ComPotts

Optimal alignment of coevolutionary models for protein sequences

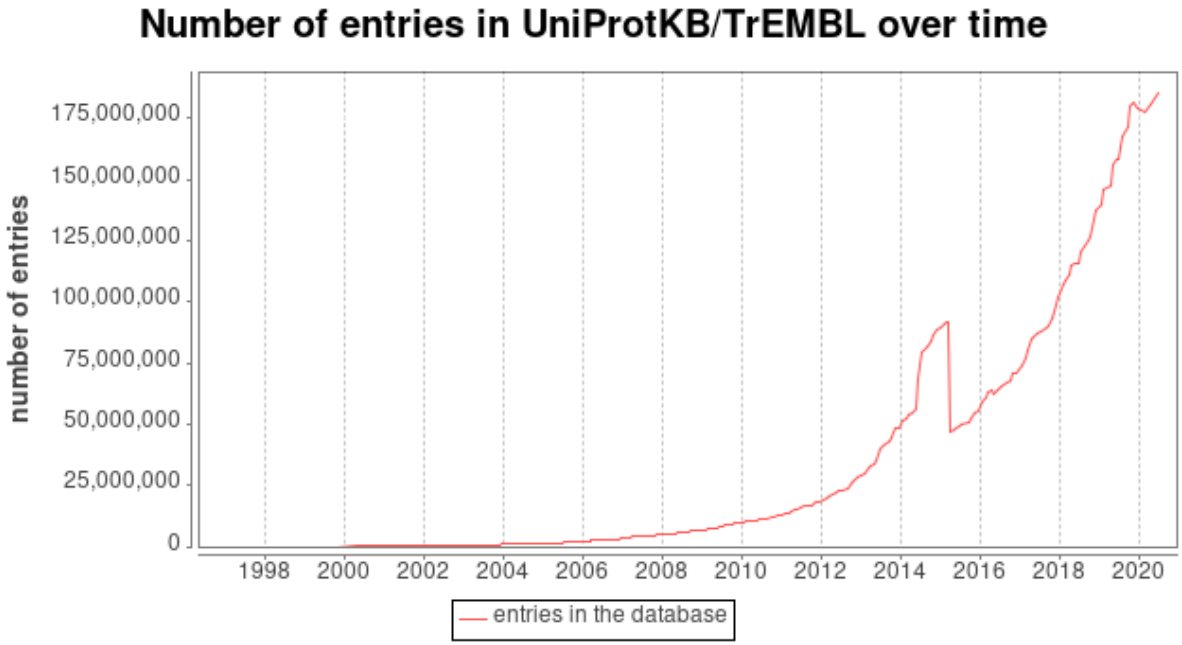
Hugo Talibart, François Coste

JOBIM 2020

Motivation: protein sequence annotation problem

Sequencing technologies $\rightarrow$ **increasing number** of **protein sequences**

Illumina NextSeq 2000 sequencer



Problem

Annotate all these sequences?

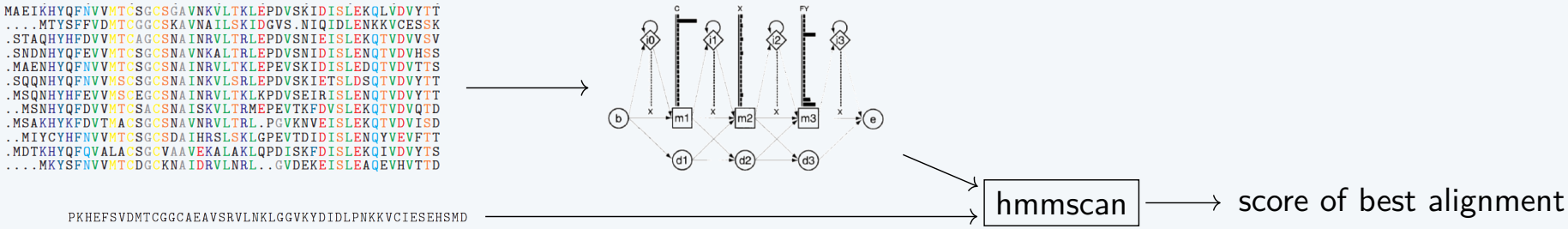
$\rightarrow$ *in silico* annotation

Annotate sequences with alignment-based homology search

Align a **sequence** to a **sequence**

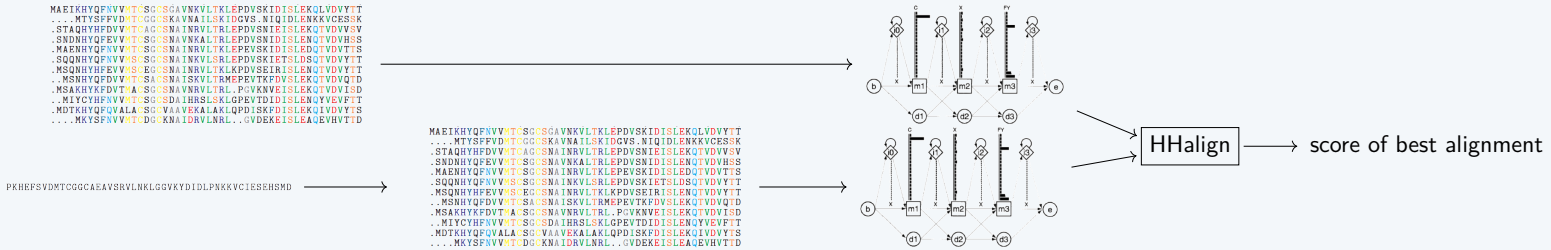
Altschul et al., "Basic local alignment search tool", 1990

