PPalign: optimal alignment of Potts models representing proteins with direct coupling information

Hugo Talibart, François Coste, Mathilde Carpentier

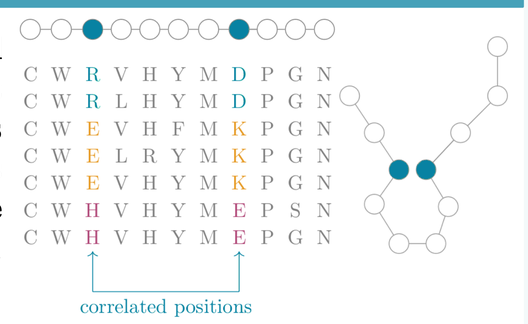
PPalign: optimal **alignment** of **Potts models** representing **proteins** with **direct coupling information**

Hugo Talibart[1,2], François Coste[2], and Mathilde Carpentier[1]

[1]Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, Sorbonne Université, EPHE, UA, CNRS
[2]Univ Rennes, Inria, CNRS, IRISA, Rennes, France

Introduction

To assign structural and functional **annotations** to the **ever increasing amount of sequenced proteins**, the main approach relies on **sequence-based homology search methods**, e.g. BLAST or the current state-of-the-art methods based on profile Hidden Markov Models, which rely on significant **alignments** of query sequences to annotated proteins or protein families. While powerful, these approaches do not **take coevolution between residues into account**. Taking advantage of recent advances in the field of contact prediction, our approach, recently published in BMC Bioinformatics [1], proposes to **represent proteins by Potts models**, which **model direct couplings between positions in addition to positional composition**, and to **compare proteins by aligning these models**. Due to non-local dependencies, the problem of aligning Potts models is hard and remains the main computational bottleneck for their use.



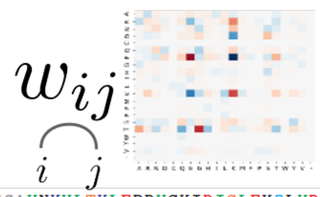
Inference of Potts models

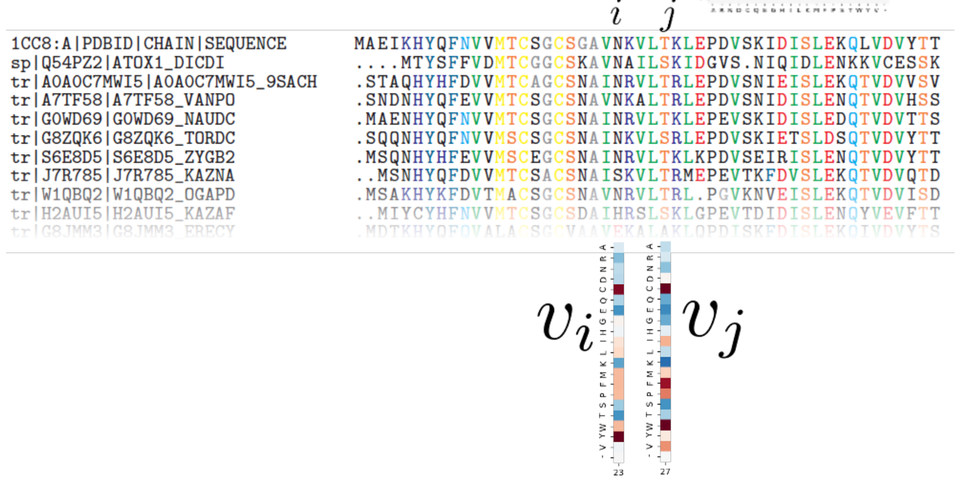
As introduced in Direct Coupling Analysis [2], a Potts model for a multiple sequence alignment (MSA) of homologous sequences can be defined as a statistical model whose probability distribution **maximizes Shannon entropy** and generates the **empirical single and double frequencies** of the MSA **as marginals**.

Its probability distribution has the following form:

Probability of sequence $x = x_1, \dots, x_L$ — Couplings — Normalization constant — Fields

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





Its parameters can be assigned a practical interpretation:

