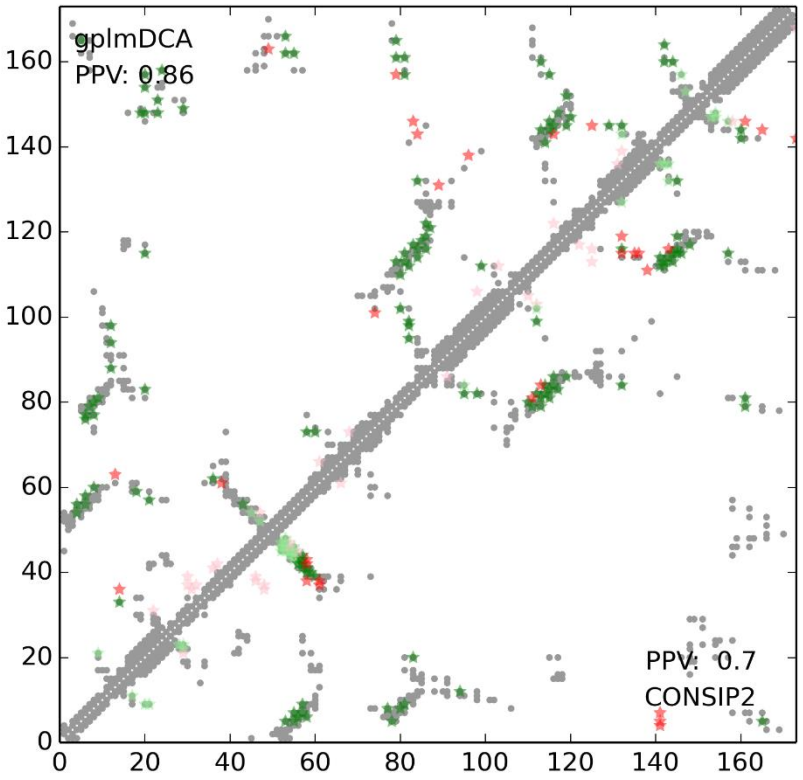
Morcos F. et al "Direct-coupling analysis of residue coevolution captures native contacts across many protein families" (PNAS 2011)