Align a **sequence** to a **model** (pHMMs: profile Hidden Markov Models)



Eddy, "Profile hidden Markov models.", 1998

Align a **model** to a **model**

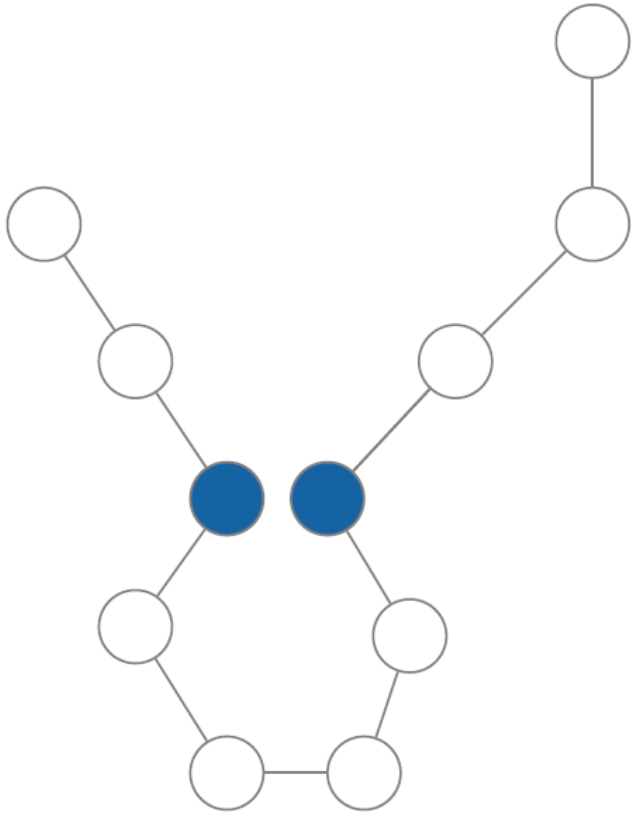


Steinegger et al., "HH-suite3 for fast remote homology detection and deep protein annotation", 2019

What about long-distance dependencies?

C W R V H Y M D P G N
C W R L H Y M D P G N
C W E V H F M K P G N
C W E L R Y M K P G N
C W E V H Y M K P G N
C W H V H Y M E P S N
C W H V H Y M E P G N

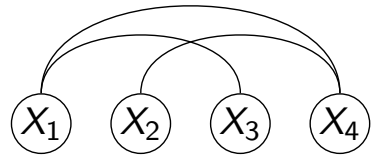
correlated positions



Goal: **Homology search** with **distant dependencies**

Homology search with distant dependencies: first attempts

Model proteins with **Markov Random Fields** (MRFs)



- SMURF[1]
  - pHMMs + dependencies between β-strands ($\Rightarrow$ limited to all-β folds)
  - aligns sequence to model
  - outperforms HMMER in propeller fold prediction
- MRFalign[2]
  - MRFs allow dependencies between all positions
  - aligns model to model
  - complex workflow for building MRFs and aligning them
  - outperforms HMMER and HHsearch in remote homology detection
    on SCOP20, SCOP40 and SCOP80 benchmarks at the superfamily level

$\rightarrow$ **shows the potential of using long-distance dependencies**

[1] Menke, Berger, and Cowen, "Markov random fields reveal an N-terminal double beta-propeller motif as part of a bacterial hybrid two-component sensor system", 2010.

[2] Ma et al., "MRFalign: protein homology detection through alignment of Markov random fields", 2014.

Our proposition: exploit the Potts model

- **Potts model**: another type of Markov Random Field
- Based on maximum-entropy principle
- First applied to proteins within **Direct Coupling Analysis**[3]
- **Successfully applied to contact prediction**
  - → dramatic improvement in CASP predictions[4]

Can Potts model **improve homology search**? [5][6]

[3] Weigt et al., "Identification of direct residue contacts in protein–protein interaction by message passing", 2009.

[4] Monastyrskyy et al., "New encouraging developments in contact prediction: Assessment of the CASP 11 results", 2016.

[5] H. Talibart and F. Coste, "Using residues coevolution to search for protein homologs through alignment of Potts models". CECAM, 2019.

[6] A. P. Muntoni et al., "Using Direct Coupling Analysis for the protein sequences alignment problem". CECAM, 2019.

