Using residue coevolution to retrieve protein homologs[1]

ComPotts

Hugo Talibart, François Coste

Co-evolutionary methods for the prediction and design of protein structure and interactions

CECAM-HQ-EPFL, June 18, 2019

[1] work in progress

The Dyliss bioinformatics team

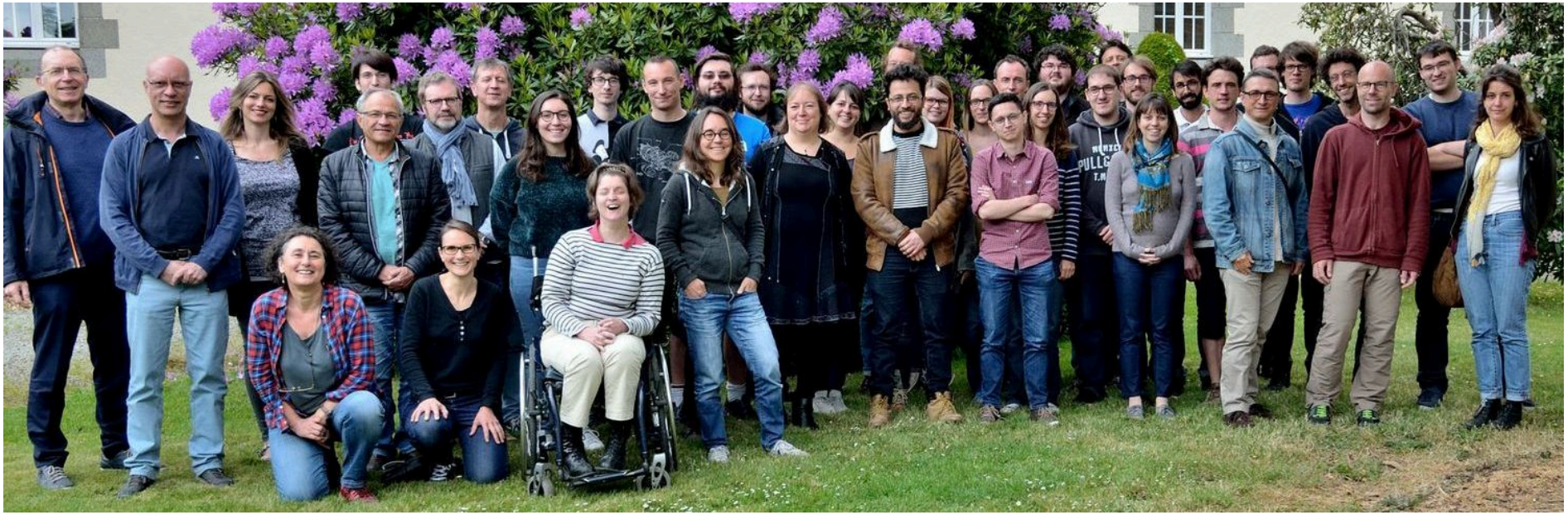
http://www.irisa.fr/dyliss

DyLiSS

informatiques mathématiques

UMR IRISA

Bioinformatics at IRISA / Inria Rennes



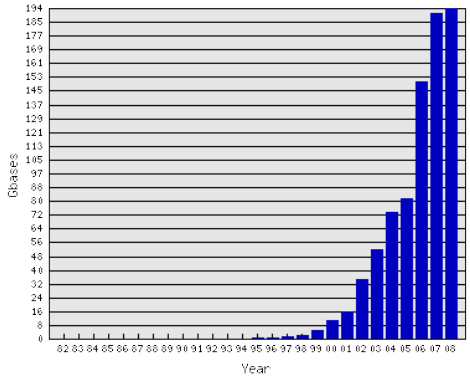
- Symbiose (bioinformatics Irisa/Inria Rennes):
  - Dyliss research team
  - Genscale research team
  - Genouest bioinformatics platform

    Seminars: `http://symbiose.irisa.fr/symbioseSeminars`
- Biogenouest western France life science and environment network
  Marine biology, agriculture/food-processing, human health, and bioinformatics.

Motivation

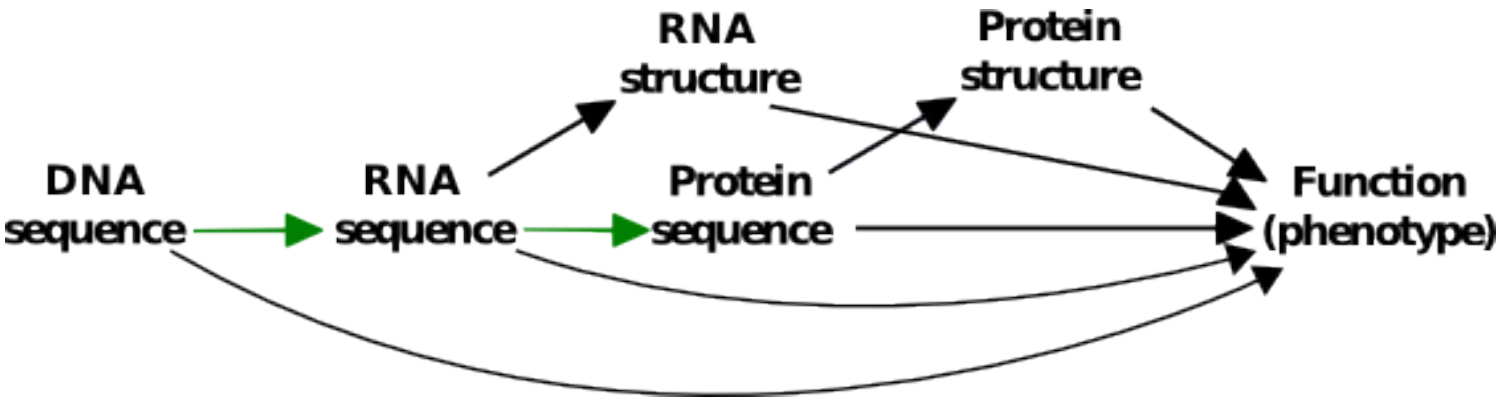
Sequences annotation problem

High throughput production of raw sequences



Problem

Function(s) of these sequences ?



Protein function?

- *In-vivo / in-vitro* experiments
  Especially on model organisms:
  - Gene knockout and others mutations
    $\rightarrow$ key sequence(s) for a function
  - Structure determination
    $\rightarrow$ key positions for a function
  - ...

Does not scale well...

- To face the (ever-increasing) amount of available sequences, automatic methods are needed $\rightsquigarrow$ *in-silico* functional or structural predictions.

Classical approach to predict the function of a new gene sequence

Search for annotated homologs ...

Retrieve the homologs of a protein gene

Search for a significant match with:

- an (already annotated) protein sequence, e.g. with BLAST[2]

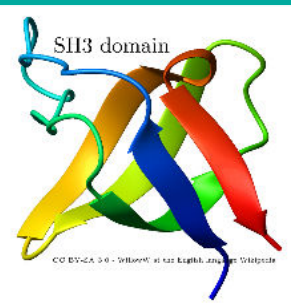
```
>1shg:A
AKELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
```

[2] S. F. Altschul et al. "Basic local alignment search tool". *Journal of molecular biology* (1990).

[3] S. R. Eddy. "Profile hidden Markov models.". *Bioinformatics* 14.9 (1998), pp. 755–763.

[4] M. Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation". *bioRxiv* (2019), p. 560029.

Retrieve the homologs of a protein gene



Search for a significant match with:

- an (already annotated) protein sequence, e.g. with BLAST[2]

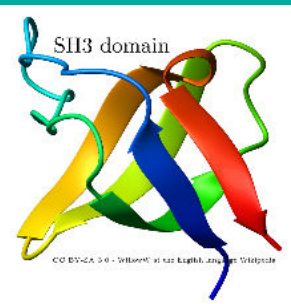
```
>1shg:A
AKELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
```

[2] S. F. Altschul et al. "Basic local alignment search tool". *Journal of molecular biology* (1990).

[3] S. R. Eddy. "Profile hidden Markov models.". *Bioinformatics* 14.9 (1998), pp. 755–763.

[4] M. Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation". *bioRxiv* (2019), p. 560029.