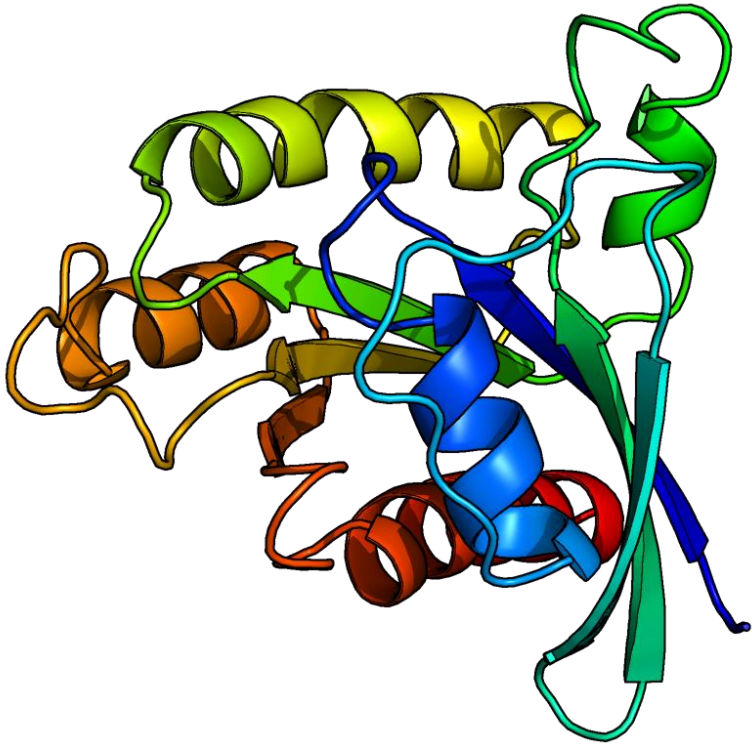
KTH/CSC

Aalto University
School of Science
and Technology

An example from CASP11 where two different DCAs seem to do well

gplmDCA *vs* CONSIP2/MetaPSICOV



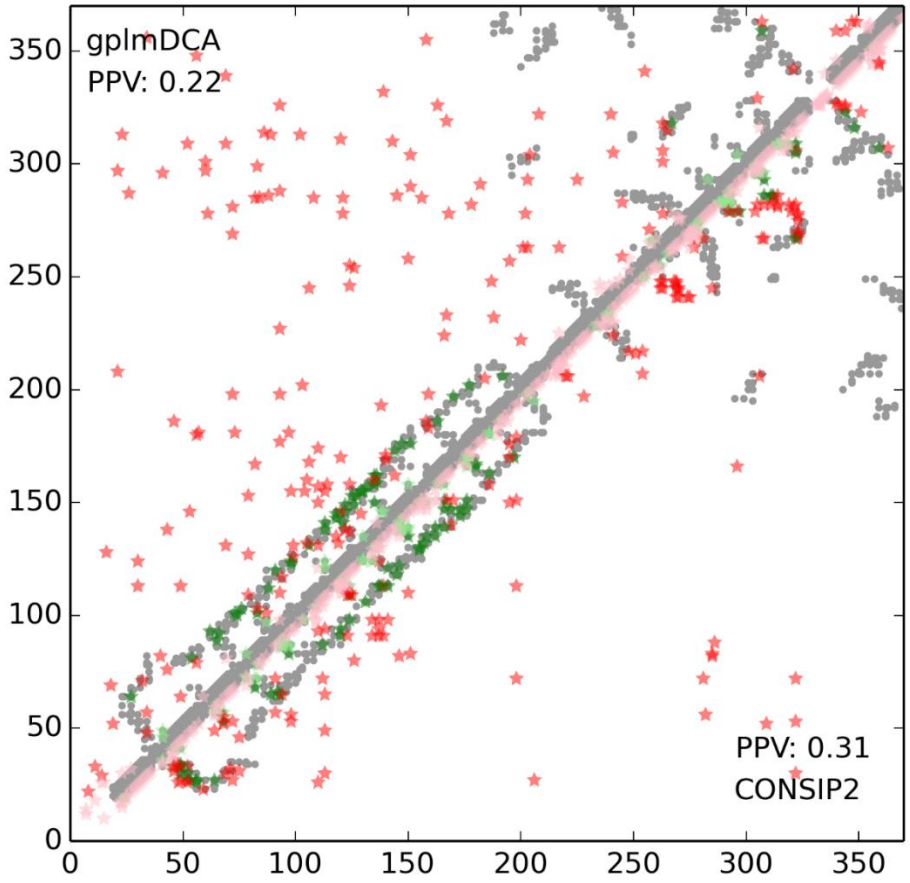


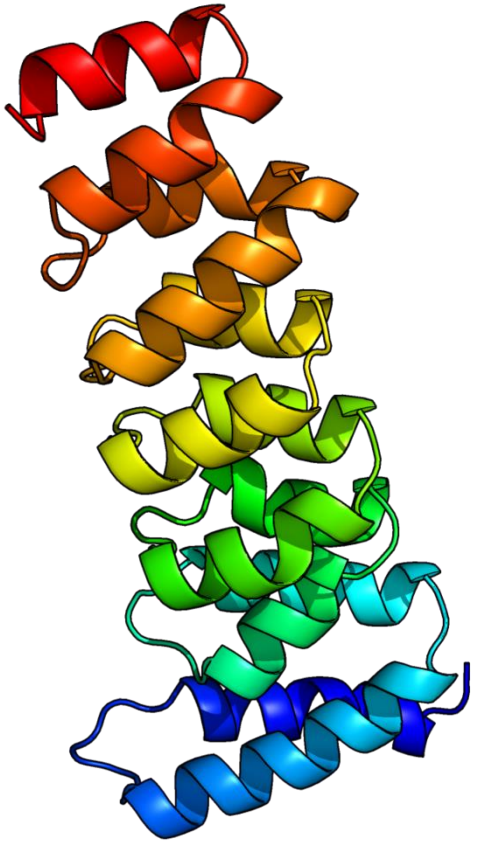
T0798: RAS11B a protein involved in membrane trafficking, 88253 sequences at 90%

KTH/CSC

Aalto University
School of Science
and Technology

Another example from CASP11 – less well predicted contacts...

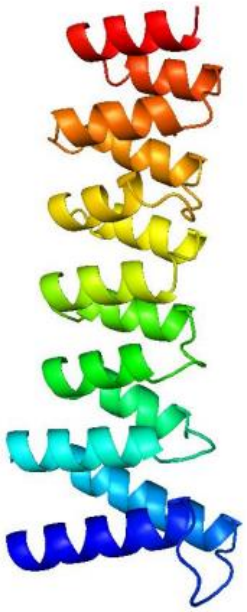




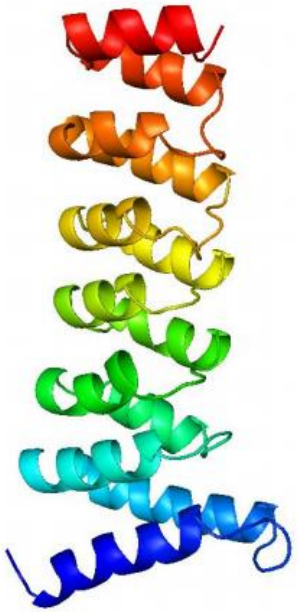
T0827: CC0948 in Caulobacter crescentus, 1834 sequences at 90%

...but still useful for structure...

Outstanding Predictions: T0827-D1



Best prediction
GDT-TS = 62.8
Group Skwark



Experimental structure

Median prediction
GDT-TS = 22.8

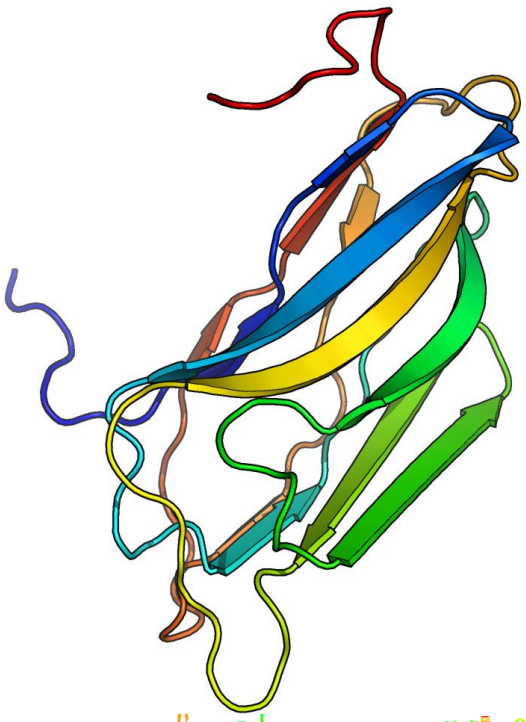
10 points better than 2nd place

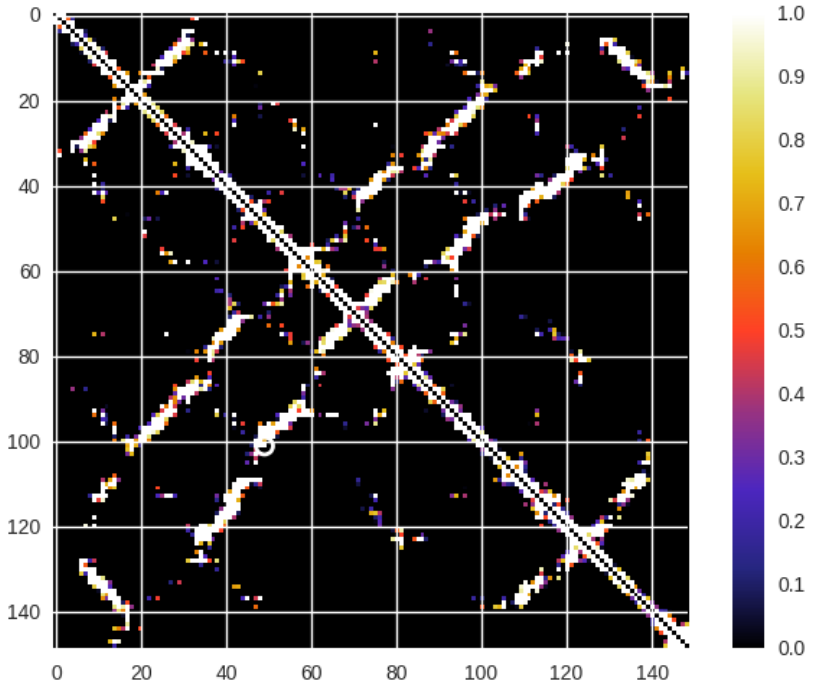
37

A CASP12 example (2016)