Represent sets of protein sequences with Potts models

- Markov Random Field **representing MSA of homologous proteins**

- Derives from the maximum-entropy principle
  - reproduces MSA empirical frequencies:

$$\sum_{x \in \Sigma^L :\, x_i = a} \mathbb{P}(x_1, \cdots, x_L) = f_i(a)$$

$$\sum_{x \in \Sigma^L :\, x_i = a, x_j = b} \mathbb{P}(x_1, \cdots, x_L) = f_{ij}(a, b)$$

Probability of sequence $x = x_1, \ldots, x_L$

Couplings

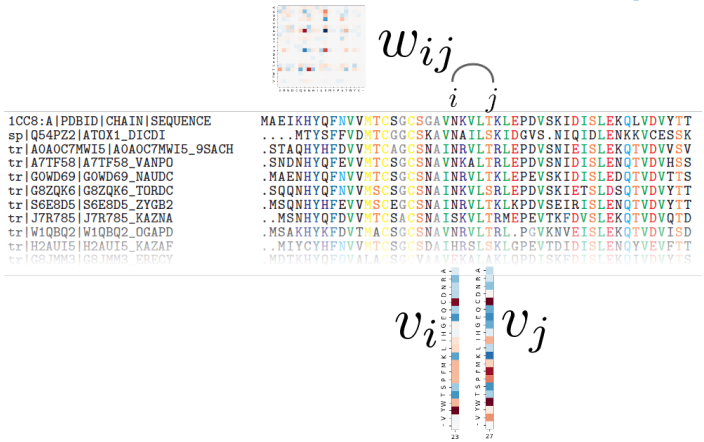
$$\mathbb{P}(x|w, v) = \frac{1}{Z} \exp\left(\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} w_{ij}(x_i, x_j) + \sum_{i=1}^{L} v_i(x_i)\right)$$

Normalization constant

Fields

Represent sets of protein sequences with Potts models

- Markov Random Field **representing MSA of homologous proteins**



- Field v_i $\sim$ positional conservation
- Coupling w_{ij} $\sim$ covariation

- Derives from the maximum-entropy principle
  - reproduces MSA empirical frequencies:

$$\sum_{x \in \Sigma^L :\, x_i = a} \mathbb{P}(x_1, \cdots, x_L) = f_i(a)$$

$$\sum_{x \in \Sigma^L :\, x_i = a, x_j = b} \mathbb{P}(x_1, \cdots, x_L) = f_{ij}(a, b)$$

Probability of sequence $x = x_1, \ldots, x_L$

Couplings

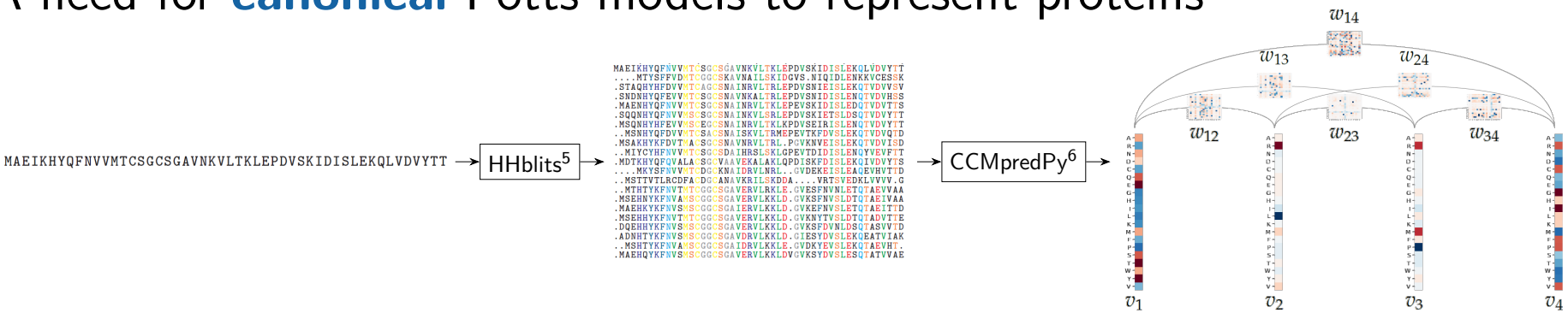
$$\mathbb{P}(x|w,v) = \frac{1}{Z} \exp\left(\sum_{i=1}^{L-1} \sum_{j=i+1}^{L} w_{ij}(x_i, x_j) + \sum_{i=1}^{L} v_i(x_i) \right)$$

Normalization constant

Fields

Use Potts models for homology search

A need for **canonical** Potts models to represent proteins



Choice of prior on positional parameters: center v at v^*: $\dfrac{\exp(v_i^*(a))}{\sum_{b=1}^{q}\exp(v_i^*(b))}$

$\rightarrow$ yields correct model if no columns are coupled, i.e. $\mathbb{P}(x|v,w) = \prod_{i=1}^{L}\mathbb{P}(x_i)$