Retrieve the homologs of a protein gene



Search for a significant match with:

- an (already annotated) protein sequence, e.g. with BLAST[2]

```
>1shg:A
AKELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
```

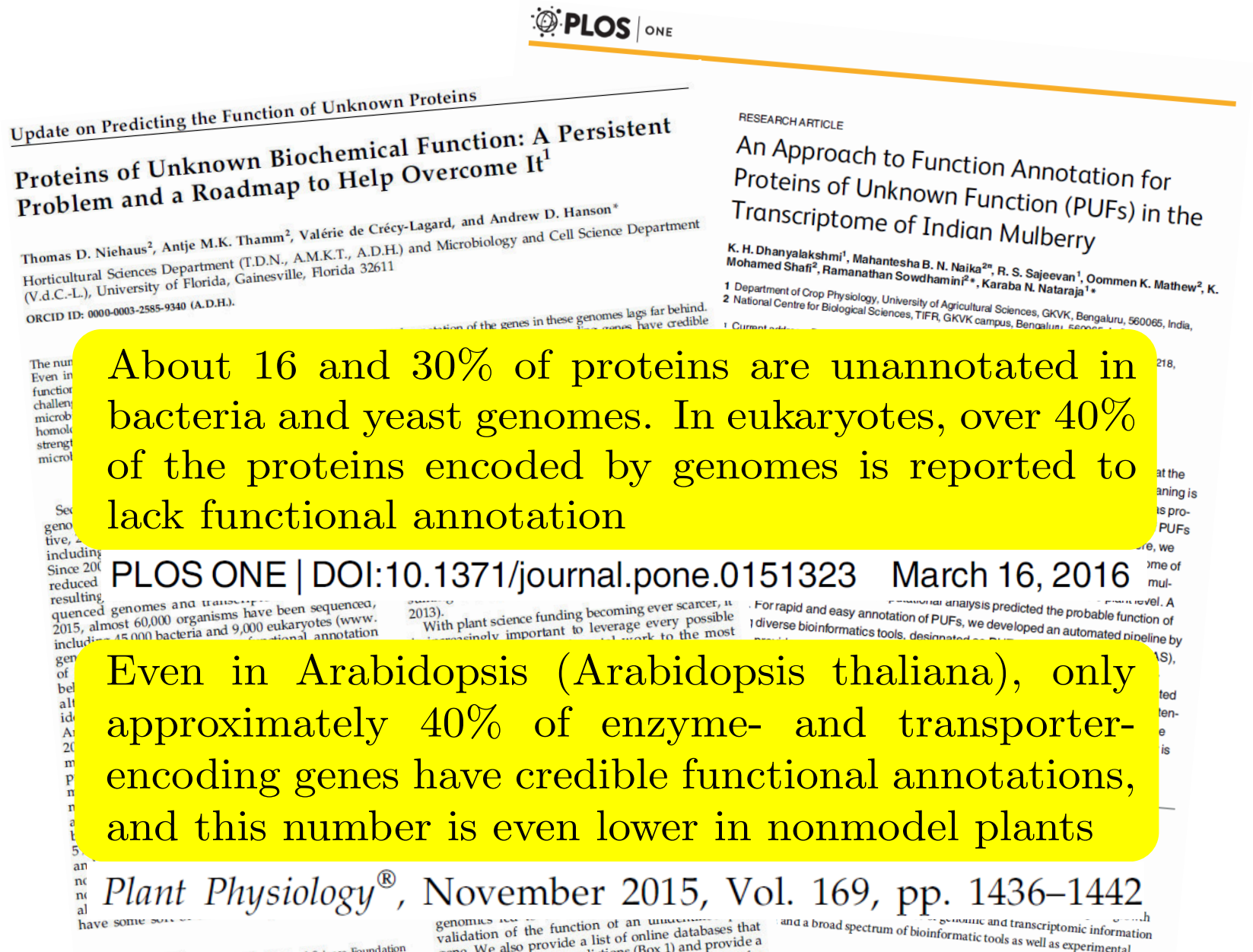
- the Profile HMM of a protein family, e.g. with HMMER[3] or HH-suite[4]

[2] S. F. Altschul et al. "Basic local alignment search tool". *Journal of molecular biology* (1990).

[3] S. R. Eddy. "Profile hidden Markov models.". *Bioinformatics* 14.9 (1998), pp. 755–763.

[4] M. Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation". *bioRxiv* (2019), p. 560029.

Yet. . .



PLOS | ONE

Update on Predicting the Function of Unknown Proteins

Proteins of Unknown Biochemical Function: A Persistent Problem and a Roadmap to Help Overcome It[1]

Thomas D. Niehaus[2], Antje M.K. Thamm[2], Valérie de Crécy-Lagard, and Andrew D. Hanson*

Horticultural Sciences Department (T.D.N., A.M.K.T., A.D.H.) and Microbiology and Cell Science Department (V.d.C.-L.), University of Florida, Gainesville, Florida 32611

ORCID ID: 0000-0003-2585-9340 (A.D.H.).

RESEARCH ARTICLE

An Approach to Function Annotation for Proteins of Unknown Function (PUFs) in the Transcriptome of Indian Mulberry

K. H. Dhanyalakshmi[1], Mahantesha B. N. Naika[2¤], R. S. Sajeevan[1], Oommen K. Mathew[2], K. Mohamed Shafi[2], Ramanathan Sowdhamini[2]*, Karaba N. Nataraja[1]*

1 Department of Crop Physiology, University of Agricultural Sciences, GKVK, Bengaluru, 560065, India,
2 National Centre for Biological Sciences, TIFR, GKVK campus, Bengaluru,

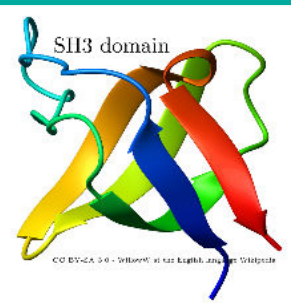
About 16 and 30% of proteins are unannotated in bacteria and yeast genomes. In eukaryotes, over 40% of the proteins encoded by genomes is reported to lack functional annotation

PLOS ONE | DOI:10.1371/journal.pone.0151323 March 16, 2016

Even in Arabidopsis (Arabidopsis thaliana), only approximately 40% of enzyme- and transporter-encoding genes have credible functional annotations, and this number is even lower in nonmodel plants

Plant Physiology®, November 2015, Vol. 169, pp. 1436–1442

Retrieve the homologs of a protein gene



Search for a significant match with:

- an (already annotated) protein sequence, e.g. with BLAST[2]

```
>1shg:A
AKELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
```

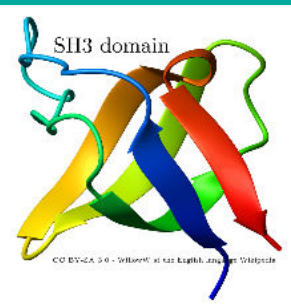
- the Profile HMM of a protein family, e.g. with HMMER[3] or HH-suite[4]

[2] S. F. Altschul et al. "Basic local alignment search tool". *Journal of molecular biology* (1990).

[3] S. R. Eddy. "Profile hidden Markov models.". *Bioinformatics* 14.9 (1998), pp. 755–763.

[4] M. Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation". *bioRxiv* (2019), p. 560029.

Retrieve the homologs of a protein gene



Search for a significant match with:

- an (already annotated) protein sequence, e.g. with BLAST[2]

```
>1shg:A
AKELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
```

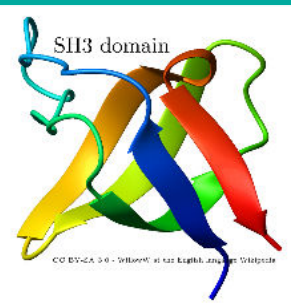
- the Profile HMM of a protein family, e.g. with HMMER[3] or HH-suite[4]

Score each position independently :-(

[2] S. F. Altschul et al. "Basic local alignment search tool". *Journal of molecular biology* (1990).

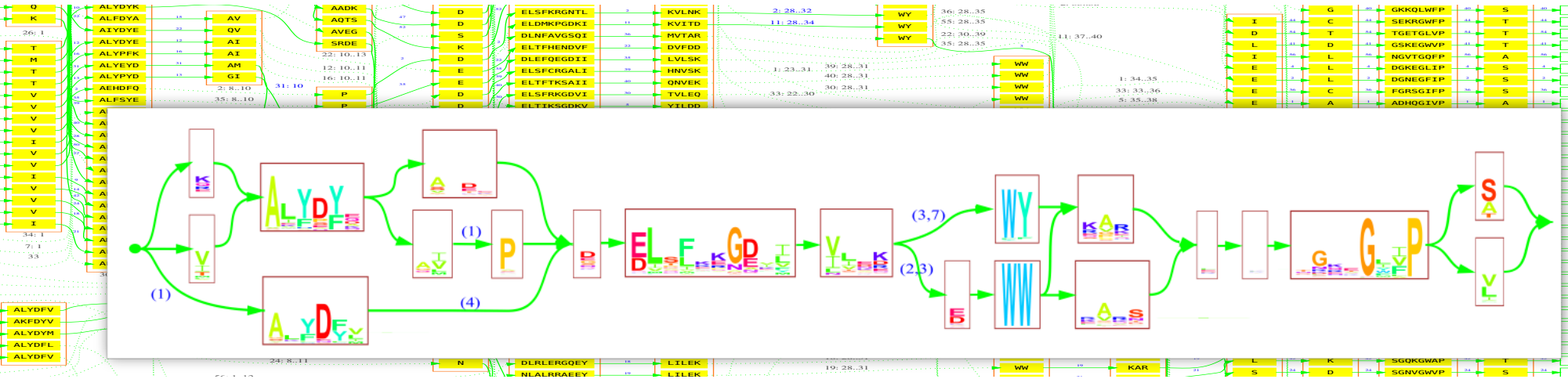
[3] S. R. Eddy. "Profile hidden Markov models.". *Bioinformatics* 14.9 (1998), pp. 755–763.