KTH/CSC

Aalto University
School of Science
and Technology





CASP12
T0921
Coh5

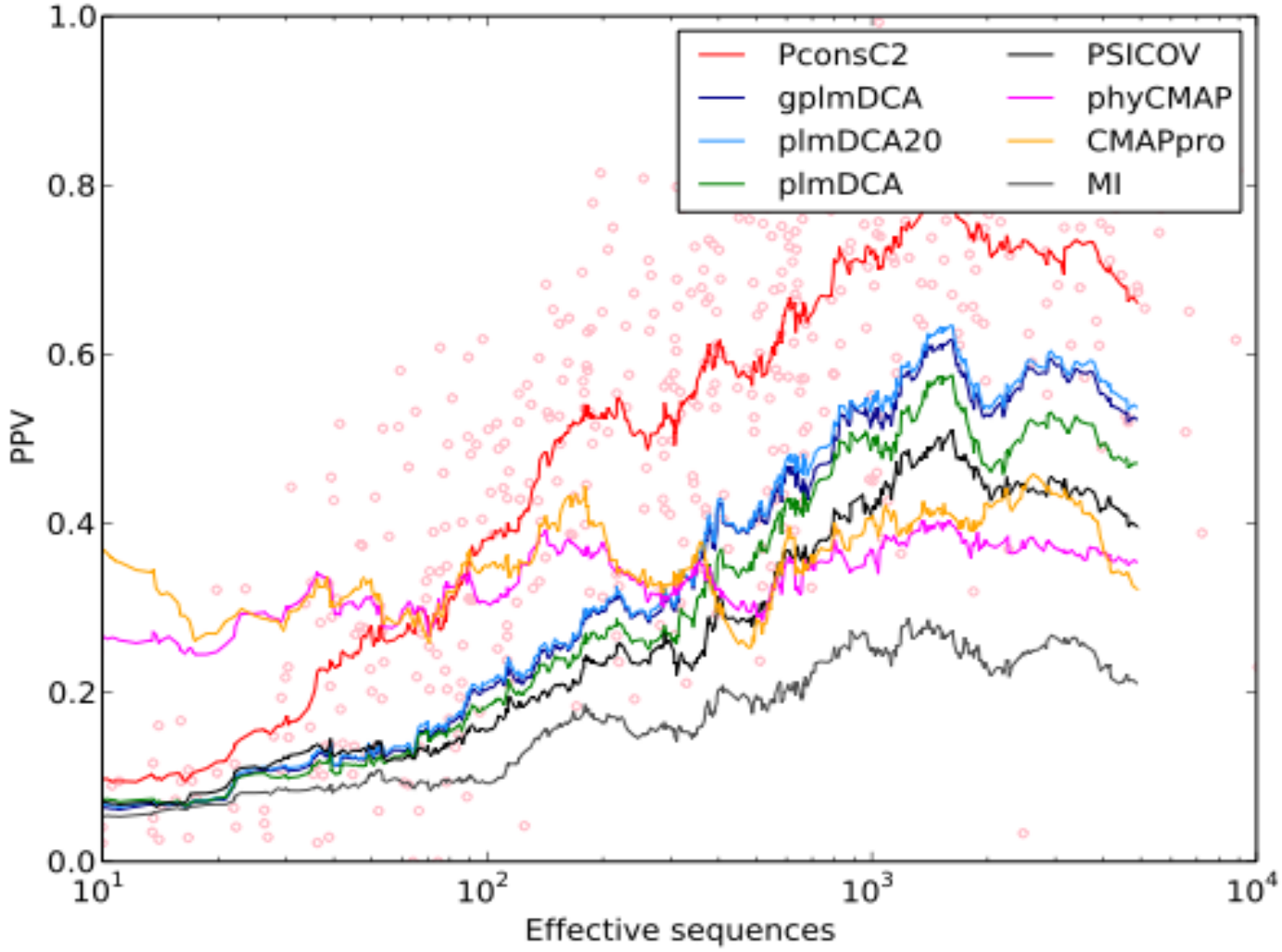
Cellulosomal scaffoldin protein

Ruminococcus flavefaciens

KTH/CSC

A State-of-the-Art from 2014

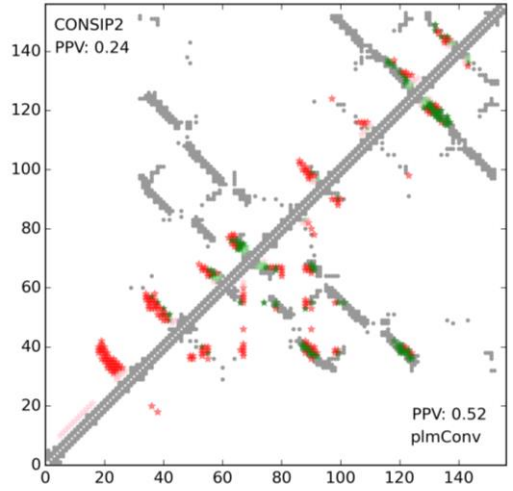
Aalto University
School of Science
and Technology