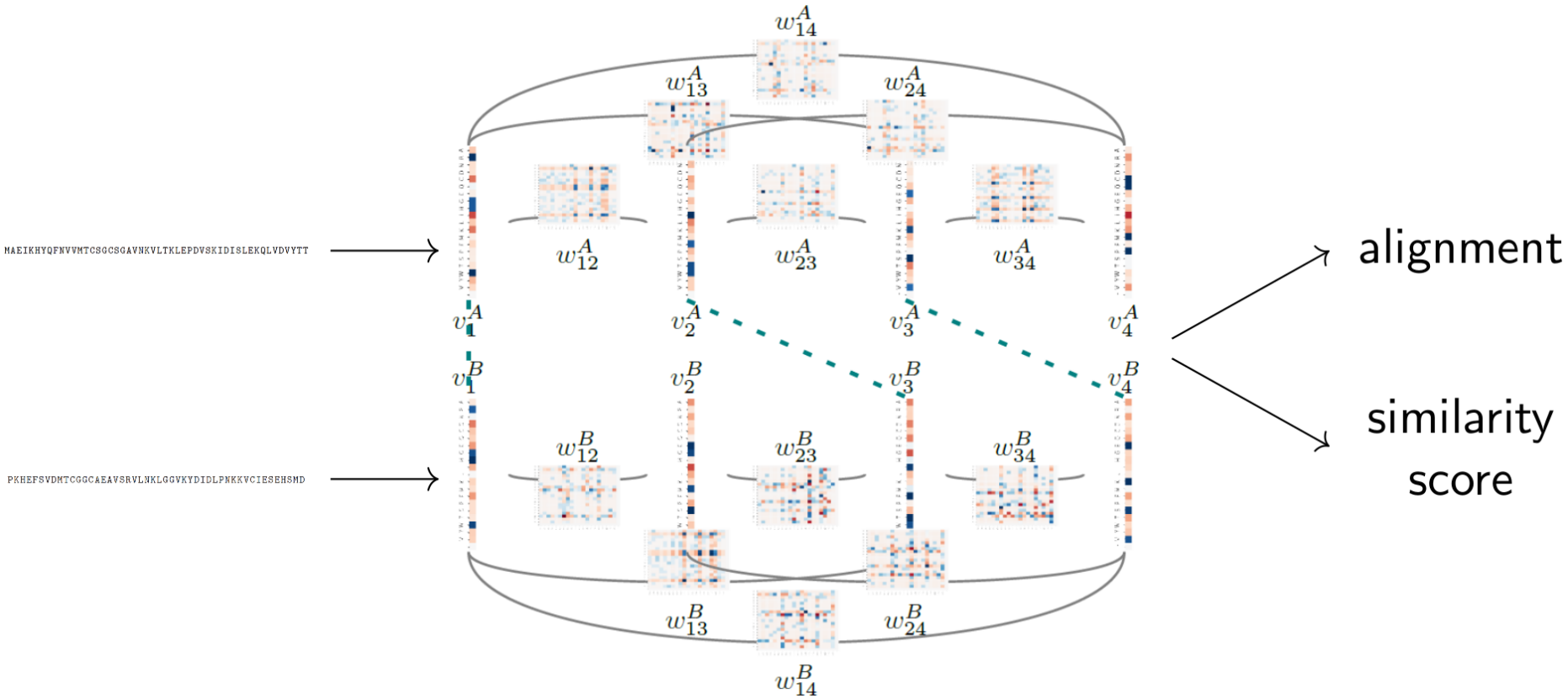
$\rightarrow$ Intuition: only necessary couplings should be added

[5] Steinegger et al., "HH-suite3 for fast remote homology detection and deep protein annotation", 2019.

[6] Vorberg, "Bayesian Statistical Approach for Protein Residue-Residue Contact Prediction", 2017.

Compare proteins by aligning Potts models

Compare two proteins by aligning Potts models: **ComPotts**

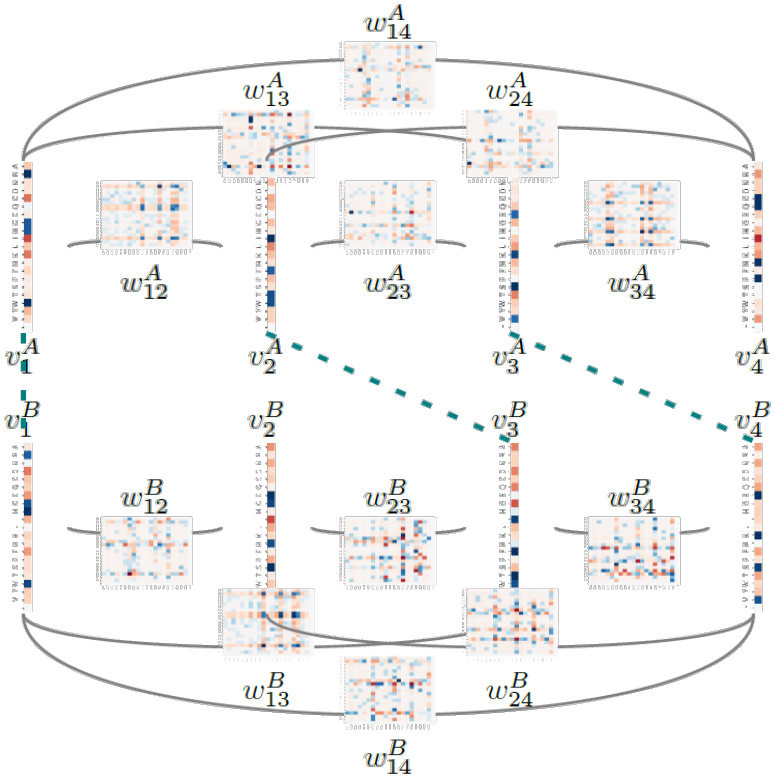


ComPotts: optimal Potts model alignment

- Formulation of Potts model alignment as an

Integer Linear Programming (ILP) problem

$$\max \sum_{i=1}^{L_A}\sum_{k=1}^{L_B}\langle v_i^A, v_k^B\rangle x_{ik} + \sum_{i=1}^{L_A-1}\sum_{j=i+1}^{L_A}\sum_{k=1}^{L_B-1}\sum_{l=k+1}^{L_B}\langle w_{ij}^A, w_{kl}^B\rangle y_{ikjl}$$



s.t.