[4] M. Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation". *bioRxiv* (2019), p. 560029.



Automatic characterization of protein sequence families with:

- Automata (Protomata-Learner[5,6])



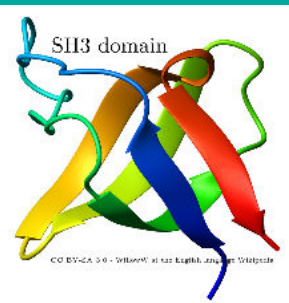
Local dependencies :-)

[5] G. Kerbellec. "Apprentissage d'automates modélisant des familles de séquences protéiques". PhD thesis. Université de Rennes 1, Apr. 2008, p. 139.

[6] A. Bretaudeau et al. "CyanoLyase: a database of phycobilin lyase sequences, motifs and functions". *Nucleic Acids Research* 41.Database-Issue (2013), pp. 396–401.

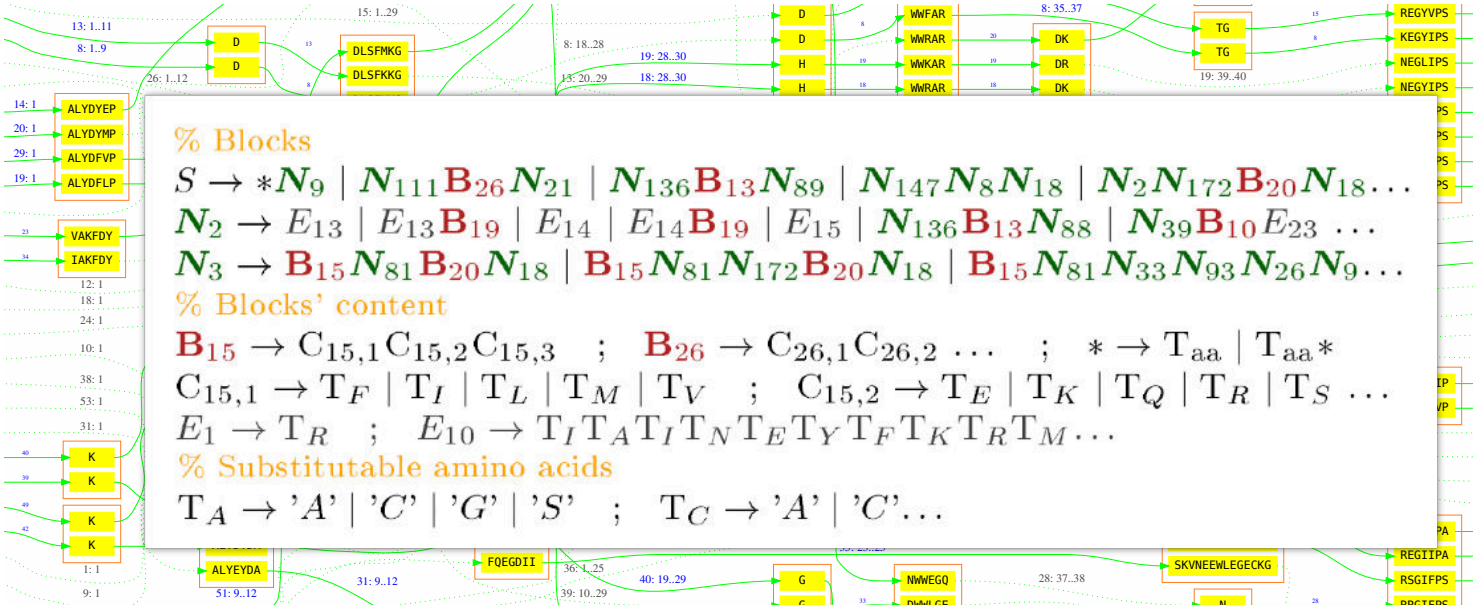
Our research

Grammatical inference on biological sequences



Automatic characterization of protein sequence families with:

- Context-free grammars (ReGLiS[7], see also[8])

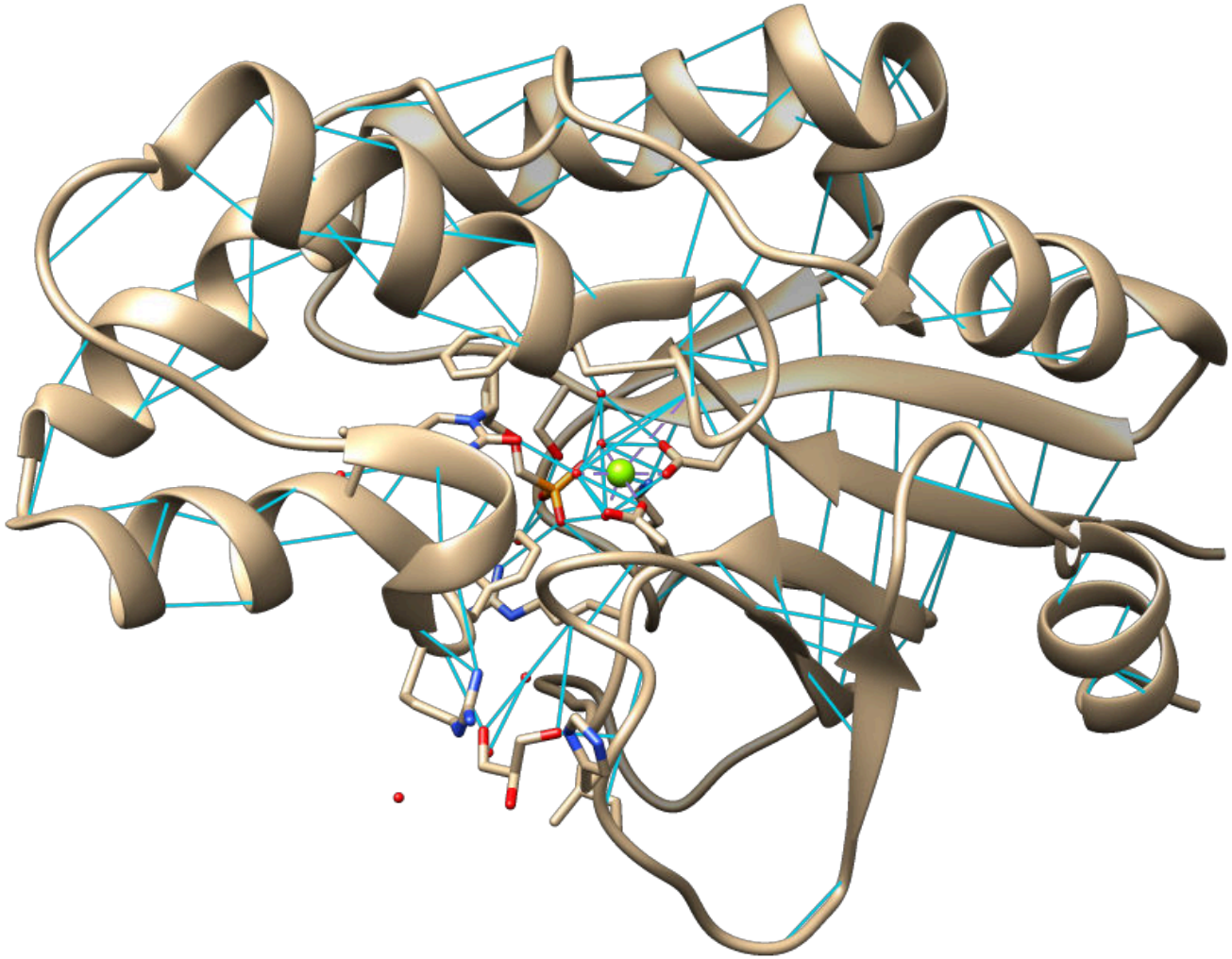


Nested dependencies :-D

[7] F. Coste, G. Garet, and J. Nicolas. "A bottom-up efficient algorithm learning substitutable languages from positive examples". *ICGI*. 2014.

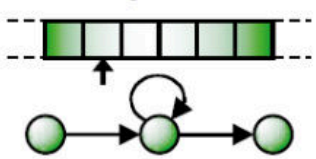
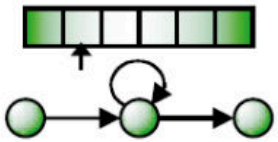
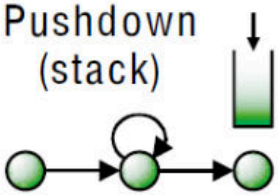
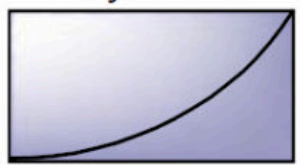
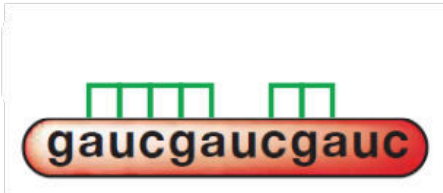
[8] W. Dyrka et al. "Estimating probabilistic context-free grammars for proteins using contact map constraints". *PeerJ* (2019).