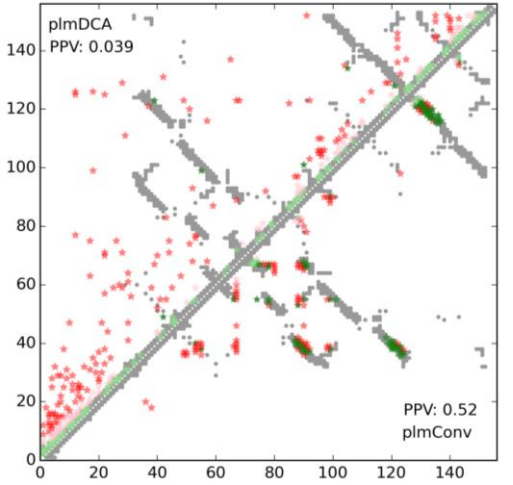


A State-of-the-Art from 2016

KTH/CSC

Aalto University
School of Science
and Technology



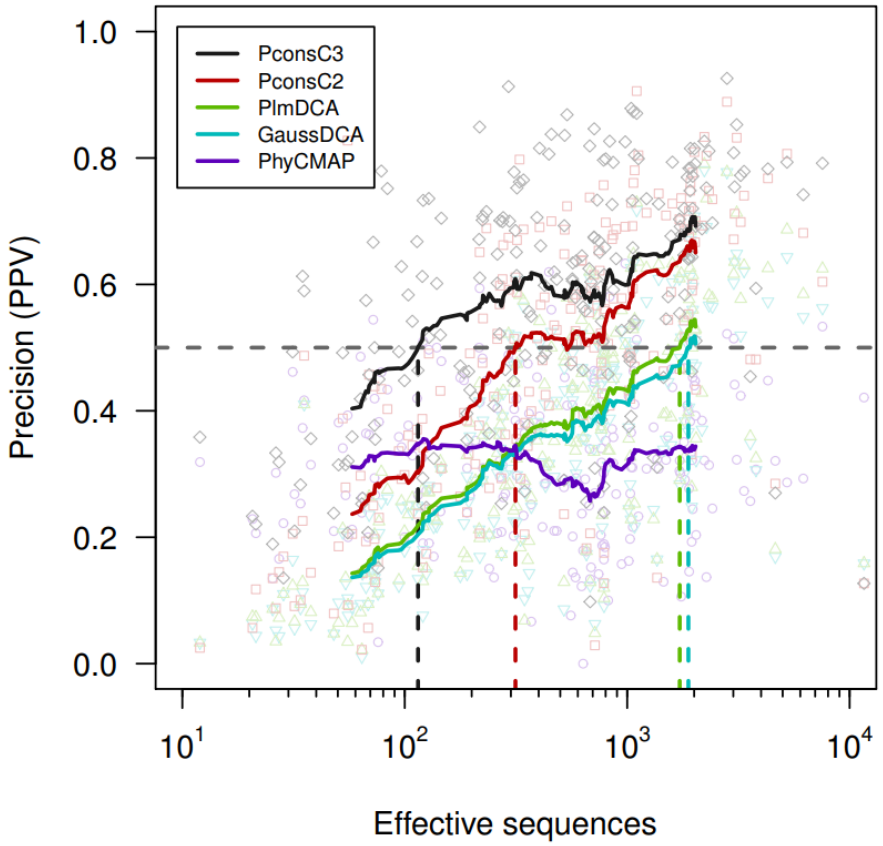


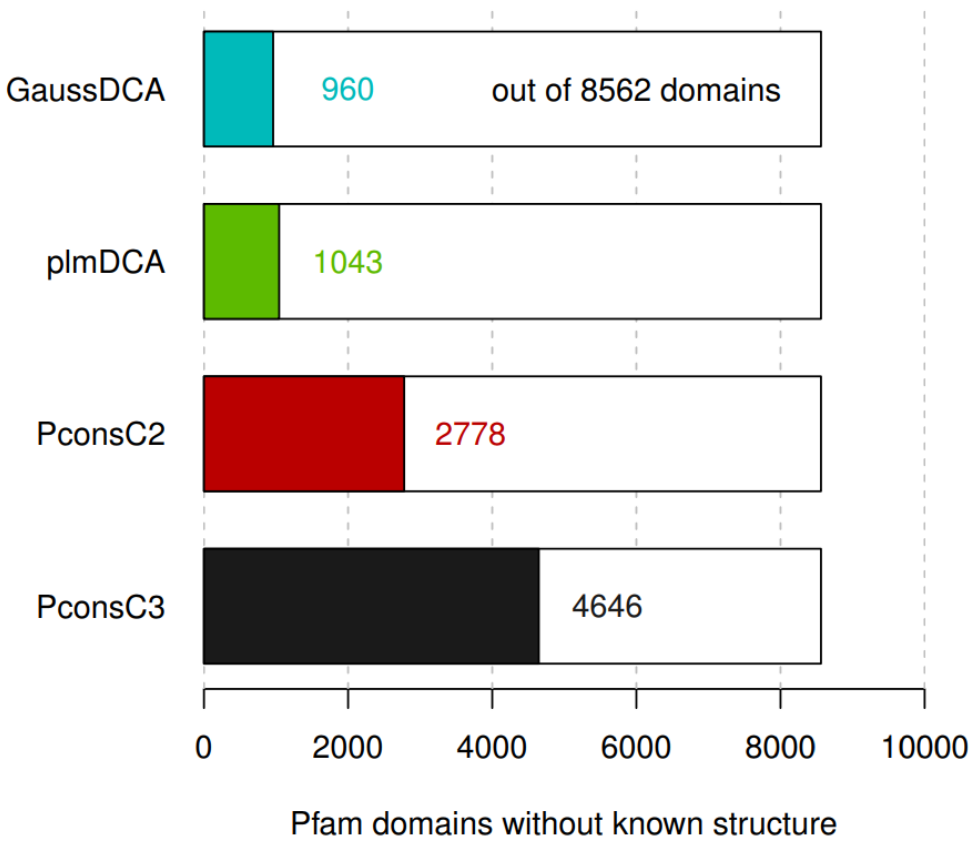
plmConv: Golkov, Skwark, Golkov, Dosovitskiy, Brox, Meiler, and Cremers, *NIPS* 2016

KTH/CSC

A State-of-the-Art from 2017

Aalto University
School of Science
and Technology





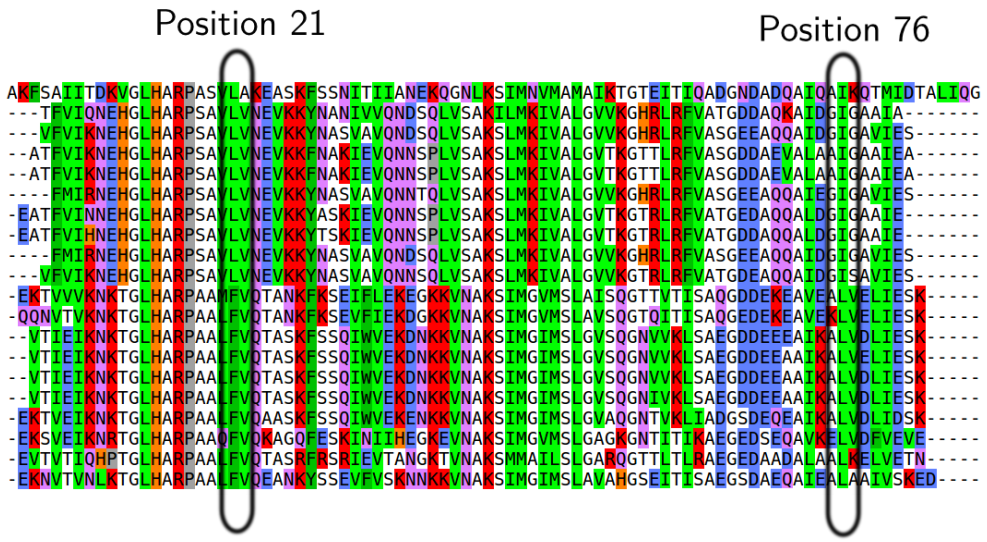
PconsC3: Skwark, Michel, Menendez Hurtado, Ekeberg, Elofsson, *Bioinformatics* **33**:2859-2866 (2017)

3. Methods. What is under the hood?

Data, data, data (and methods)

KTH/CSC

Aalto University
School of Science
and Technology



First papers used sequence data from the PFAM data base. Better performance from an unbiased set of homologous proteins (Feinauer et al, PLoS Comput Biol (2014)).

Maximum likelihood learning is computationally intractable. One must look for approximations to maximum likelihood, or to weaker learning criteria.

These problems are very under-sampled. Typically millions of parameters are learnt from tens of thousands of examples. Scoring and regularization.

Combination with other methods. This is the current state-of-the-art.