$$x_{ik} \geq \sum_{j.l\in C} y_{ikjl} \quad \forall C \in \mathcal{C}_{i.k}^+, i \in [1, L_A - 1], k \in [1, L_B - 1]$$

$$x_{ik} \geq \sum_{j.l\in C} y_{jlik} \quad \forall C \in \mathcal{C}_{i.k}^-, i \in [2, L_A], k \in [2, L_B]$$

$$x_{ik} \leq 1 + \sum_{\substack{j.l\in C \\ \langle w_{ij}^A, w_{kl}^B\rangle \leq 0}} (y_{ikjl} - x_{jl}) \quad \forall C \in \mathcal{C}_{i.k}^+, i \in [1, L_A - 1], k \in [1, L_B - 1]$$

$$\sum_{i.k\in C} x_{ik} \leq 1 \quad \forall C \in \mathcal{C}$$

$$y \geq 0$$

$$x \text{ binary}$$

- Based on Wohlers, Andonov, Malod-Dognin and Klau's solver[7]

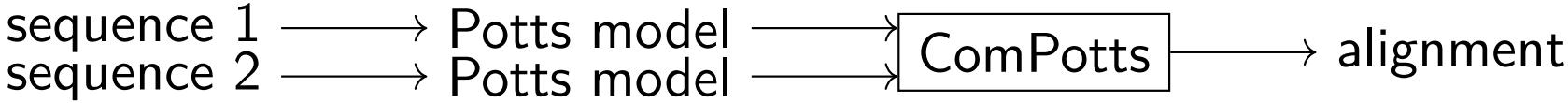
[7] Wohlers, Andonov, and Klau, "DALIX: optimal DALI protein structure alignment", 2012.

Preliminary experiments to assess **alignment quality**

(homology search = **align** + score)

Preliminary experiments to assess **alignment quality**

- **59 sequence pairs**
  - extracted from reference structural multiple sequence alignments from BAliBASE[8] and sisyphus[9]
  - low sequence identity (6 - 12%)
  - length(training MSAs) < 200
- **Build Potts models** sequence → HHblits → MSA → CCMpredPy → Potts model
- **Align with ComPotts** (stop when $\frac{2(UB-LB)}{s(A,A)+s(B,B)} < 0.005$)



- **Compare our alignment with reference alignment**

[8] Thompson, Plewniak, and Poch, "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs.", 1999.

[9] Andreeva et al., "SISYPHUS—structural alignments for proteins with non-trivial relationships", 2007.

Compare with main (alignment-based) homology search rivals

- BLASTp v2.9.0+[10] (without E-value cutoff)
  - aligns sequences
- HHalign v3.03[11]
  - aligns HMMs inferred from MSAs outputted by HHblits
- MRFalign v0.90[12]
  - aligns MRFs built from sequences

\+ Matt v1.00[13]
  - aligns corresponding PDB structures

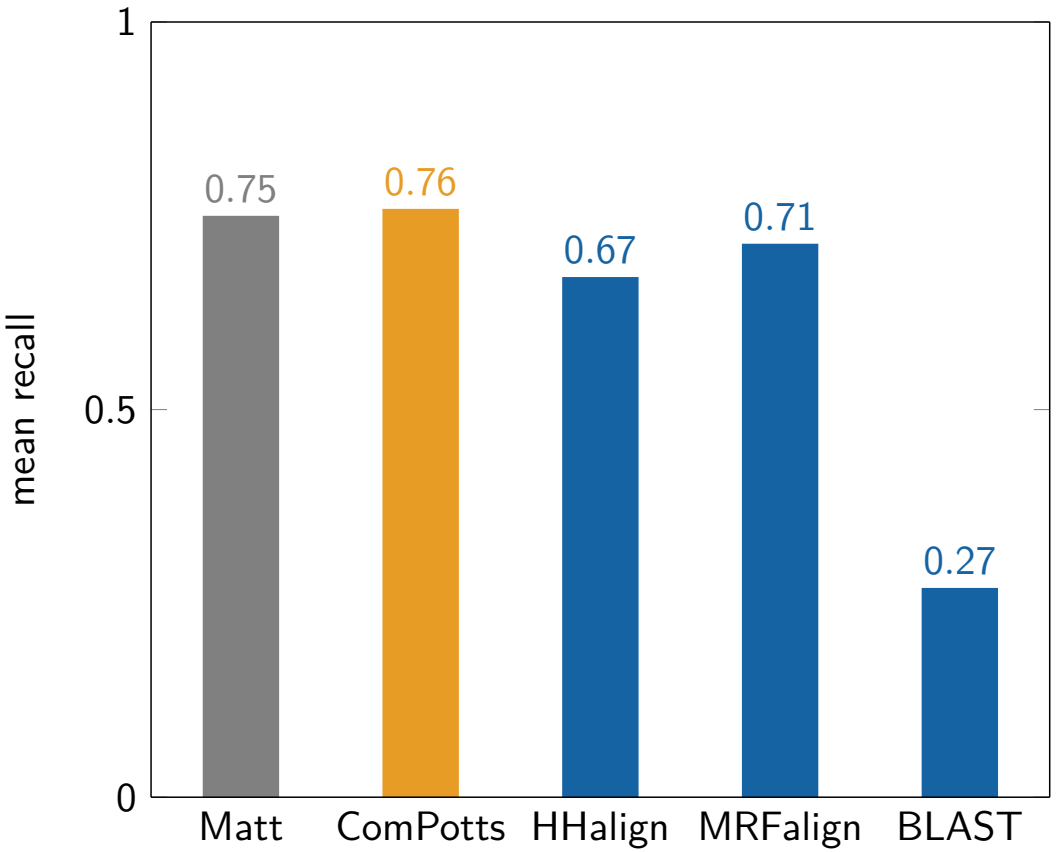
[10] Altschul et al., "Basic local alignment search tool", 1990.