Proteins are 3D objects



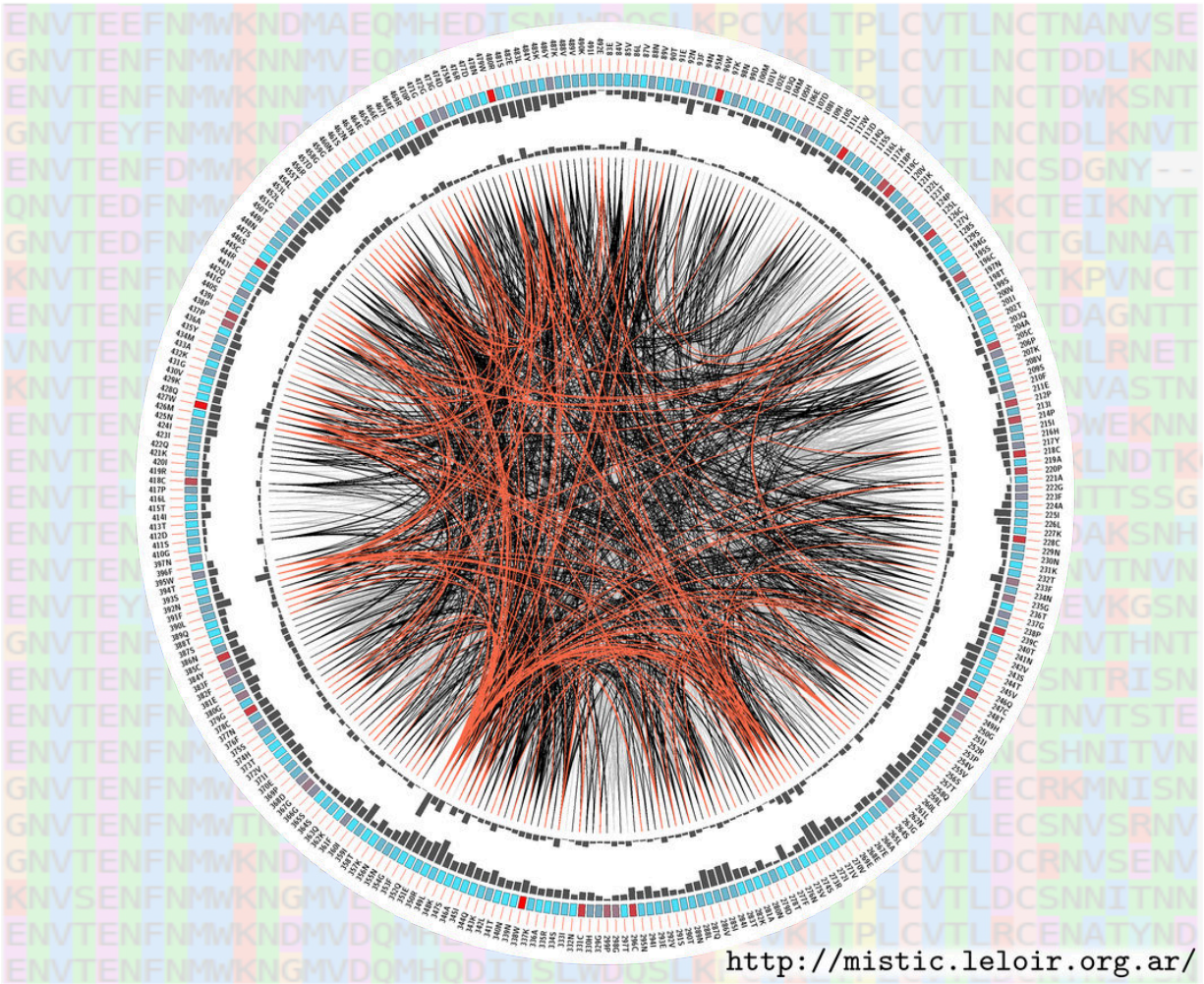
Many crossing interactions between amino-acids distant in the sequence but close in the structure

The Chomsky Hierarchy

Language	Automaton	Grammar	Recognition	Dependencies
Recursively enumerable languages	Turing machine 	Unrestricted $Baa \rightarrow A$	Undecidable 	
∪ Context-sensitive languages	Linear bounded 	Context sensitive $At \rightarrow aA$	Exponential? 	
∪ Context-free languages	Pushdown (stack) 	Context free $S \rightarrow gSc$	Polynomial 	
∪ Regular languages	Finite-state automaton 	Regular $A \rightarrow cA$	Linear 	

Adapted from D. Searls

Mutual information in HIV-1 gp120 homologs



Many (crossing) correlations between MSA columns

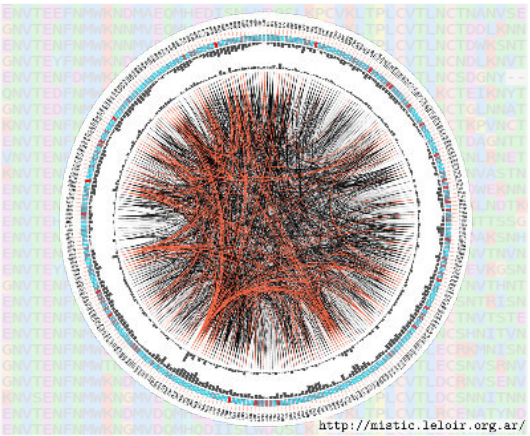
Direct Coupling Analysis to the rescue

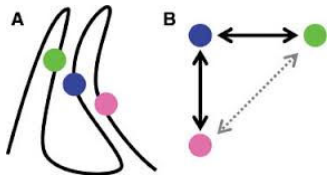
Direct Coupling Analysis (DCA) to the rescue

Hugo Talibart's PhD

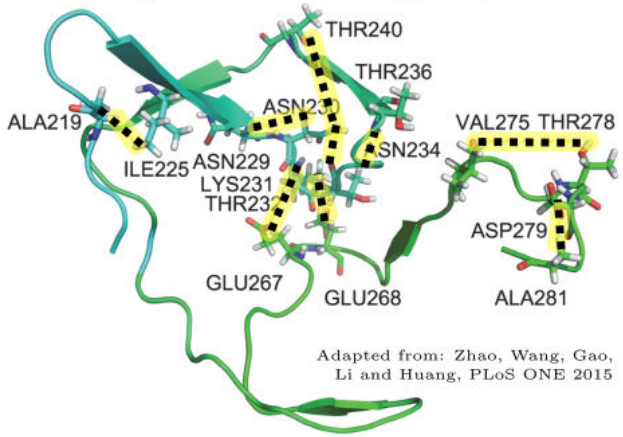
- A recent breakthrough for the prediction of 3D structures by prediction of contacts
- Principle : disentangle direct from indirect effects

Mutual information on HIV-1 gp120 protein





Coevolving residues in HIV-1 gp120 protein



- Idea: Use DCA for automatic characterization of protein families:
  - Identify important (crossing) dependencies with DCA
  - Build accordingly a syntactic model that can be used in practice...

Choice of DCA method

CCMpred[9]

- Best one-model precision for contact prediction[10]
- "Structuring" couplings

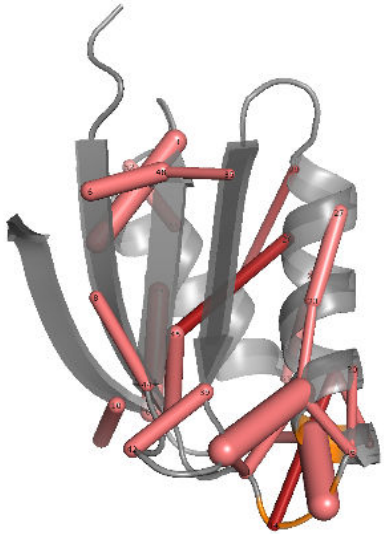


Figure: Top 25 PSICOV predictions

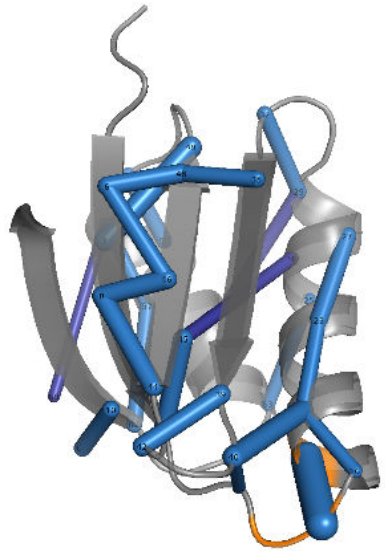


Figure: Top 25 CCMpred predictions

[9] S. Seemayer, M. Gruber, and J. Söding. "CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations". *Bioinformatics* 30.21 (2014), pp. 3128–3130.

[10] S. H. P. de Oliveira, J. Shi, and C. M. Deane. "Comparing co-evolution methods and their application to template-free protein structure prediction". *Bioinformatics* 33.3 (2017), pp. 373–381.