1st main method: mean-field

KTH/CSC

Aalto University
School of Science
and Technology

$$E(s) = \sum_i h_i S_i + \sum_{ij} J_{ij} S_i S_j \qquad P^{\text{trial}}(s) = \Pi_i P_i(S_i)$$

$$F^{nMF} = \sum_i H\left(\frac{1+m_2}{s}\right) + H\left(\frac{1-m_2}{s}\right) + \sum_i h_i m_i + \sum_{ij} J_{ij} m_i m_j \qquad H(x) = -x \log x$$

$$\frac{\partial F^{nMF}}{\partial m_i} = 0 \quad \Longrightarrow \quad m_i = \tanh(h_i^{nMF} + \sum_j J_{ij} m_j)$$

$$\chi_{ij} = \frac{\partial m_i}{\partial h_j} = c_{ij}$$

Exact, a fluctuation-dissipation relation. An immediate result for pairwise exponential models.

$$\left(\chi^{nMF}\right)^{-1}_{ij} = \frac{\partial h_i^{nMF}}{\partial m_j} \approx \left(c^{-1}\right)_{ij} \quad \Longrightarrow \quad \left(c^{-1}\right)_{ij} \approx \frac{1}{1-m_i^2} 1_{ij} - J_{ij}$$

Use in DCA: Weigt et al (2009), Morcos et al (2011) + many later contributions
Theory in Kappen & Rodriguez, 1998, Kappen & Spanjers, 2001, F Ricci-Tersenghi, 2013

KTH/CSC

2nd main method: pseudo-likelihood maximization

Aalto University
School of Science and Technology

Maximum likelihood $P(\mathbf{S}) = \frac{1}{Z(\mathbf{h},\mathbf{J})} \exp\left(\sum_i h_i S_i + \sum_{ij} J_{ij} S_i S_j \right)$

$$\Pr(\mathbf{S}^{(1)},\ldots,\mathbf{S}^{(n)};\mathbf{h},\mathbf{J}) = P(\mathbf{S}^{(1)};\mathbf{h},\mathbf{J}) \cdots P(\mathbf{S}^{(n)};\mathbf{h},\mathbf{J})$$

$$\mathbf{h}^*, \mathbf{J}^* \in \arg\max \left[\sum_{ij} h_i \frac{1}{n} \sum_{s=1}^{n} x_i^{(s)} + \sum_{ij} J_{ij} \frac{1}{n} \sum_{s=1}^{n} x_i^{(s)} x_j^{(s)} - \log Z(\mathbf{h},\mathbf{J}) \right]$$

Pseudo-maximum likelihood (avoids computing Z):

$$P(S_r \mid S_{\setminus r}) = \exp\left(h_r S_r + \sum_l J_{rl} S_r S_l \right) \Big/ \sum_y \exp\left(h_r y + \sum_l J_{rl} y S_l \right)$$

$$h_r^{plm}, J_{rl}^{plm} \in \arg\max \left[\sum_{ij} h_i \frac{1}{n} \sum_{s=1}^{n} x_i^{(s)} + \sum_{ij} J_{ij} \frac{1}{n} \sum_{s=1}^{n} x_i^{(s)} x_j^{(s)} - f(h_r, J_{rl}, S_{\setminus r}) \right]$$

Besag (1974), Wainwright-Ravikumar-Lafferty (2010)

Ekeberg et al *Phys. Rev. E* (2013) + github.com/magnusekeberg/plmDCA

Regularization and scoring

KTH/CSC

Aalto University
School of Science
and Technology

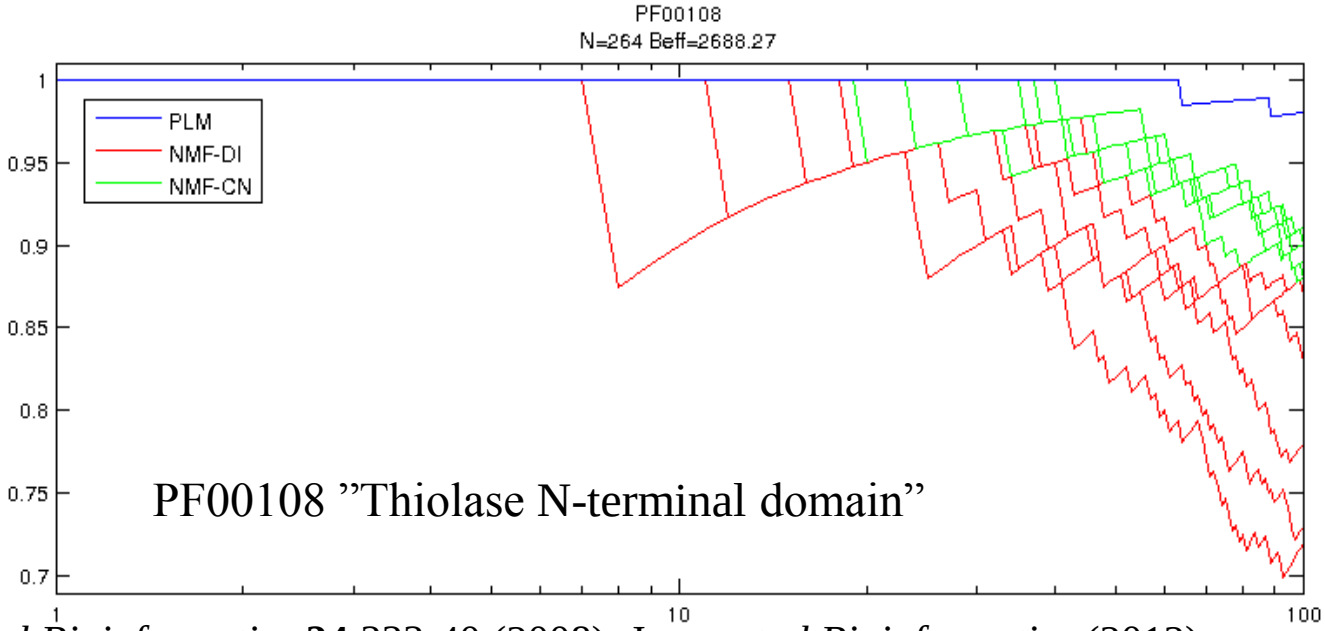
In mean-field the regularization is through pseudo-counts, in pseudo-likelihood by a simple L_2-regularizatio. The outcome is then matrices of coupling strengths (J's). How to score them? Weigt et al (2009) and Morcos et al (2011) scored by the mutual information of the direct interaction, the model restricted to two variables. Better performance is obtained by scoring by a *corrected norm* (CN).