[11] Steinegger et al., "HH-suite3 for fast remote homology detection and deep protein annotation", 2019.

[12] Ma et al., "MRFalign: protein homology detection through alignment of Markov random fields", 2014.

[13] Menke, Berger, and Cowen, "Matt: local flexibility aids protein multiple structure alignment", 2008.

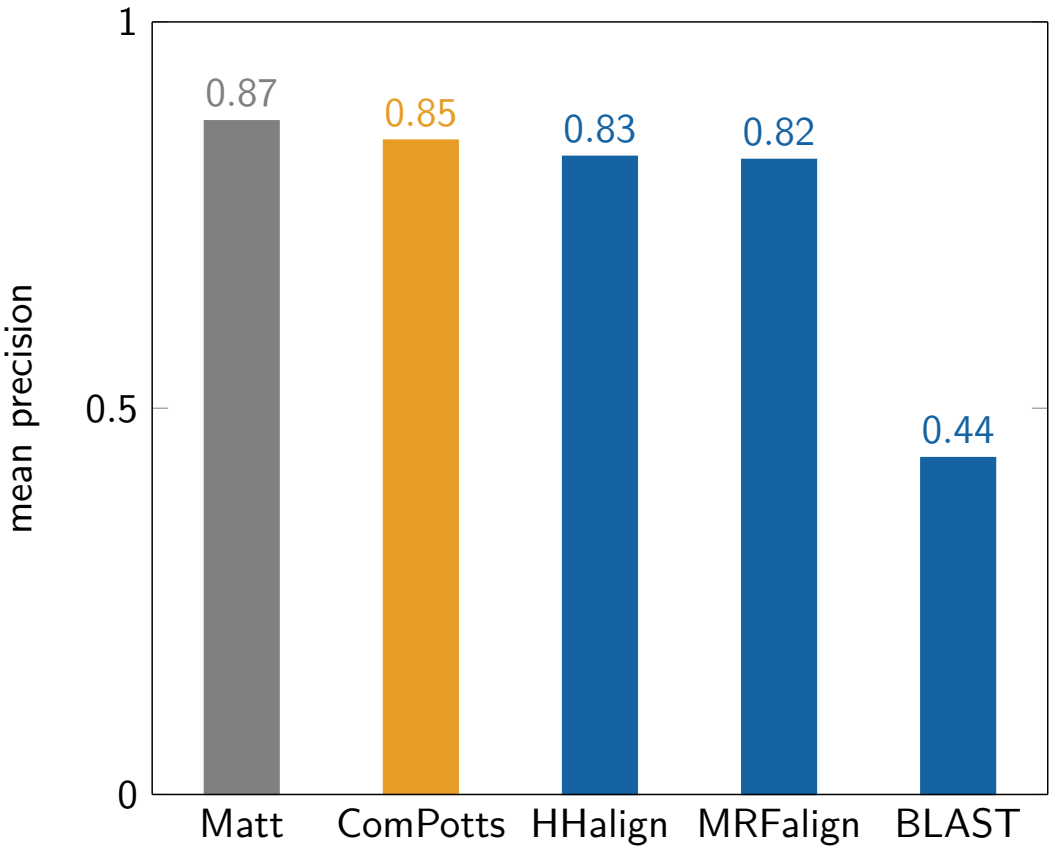
A better recall than our competitors...



$$\frac{\#\textit{correctly aligned columns}}{\#\textit{columns in reference alignment}}$$

- Better than HHalign in most cases (52/59)
- Better than MRFalign in 39 cases
- On average better than Matt !