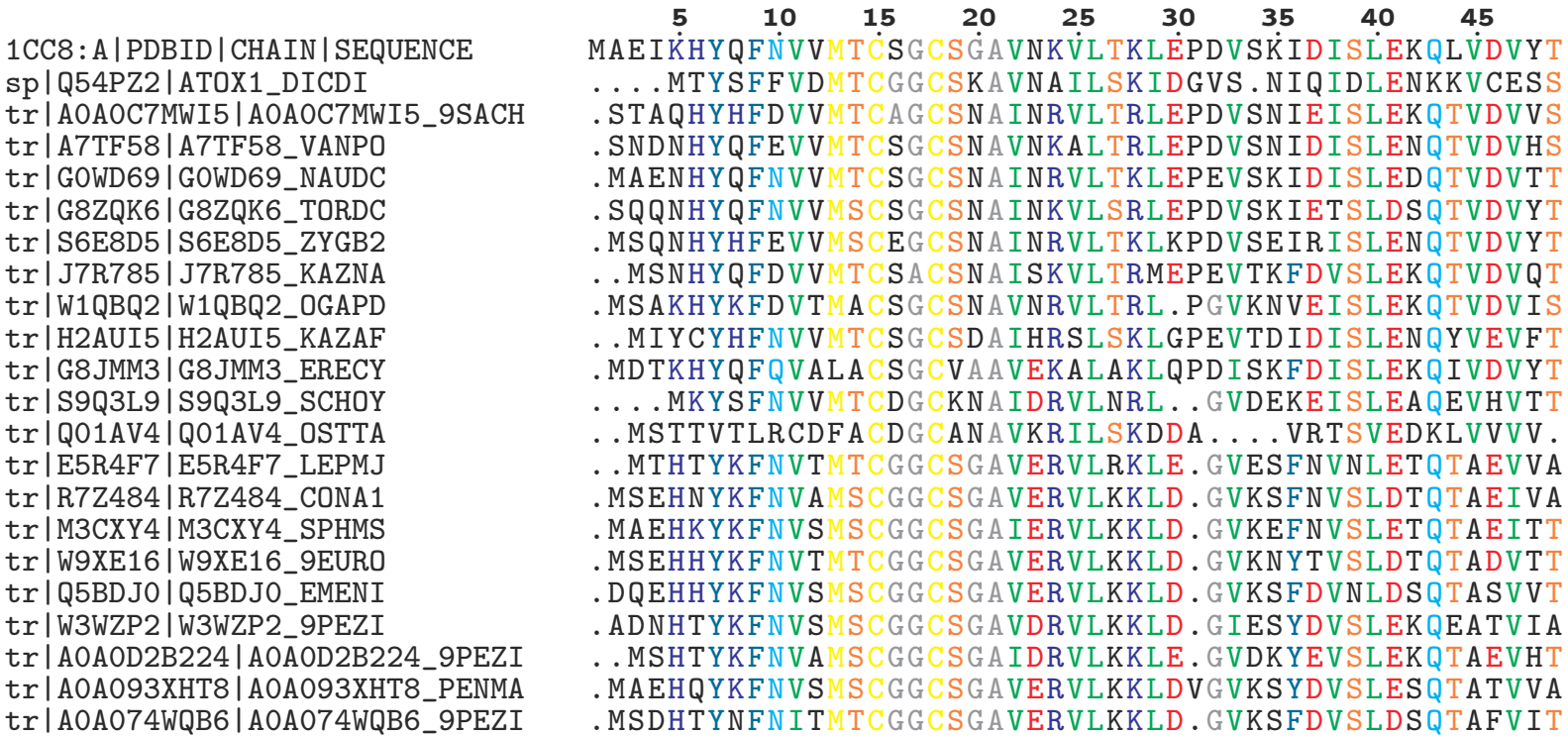
DCA workflow

1. Protein sequence query q

```
1CC8:A|PDBID|CHAIN|SEQUENCE    MAEIKHYQFNVVMTCSGCSGAVNKVLTKLEPDVSKIDISLEKQLVDVYT
```

DCA workflow

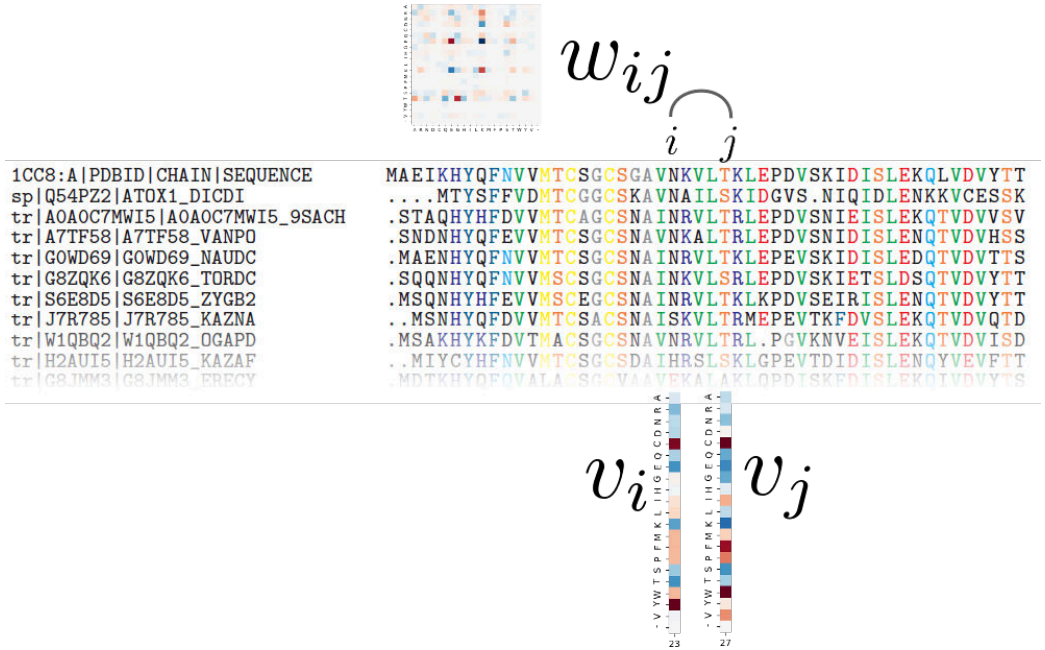
2. Retrieve close homologs and build a MSA (e.g. with HHblits[11])



[11] M. Remmert et al. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment". *Nature methods* 9.2 (2012), p. 173.

DCA workflow

3. Infer a Potts model from MSA



Probability of sequence

$a = a_1, \ldots, a_L$

Couplings

Fields

$$P(a|w,v) = \frac{1}{Z} \exp\left(\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} w_{ij}(a_i,a_j) + \sum_{i=1}^{L} v_i(a_i)\right)$$

Normalization constant

Inference of Potts model from MSA (CCMpred)

Maximise *pseudo*-likelihood of N aligned sequences, i.e.:

$$(w, v) = \underset{w,v}{\operatorname{argmax}} \sum_{n=1}^{N} \sum_{i=1}^{L} \log P(A_i = a_i^n | a_1^n, \cdots, a_{i-1}^n, a_{i+1}^n, \cdots, a_L^n, v, w)$$

(more tractable and still good precision)

while respecting empirical frequencies:

- $P_i(a) = f_i(a)$
- $P_{ij}(a, b) = f_{ij}(a, b)$

Using Potts model for contact prediction

4. Contacts in q are predicted using Frobenius norm of the couplings

$$||w_{ij}|| = \sqrt{\sum_a \sum_b w_{ij}(a,b)^2}$$

A larger norm is interpreted as a likelier contact between positions

Using Potts model for homology search

Using Potts model for homology search

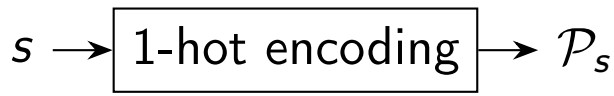
Use whole Potts model $\mathcal{P}_q$ of q instead of Frobenius norms

- Use $\mathcal{P}_q$ to score each possibly homologous sequence s
- Require to compute best alignment of s in $\mathcal{P}_q$
- As HHalign for pairs of HMMs[12], align directly pairs of Potts models

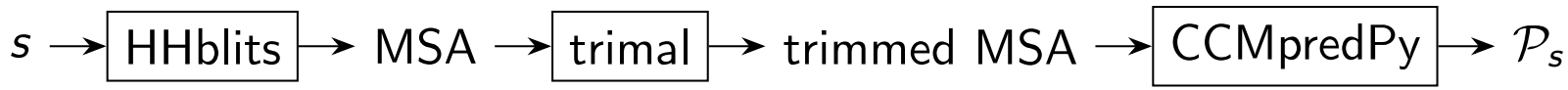
  $\rightsquigarrow$ A new tool: ComPotts

with two options to get Potts model $\mathcal{P}_s$ of s:

- One-hot encoding $\quad v_i(a_i) = 1,\ w_{ij}(a_i, a_j) = 1$, others are 0

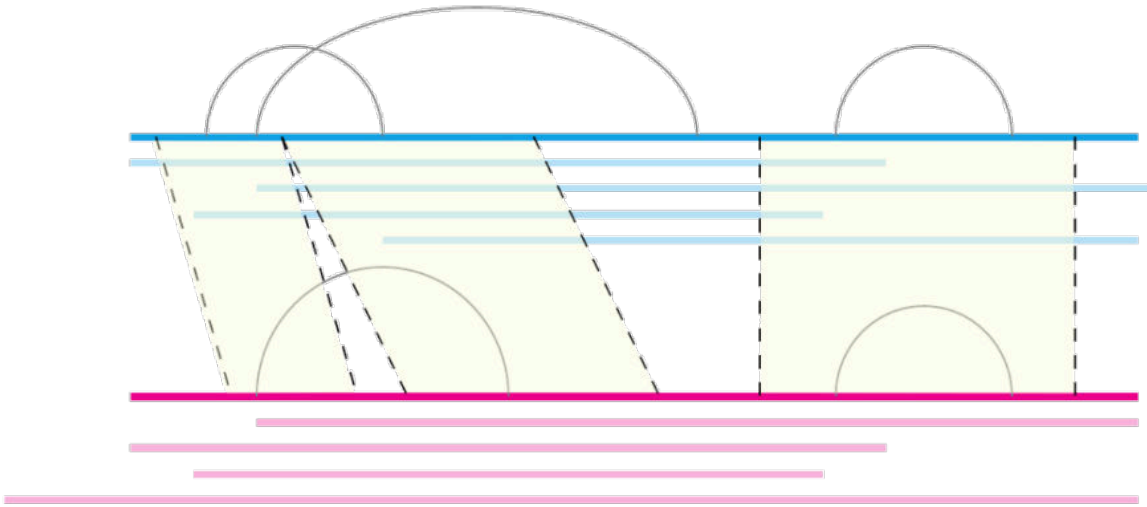


- From close homologs of s $\quad$ as for $\mathcal{P}_q$



[12] J. Söding. "Protein homology detection by HMM–HMM comparison". *Bioinformatics* 21.7 (2004), pp. 951–960.

ComPotts (**Com**paring **Potts** models)



$$s(A,B) = \sum_{i=1}^{L_A}\sum_{k=1}^{L_B} s_v(v_i^A, v_k^B)x_{ik} + \sum_{i=1}^{L_A-1}\sum_{j=i+1}^{L_A}\sum_{k=1}^{L_B-1}\sum_{l=k+1}^{L_B} s_w(w_{ij}^A, w_{kl}^B)y_{ikjl}$$

$$x_{ik} \geq \sum_{j,l\in C} y_{ikjl} \qquad \forall C \in \mathcal{C}^+_{i,k}, i \in [1, n_A - 1], k \in [1, n_B - 1]$$

$$x_{ik} \geq \sum_{j,l\in C} y_{jlik} \qquad \forall C \in \mathcal{C}^-_{i,k}, i \in [2, n_A], k \in [2, n_B]$$

$$x_{ik} \leq 1 + \sum_{\substack{j,l\in C \\ s(A_{ij},B_{kl})\leq 0}} (y_{ikjl} - x_{jl}) \qquad \forall C \in \mathcal{C}^+_{i,k}, i \in [1, n_A - 1], k \in [1, n_B - 1]$$

$$\sum_{i,k\in C} x_{ik} \leq 1 \qquad \forall C \in \mathcal{C}$$

$$\mathbf{y} \geq 0$$

$$\mathbf{x} \text{ binary.}$$

- Formulation of Potts model alignment as an Integer Linear Programming (ILP) problem
- Based on Inken Wohlers' solver[13]

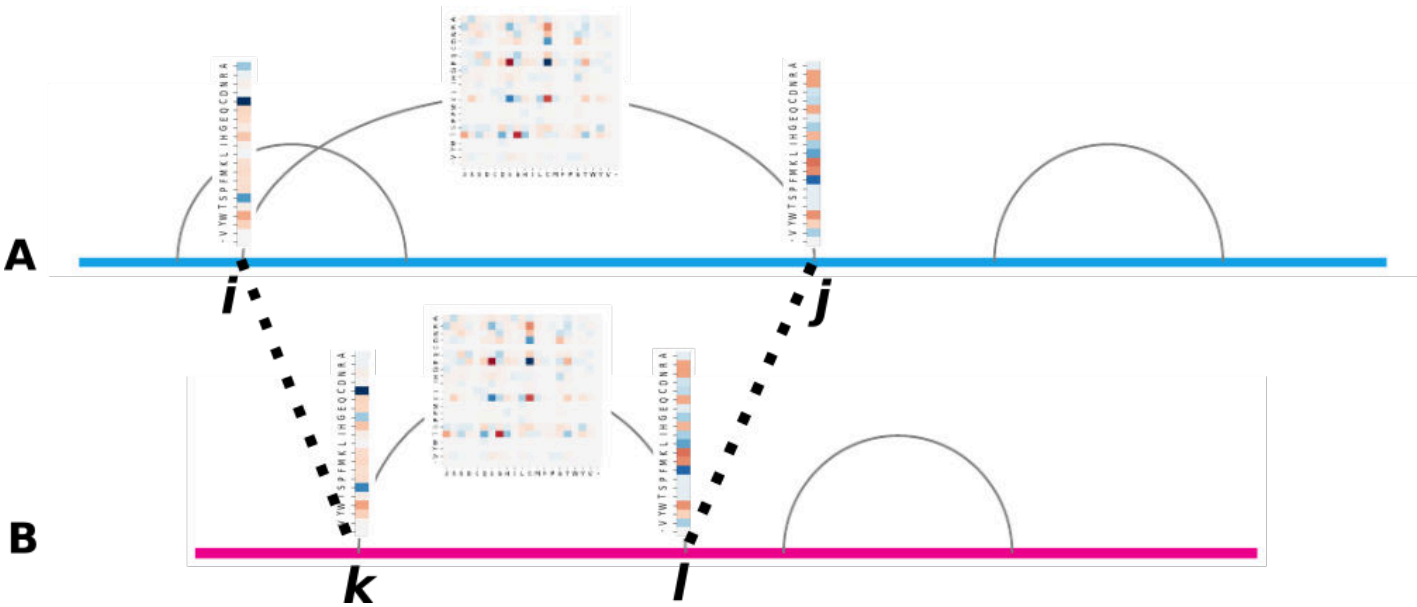
[13] I. Wohlers. "Exact Algorithms For Pairwise Protein Structure Alignment". PhD thesis. Vrije Universiteit, Jan. 2012, pp. 1 –147.

Scoring alignment of Potts models A and B

$$s(A,B) = \sum_{i=1}^{L_A} \sum_{k=1}^{L_B} s_v(v_i^A, v_k^B) x_{ik} + \sum_{i=1}^{L_A-1} \sum_{j=i+1}^{L_A} \sum_{k=1}^{L_B-1} \sum_{l=k+1}^{L_B} s_w(w_{ij}^A, w_{kl}^B) y_{ikjl}$$

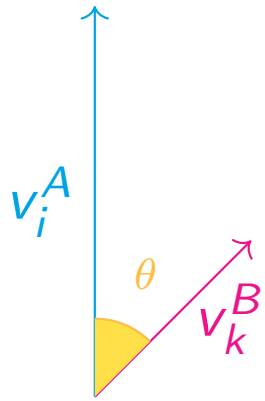
where

- $x_{ik} = 1$ iff position i of A and position k of B are aligned (otherwise, $x_{ik} = 0$)
- $y_{ikjl} = 1$ iff $x_{ik} = 1$ and $x_{jl} = 1$ (otherwise, $y_{ikjl} = 0$)



Choice of $s_v(v_i, v_k)$ and $s_w(w_{ij}, w_{kl})$: scalar products

- $s_v(v_i^A, v_k^B) = \langle v_i^A, v_k^B \rangle$
  - $\rightarrow$ standard scalar product : $\langle x, y \rangle = \sum_i x_i y_i$
- $s_w(w_{ij}^A, w_{kl}^B) = \langle w_{ij}^A, w_{kl}^B \rangle_F$
  - $\rightarrow$ Frobenius scalar product : $\langle X, Y \rangle_F = \sum_i \sum_j X_{ij} Y_{ij}$



Geometric insight

$$\langle v_i^A, v_k^B \rangle = \left\| v_i^A \right\| \left\| v_k^B \right\| \cos\theta$$

importance of position i

importance of position k

similarity measure

Natural extension of the 1D score of a sequence

$$P(a|w,v) = \frac{1}{Z}\exp\left(\mathcal{H}(a|v,w)\right)$$

$$\mathcal{H}(a|v,w) = \sum_{i=1}^{L} v_i(a_i) + \sum_{i=1}^{L-1}\sum_{j=i+1}^{L} w_{ij}(a_i,a_j)$$

$$= \sum_{i=1}^{L} \langle v_i, e_{a_i}\rangle + \sum_{i=1}^{L-1}\sum_{j=i+1}^{L} \langle w_{ij}, e_{a_i a_j}\rangle_F$$

a_j

$$e_{a_i a_j} = \begin{pmatrix} 0 & \dots & \dots & 0 & \dots & \dots & 0 \\ \vdots & & & \vdots & & & \vdots \\ & & & 0 & & & \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ & & & 0 & & & \\ \vdots & & & \vdots & & & \vdots \\ 0 & \dots & \dots & 0 & \dots & \dots & 0 \end{pmatrix} \leftarrow a_i$$

$$e_{a_i} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow a_i$$

First experiments

- PDB 1CC8 : Atx1 metallochaperone (*Saccharomyces cerevisiae*)
  $\times$ one homolog s (150 sequences sampling, identity with 1CC8 : 25%-50%)

One-hot encoding of $\mathcal{P}_s$ Timeout: 6 hours

	trimmed	not trimmed
$\epsilon = machine\ epsilon$	$t \in [11s, 6h]$, avg: 2h	$t \in [25s, 6h]$, avg: $4h30$
$\epsilon = 1$[14]	$t \in [8s, 6h]$, avg: $1h30$	$t \in [16s, 6h]$, avg: 2h

Build $\mathcal{P}_s$ from homologs of s Timeout: 6 hours

	trimmed	not trimmed
$\epsilon = machine\ epsilon$	$t \in [3s, 6h]$, avg: 3 min	$t \in [3s, 6h]$, avg: 8 min
$\epsilon = 1$	$t \in [2s, 6s]$, avg: 5s	$t \in [3s, 50s]$, avg: 21s

Tractable time! not for the simpler models?
Small proteins, easy to align. . .