Start from the regularized norm of the J's:

$$FN_{ij} = \left\| J'_{ij} \right\|_2$$

Then "correct" the norm

$$CN_{ij} = FN_{ij} - \frac{FN_{\bullet j} FN_{i\bullet}}{FN_{\bullet\bullet}}$$



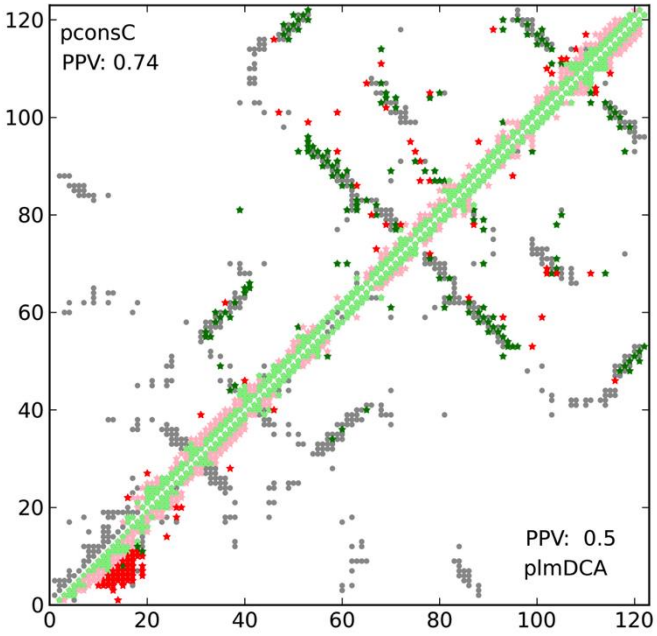
CN introduced/used in Dunn *et al Bioinformatics* **24**:333-40 (2008); Jones *et al Bioinformatics* (2012)
Detailed presentation for DCA in Ekeberg *et al Journal of Computational Physics* **276**, 341-356 (2014)

KTH/CSC

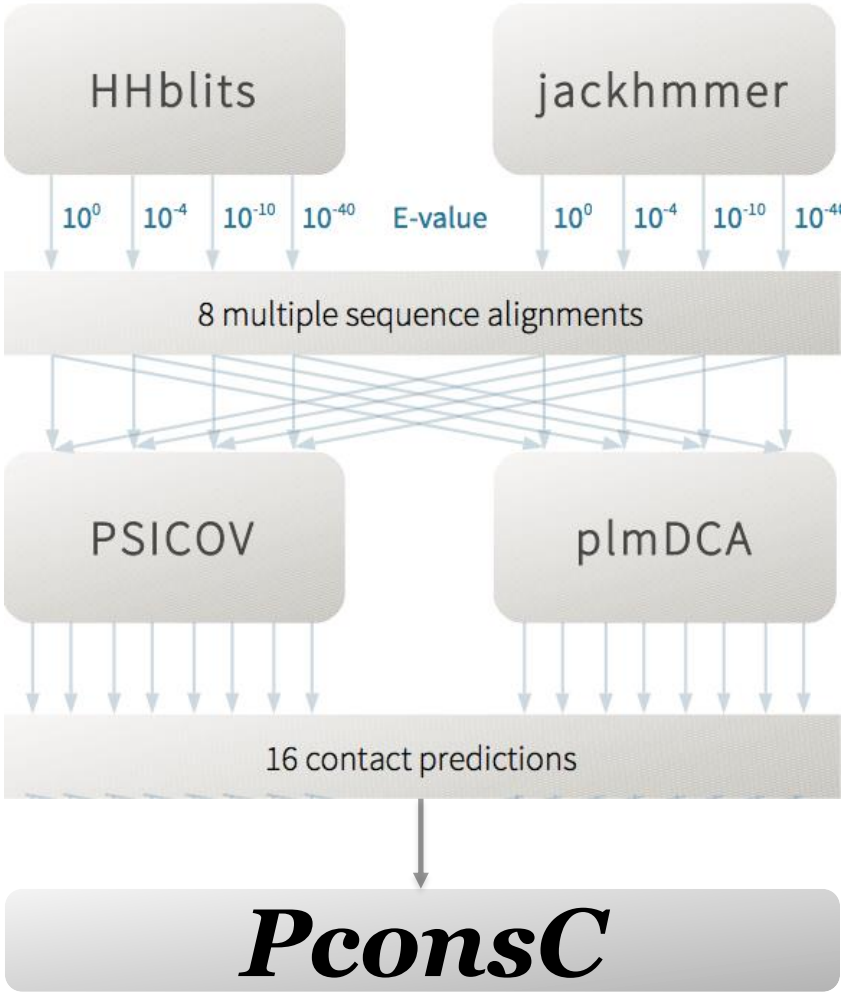
3rd main method: machine-learning by ”pooling” predictions...

Aalto University
School of Science
and Technology

A machine learning method combining different alignment sources and inference schemes