... while still having a slightly better precision



$$\frac{\#\textit{correctly aligned columns}}{\#\textit{columns in test alignment}}$$

- Better than HHalign in 46 out of 59
- Better than MRFalign in 30 out of 59

Time considerations (new results after code optimization)

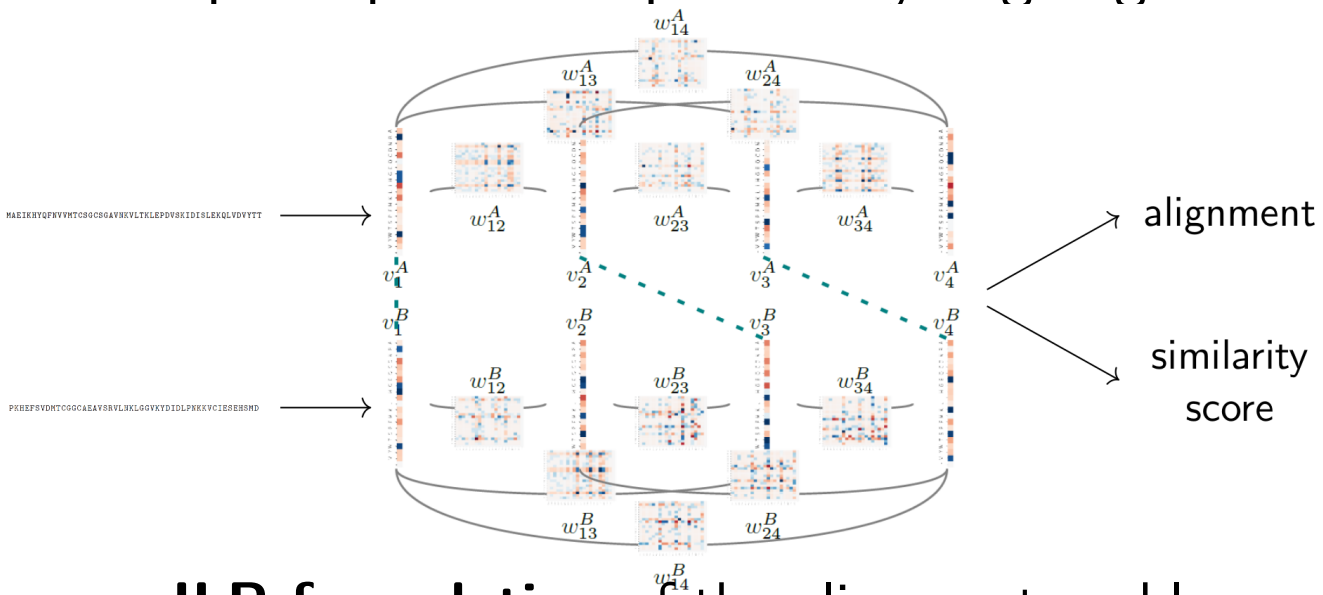
- On a Debian 9 virtual machine with 4 vCPUs, 8GB RAM:

	time (s)	**model dimension**	**alignment algorithm**
ComPotts	$3 < t < 58$	2D	exact
HHalign	$0.7 < t < 3.3$	1D	exact
MRFalign	$t < 0.2$	2D	heuristics

$\rightarrow$ As expected: higher computation time for an exact solution but **tractable** despite NP-completeness

Conclusion

ComPotts: compares protein sequences by aligning Potts models



- Based on an **ILP formulation** of the alignment problem
- Can yield **exact** solution in **tractable** time
- Encouraging preliminary results on quality of its alignments
- Suggest that **direct coupling information** can **improve** protein sequence **alignment**...
  - → ongoing work: validation at a larger scale
- ...and might **improve** sequence-based **homology search**
  - → discriminatory power of similarity score between two Potts models?

Thank you !

P.S. I'm looking for a postdoc as of 2021

✉ hugo.talibart@irisa.fr

Altschul, S. F. et al. “Basic local alignment search tool”. *Journal of molecular biology* 215.3 (1990), pp. 403–410.

Eddy, S. R. “Profile hidden Markov models.” *Bioinformatics (Oxford, England)* 14.9 (1998), pp. 755–763.

Steinegger, M. et al. “HH-suite3 for fast remote homology detection and deep protein annotation”. *BMC bioinformatics* 20.1 (2019), pp. 1–15.

Menke, M., B. Berger, and L. Cowen. “Markov random fields reveal an N-terminal double beta-propeller motif as part of a bacterial hybrid two-component sensor system”. *Proceedings of the National Academy of Sciences* 107.9 (2010), pp. 4069–4074.

Ma, J. et al. “MRFalign: protein homology detection through alignment of Markov random fields”. *PLoS computational biology* 10.3 (2014), e1003500.

Weigt, M. et al. “Identification of direct residue contacts in protein–protein interaction by message passing”. *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 67–72.

Monastyrskyy, B. et al. “New encouraging developments in contact prediction: Assessment of the CASP 11 results”. *Proteins: Structure, Function, and Bioinformatics* 84 (2016), pp. 131–144.