[14] $\simeq$ Energy needed to change one a.a. into another

Testing the limits on thioredoxins

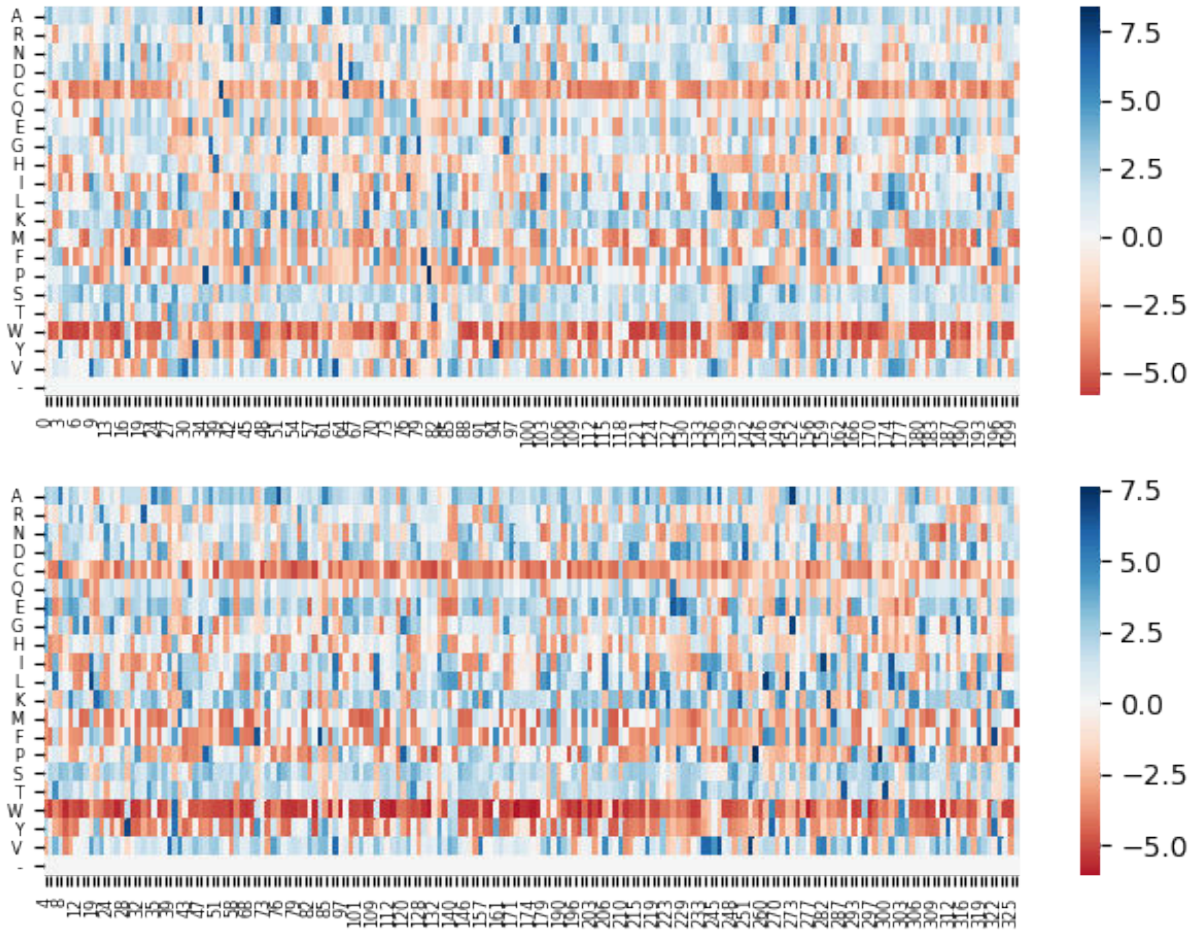
- Enzymes involved in reduction–oxidation reactions through oxidation of their active site

Figure: 3D structure of thioredoxin-1 (*Caenorhabditis elegans*) (Q09433)

- 100 amino acids on average
- Between 15 and 20% sequence identity within the family
  $\rightarrow$ known to be hard to align

An example of failure

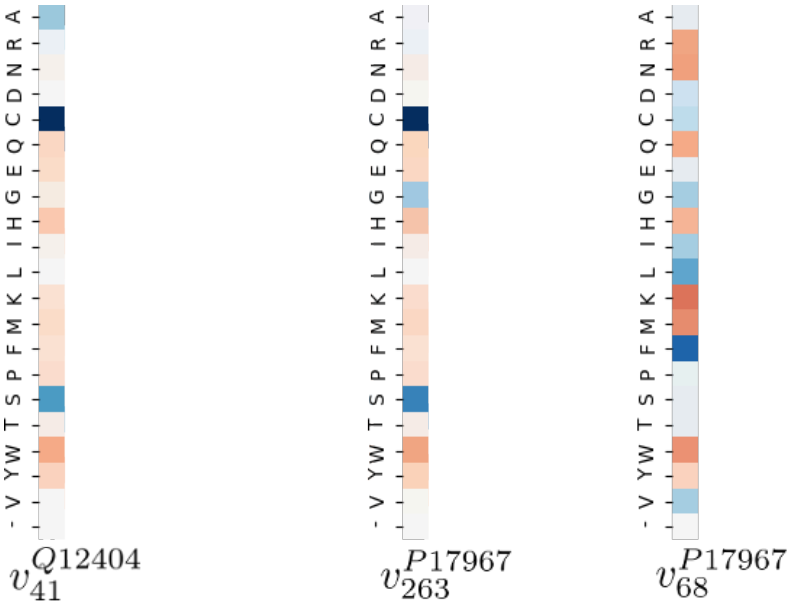
$\{v_i^{Q12404}\}_i$ and $\{v_i^{P17967}\}_i$ aligned by ComPotts:



Even well-conserved positions of the active site are not aligned

The trouble with scalar product alone

- A well-conserved column i may have a smaller $||v_i||$ than a less conserved column j



v_{41}^{Q12404} v_{263}^{P17967} v_{68}^{P17967}

$$||v_{262}^{P17967}|| \simeq 10.5 < ||v_{68}^{P17967}|| \simeq 11.4$$

- It may be more profitable to align many less conserved columns than to align fewer well-conserved columns with each other