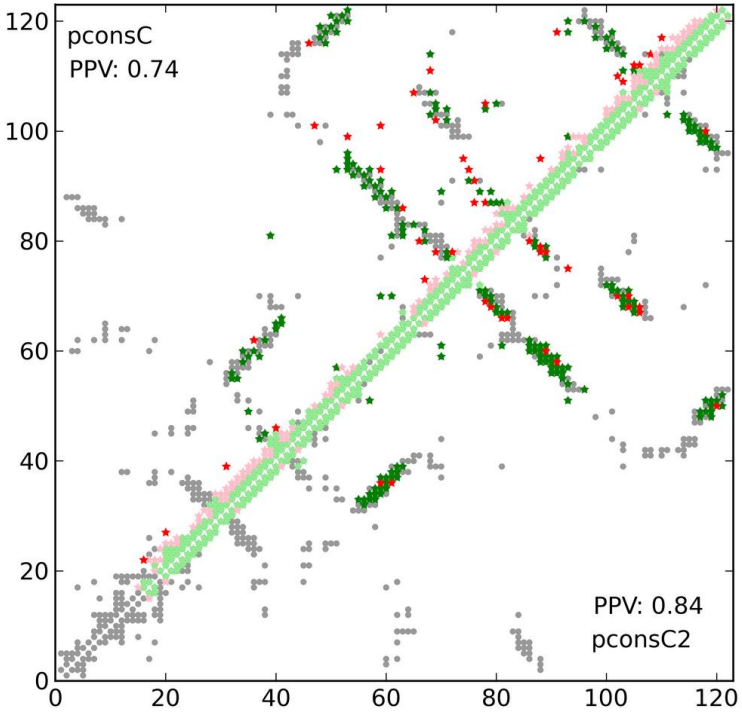


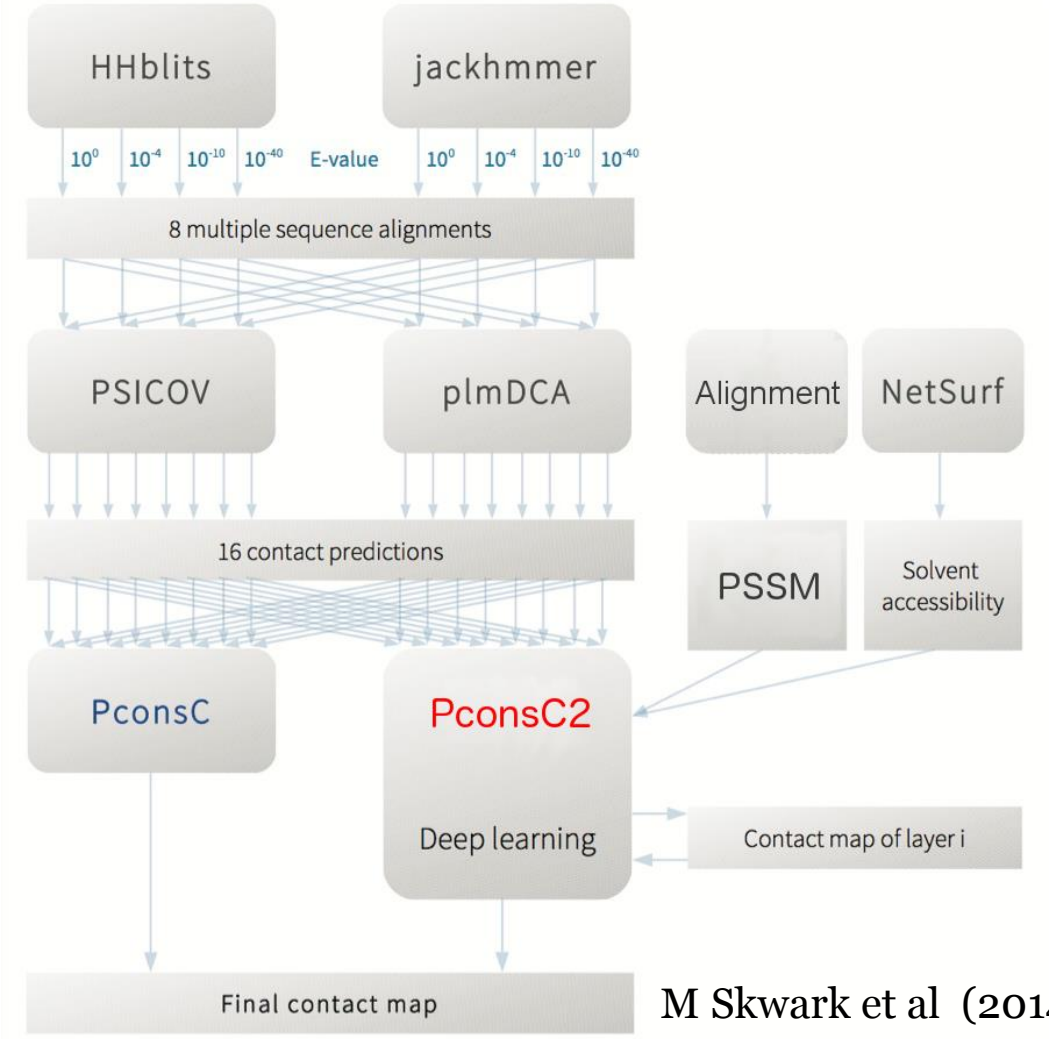
M Skwark et al, Bioinformatics (2013)



...recently much improved; also by..

A more advanced machine learning method combining DCA with information on solvent accessibility and secondary structure





M Skwark et al (2014)