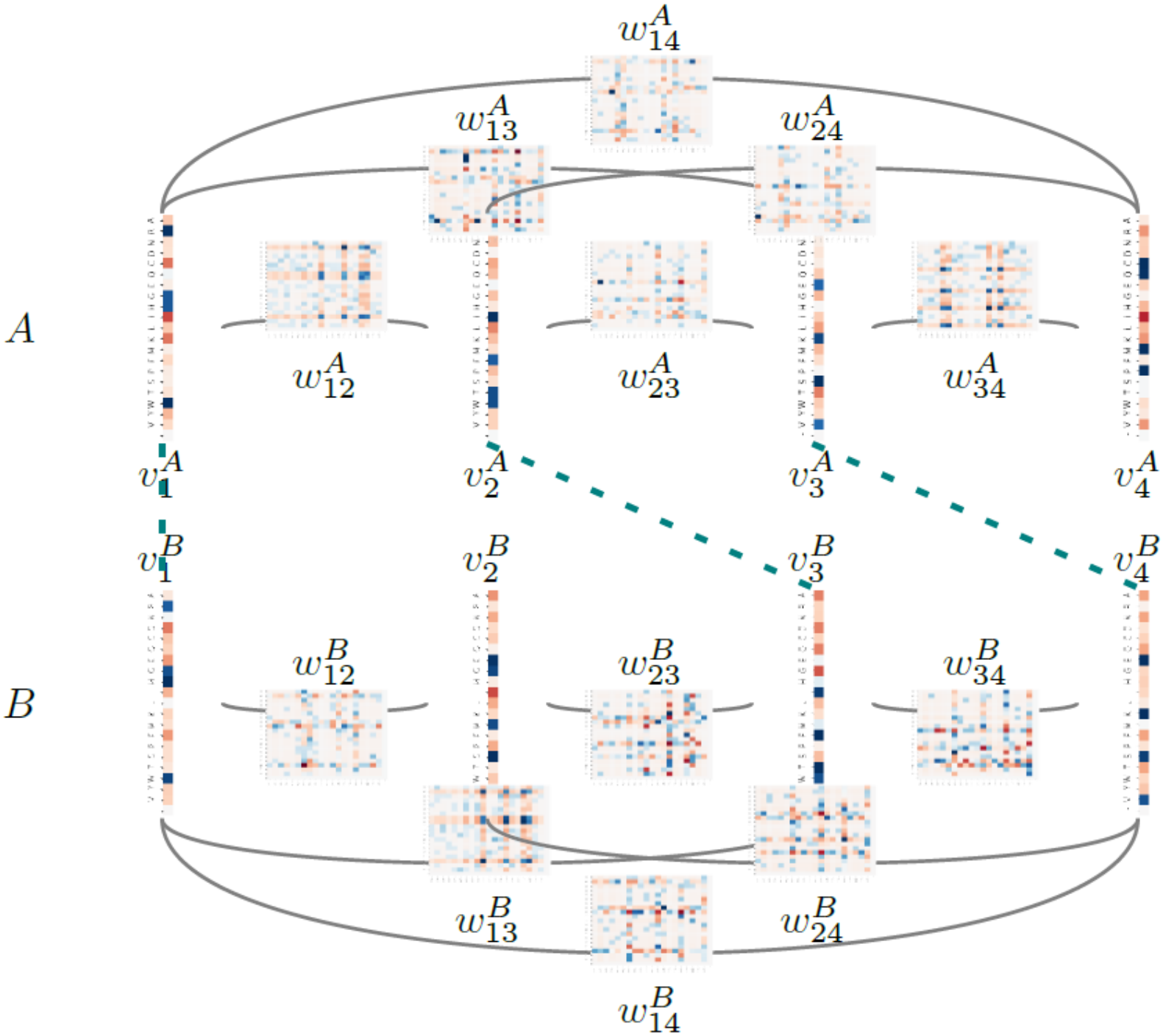
Vorberg, S. "Bayesian Statistical Approach for Protein Residue-Residue Contact Prediction". PhD thesis. Ludwig-Maximilians-Universität, 2017.

Wohlers, I., R. Andonov, and G. W. Klau. "DALIX: optimal DALI protein structure alignment". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10.1 (2012), pp. 26–36.

Thompson, J. D., F. Plewniak, and O. Poch. "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs." *Bioinformatics (Oxford, England)* 15.1 (1999), pp. 87–88.

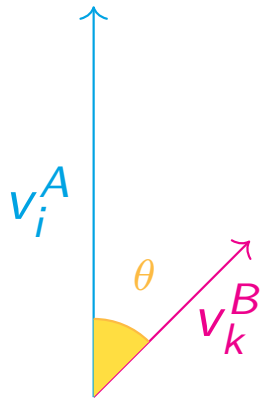
Andreeva, A. et al. "SISYPHUS—structural alignments for proteins with non-trivial relationships". *Nucleic acids research* 35.suppl_1 (2007), pp. D253–D259.

Menke, M., B. Berger, and L. Cowen. "Matt: local flexibility aids protein multiple structure alignment". *PLoS computational biology* 4.1 (2008).


w^A_{14}
w^A_{13}
w^A_{24}
A
w^A_{12}
w^A_{23}
w^A_{34}
v^A_1
v^A_2
v^A_3
v^A_4
v^B_1
v^B_2
v^B_3
v^B_4
B
w^B_{12}
w^B_{23}
w^B_{34}
w^B_{13}
w^B_{24}
w^B_{14}

Choice for $s_v(v_i, v_k)$ et $s_w(w_{ij}, w_{kl})$: scalar product

- $s_v(v_i^A, v_k^B) = \langle v_i^A, v_k^B \rangle$
  - $\rightarrow$ 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