An idea

- Use rescaling function : $f(x) = sign(x)(e^{|x|} - 1)$

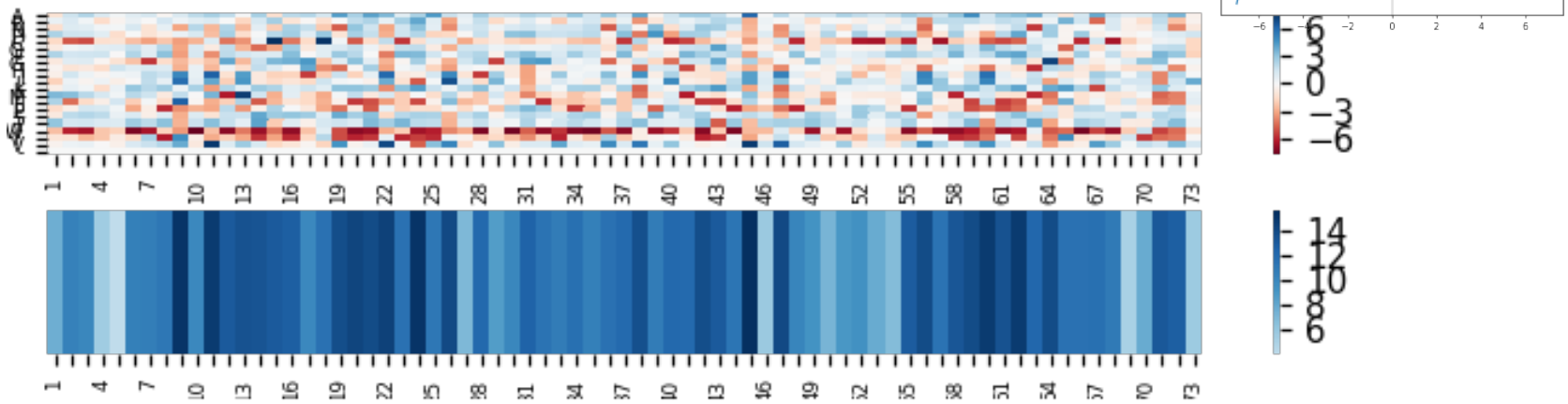


Figure: Before rescaling

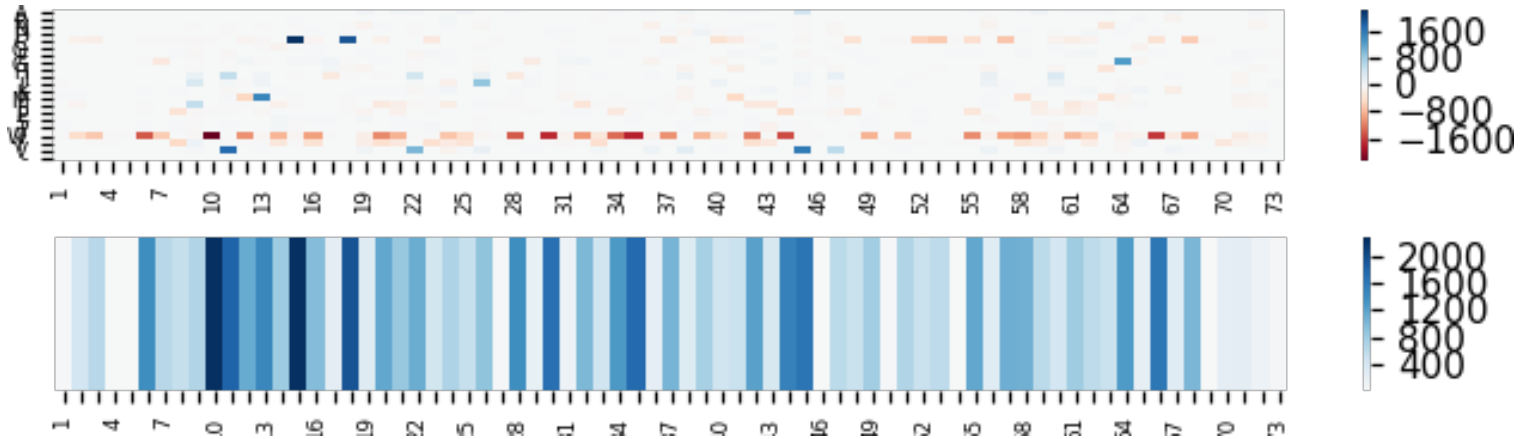
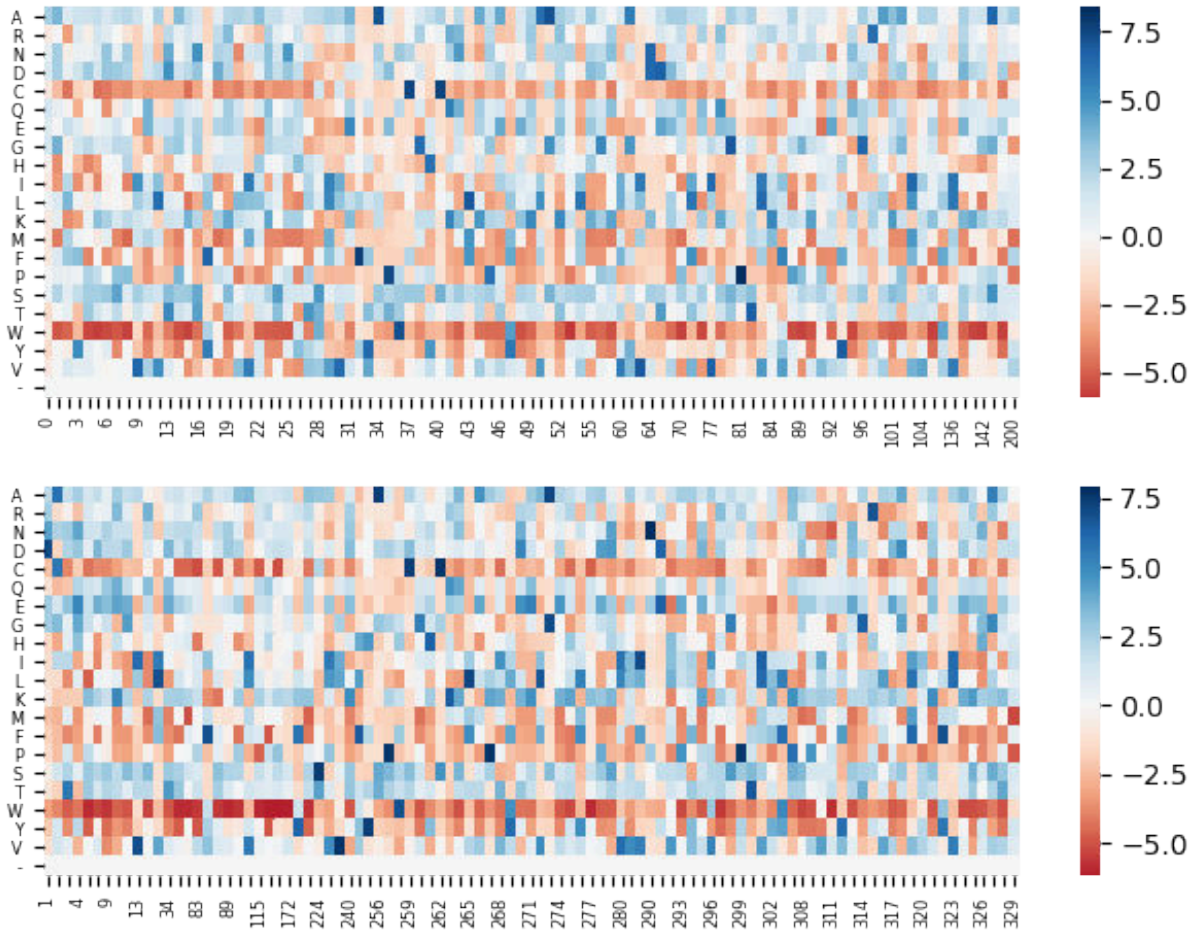


Figure: After rescaling

It's better :-)

$\{v_i^{Q12404}\}_i$ and $\{v_i^{P17967}\}_i$ aligned by ComPotts (wo couplings!)



To be continued...

- How to rescale also consistently the couplings w_{ij}?
  - $\rightarrow$ Slightly change the rescaling function $f(x) = sign(x)(\beta e^{\alpha|x|} - \gamma)$?
- Other similarity functions?...
- Introduce gap costs
- Constrain Potts model inference?
  - Canonical Potts model?
  - Better control amplitude of vectors and matrices?

Conclusion so far...

- Good news: alignment to Potts model is tractable
- A surprise: may require a transformation to a Potts model
- A working efficient implementation
- Quality of alignment can still be improved...

Thanks for your attention!

Ideas, remarks, suggestions are welcome.
See you next to our poster. . .

Bibliography I

[Alt+90] S. F. Altschul et al. "Basic local alignment search tool". *Journal of molecular biology* (1990).

[Edd98] S. R. Eddy. "Profile hidden Markov models.". *Bioinformatics* 14.9 (1998), pp. 755–763.

[Ste+19] M. Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation". *bioRxiv* (2019), p. 560029.

[Ker08] G. Kerbellec. "Apprentissage d'automates modélisant des familles de séquences protéiques". PhD thesis. Université de Rennes 1, Apr. 2008, p. 139.

[Bre+13] A. Bretaudeau et al. "CyanoLyase: a database of phycobilin lyase sequences, motifs and functions". *Nucleic Acids Research* 41.Database-Issue (2013), pp. 396–401.

Bibliography II

[CGN14] F. Coste, G. Garet, and J. Nicolas. "A bottom-up efficient algorithm learning substitutable languages from positive examples". *ICGI*. 2014.

[Dyr+19] W. Dyrka et al. "Estimating probabilistic context-free grammars for proteins using contact map constraints". *PeerJ* (2019).

[SGS14] S. Seemayer, M. Gruber, and J. Söding. "CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations". *Bioinformatics* 30.21 (2014), pp. 3128–3130.

[OSD17] S. H. P. de Oliveira, J. Shi, and C. M. Deane. "Comparing co-evolution methods and their application to template-free protein structure prediction". *Bioinformatics* 33.3 (2017), pp. 373–381.

Bibliography III

[Jon+11] D. T. Jones et al. "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments". *Bioinformatics* 28.2 (2011), pp. 184–190.

[Rem+12] M. Remmert et al. "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment". *Nature methods* 9.2 (2012), p. 173.

[Söd04] J. Söding. "Protein homology detection by HMM–HMM comparison". *Bioinformatics* 21.7 (2004), pp. 951–960.

[Woh12] I. Wohlers. "Exact Algorithms For Pairwise Protein Structure Alignment". PhD thesis. Vrije Universiteit, Jan. 2012, pp. 1 –147.