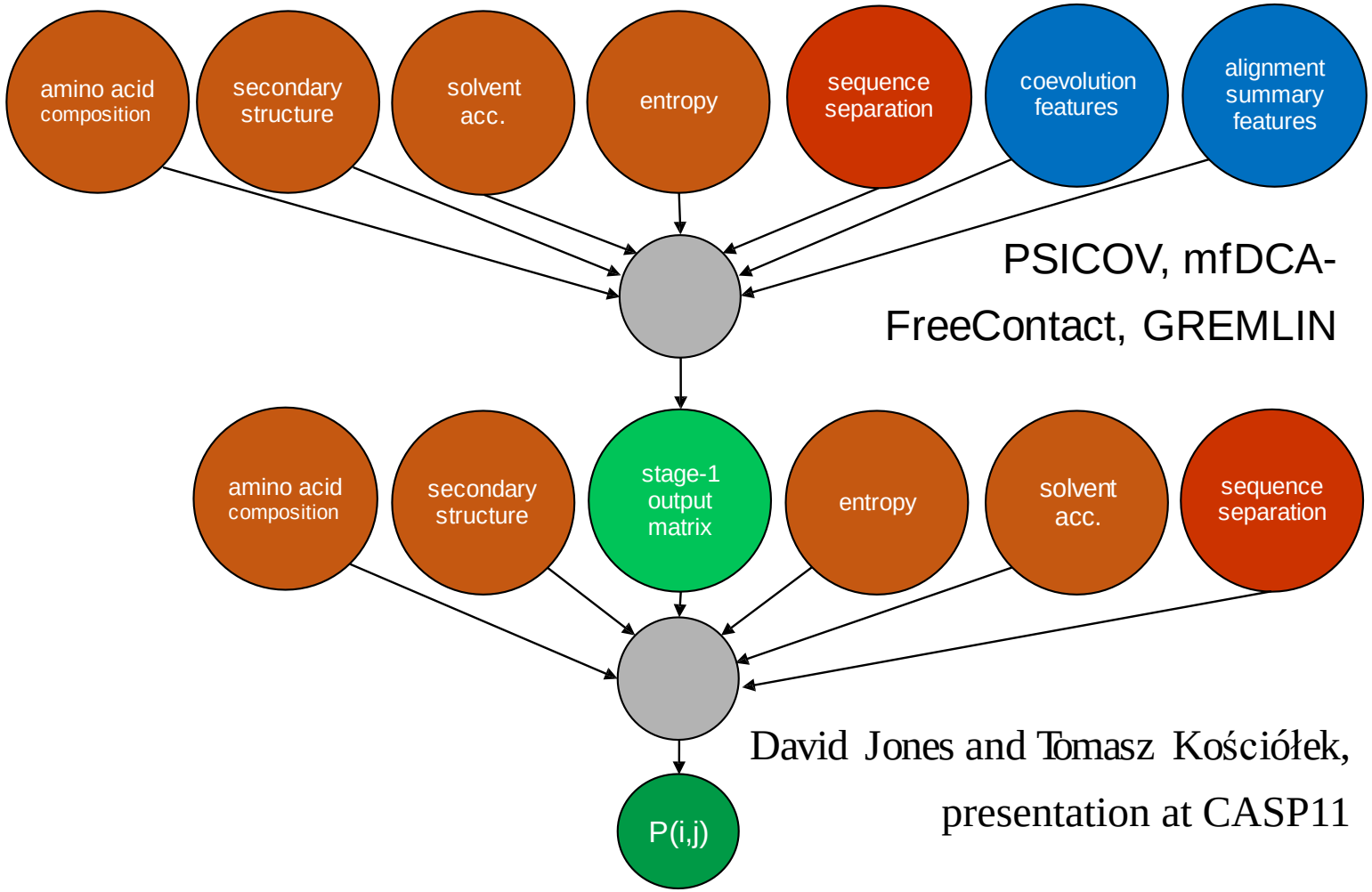
KTH/CSC

Aalto University
School of Science and Technology

...CONSIP2 / MetaPSICOV, which won contact prediction at CASP11

Stage 1
672 features

55 hidden units



Stage 2
731 features

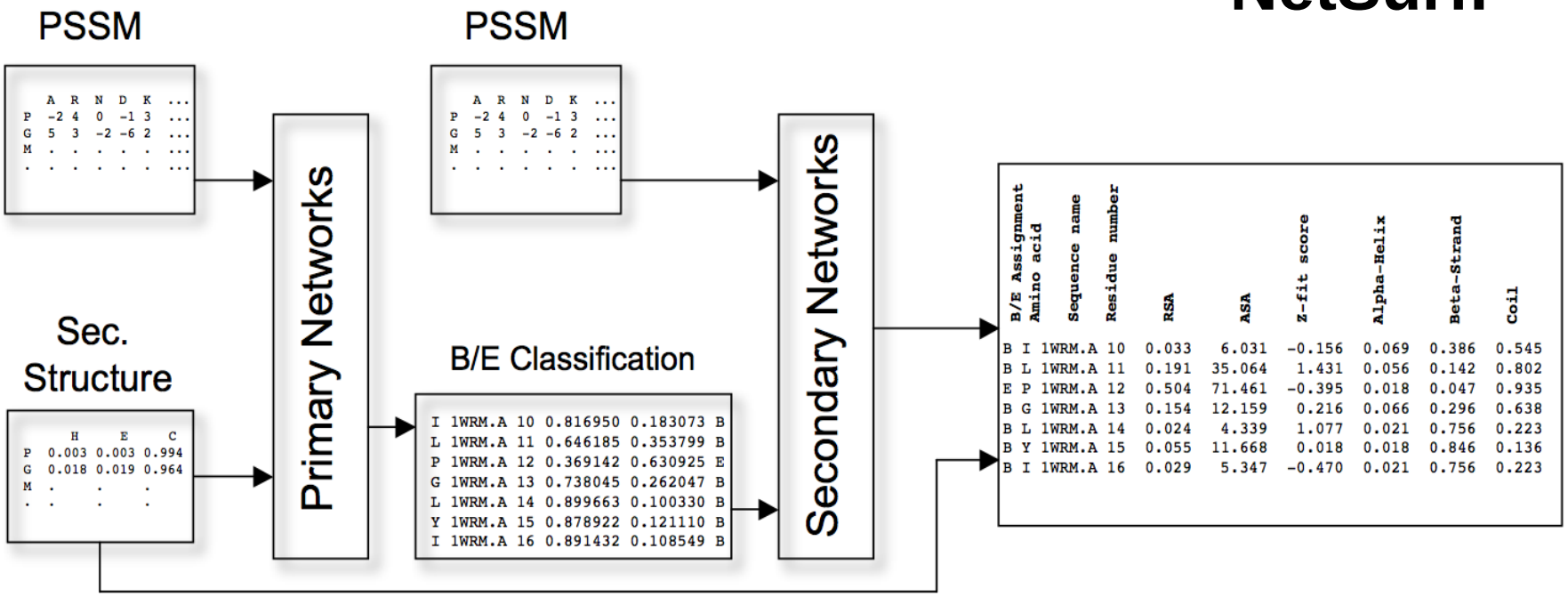
55 hidden units

KTH/CSC

To make the point on other information: solvent accessibility

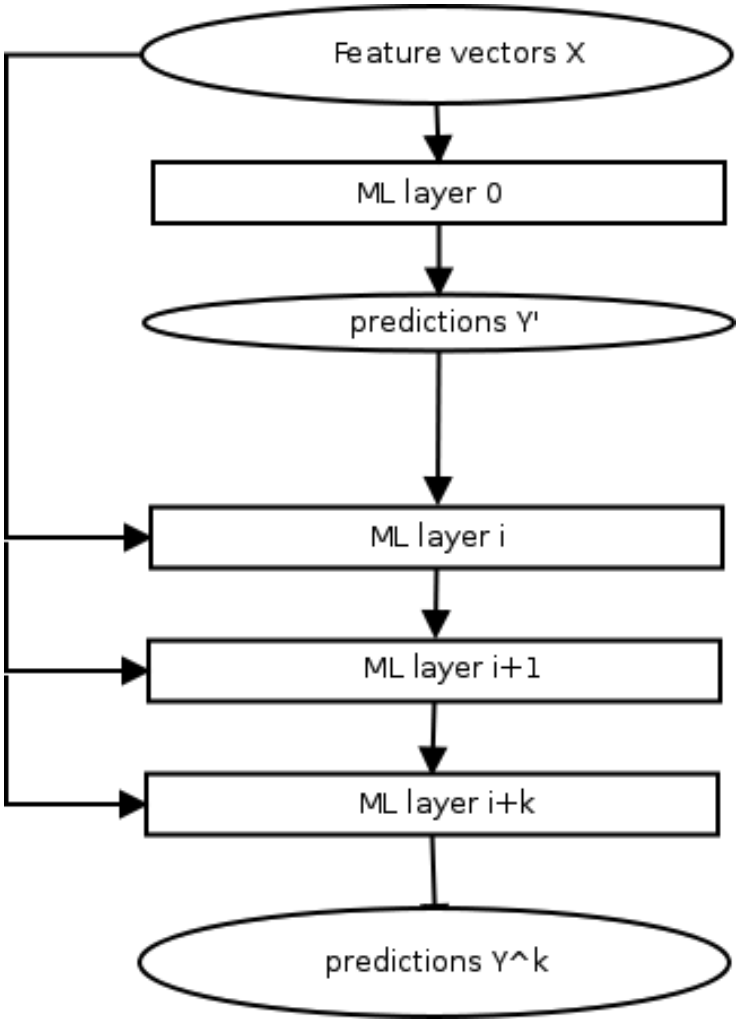
Aalto University
School of Science
and Technology

NetSurfP



Petersen et al. *BMC Structural Biology* (2009) doi:10.1186/1472-6807-9-51

The point again: ”deep learning”



Predicting structures from sequences is currently best done using deep learning.

Do not worry: these methods use the output of one or more DCA schemes as input to a total predictor.

We are not out of business: better DCAs gives better total predictors. After PconsC2 there will be PconsC3.

But we are now far from a pure DCA approach, and the other input is also important.

KTH/CSC

Thanks to

Marcin Skwark

Vetenskapsrådet

ACADEMY OF FINLAND

Magnus Ekeberg

Aalto University

UNIVERSITY OF HELSINKI

COIN

Finnish Centre of Excellence in Computational Inference Research

Martin Weigt

Christoph Feinauer

Andrea Pagnani

中国科学院

CHINESE ACADEMY OF SCIENCES