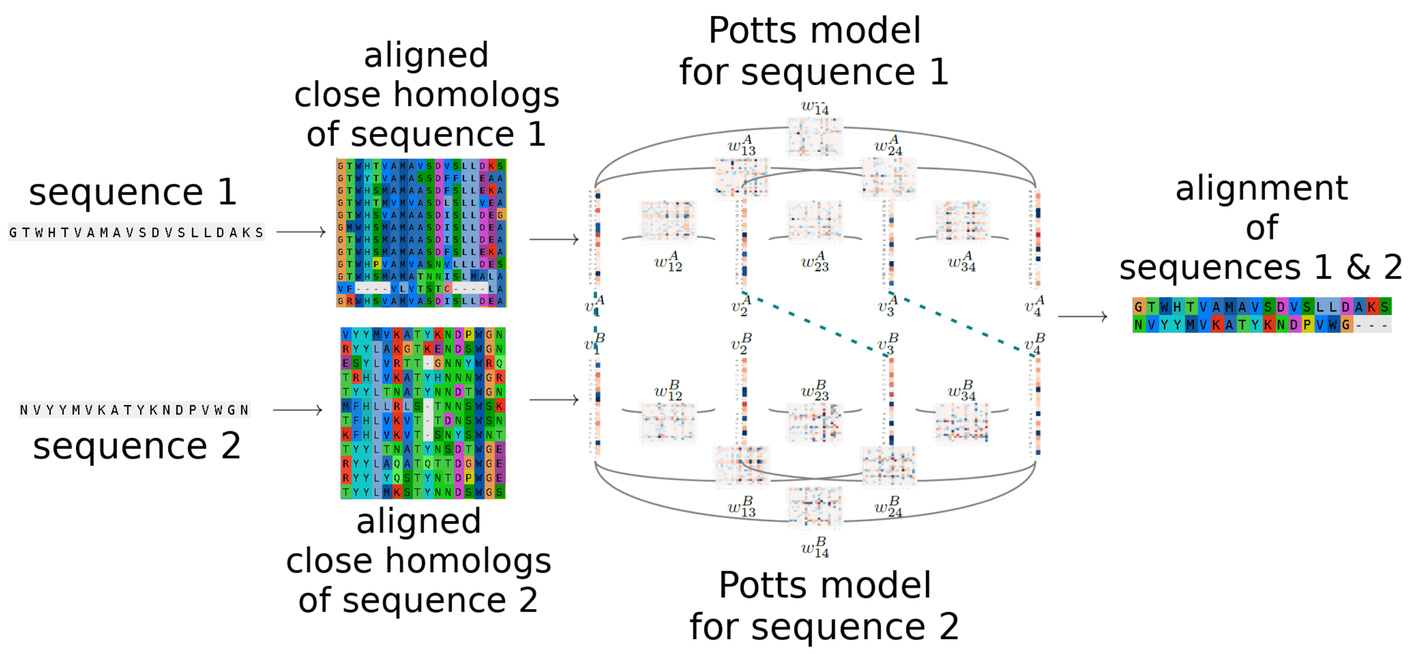
- $\boldsymbol{v} = \{v_i\}_{i=1,\cdots,L}$ are positional parameters termed "fields". ($v_i \in \mathbb{R}^q$)
  $v_i(a) \sim$ propensity of letter a to be found at position i.
- $\boldsymbol{w} = \{w_{ij}\}_{i,j=1,\cdots,L}$ are pairwise "coupling" parameters. ($w_{ij} \in \mathbb{R}^{q\times q}$)
  $w_{ij}(a,b) \sim$ compatibility of letters a and b at positions i and j

From protein sequence to Potts model:
One can get a Potts model for a sequence by inferring it on a MSA of its close homologs. In this study, homologs were retrieved using HHblits [3], then MSAs were processed by filtering at 80% identity, setting a depth threshold at 1000, trimming columns with $> 50\%$ gaps, and fed to CCMpredpy [4] to infer Potts models.

sequence → HHblits → MSA → trim filter → train MSA → CCMpredPy → Potts model

Alignment of Potts models

Potts models provide an interesting alternative to pHMMs for sequence comparison since they can model pairwise dependencies in addition to positional conservation. To investigate on their performances in alignment-based homology detection, we introduced PPalign, a pairwise Potts model alignment method. PPalign can provide in tractable time an alignment of two protein sequences which takes both positional composition and pairwise dependencies into account by aligning Potts models representing them.



Alignment as an ILP problem

We built on the work of Wohlers et al. [5], initially dedicated to protein structure alignment, to propose an Integer Linear Programming formulation for the alignment of two Potts models A and B of parameters (v^A, w^A) and (v^B, w^B).

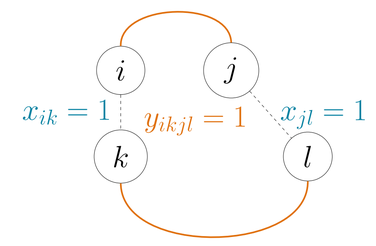
$$\max \sum_{i=1}^{L_A}\sum_{k=1}^{L_B}\langle v_i^A, v_k^B\rangle \boldsymbol{x_{ik}} + \sum_{i=1}^{L_A-1}\sum_{j=i+1}^{L_A}\sum_{k=1}^{L_B-1}\sum_{l=k+1}^{L_B}\langle w_{ij}^A, w_{kl}^B\rangle \boldsymbol{y_{ikjl}}$$

alignment constraints: one-to-one and non-crossing

$$\text{s.t. } x_{ik} \geq \sum_{r.s\in row_{ik}(j)} y_{ikrs} \quad j\in[i+1,L_A], i\in[1,L_A-1], k\in[1,L_B-1]$$
$$x_{ik} \geq \sum_{r.s\in col_{ik}(l)} y_{ikrs} \quad l\in[k+1,L_B], i\in[1,L_A-1], k\in[1,L_B-1]$$
$$x_{ik} \geq \sum_{r.s\in row_{ik}(j)} y_{rsik} \quad j\in[1,i-1], i\in[2,L_A], k\in[2,L_B]$$
$$x_{ik} \geq \sum_{r.s\in col_{ik}(l)} y_{rsik} \quad l\in[1,k-1], i\in[2,L_A], k\in[2,L_B]$$
$$x_{ik} \leq \sum_{\substack{r.s\in row_{ik}(j)\\ s(A_{ri},B_{sk})\leq 0}} (y_{rsik} - x_{rs}) + 1 \quad j\in[1,i-1], i\in[2,L_A], k\in[2,L_B]$$
$$\sum_{l=1}^{k} x_{il} + \sum_{j=1}^{i-1} x_{jk} \leq 1 \quad i\in[1,L_A], k\in[1,L_B]$$
$$x, y \text{ binary}$$

Decision variables: x, y binary

- $x_{ik} = 1$ iff node i and node k aligned
- $y_{ikjl} = 1$ iff edges (i,j) and (k,l) aligned



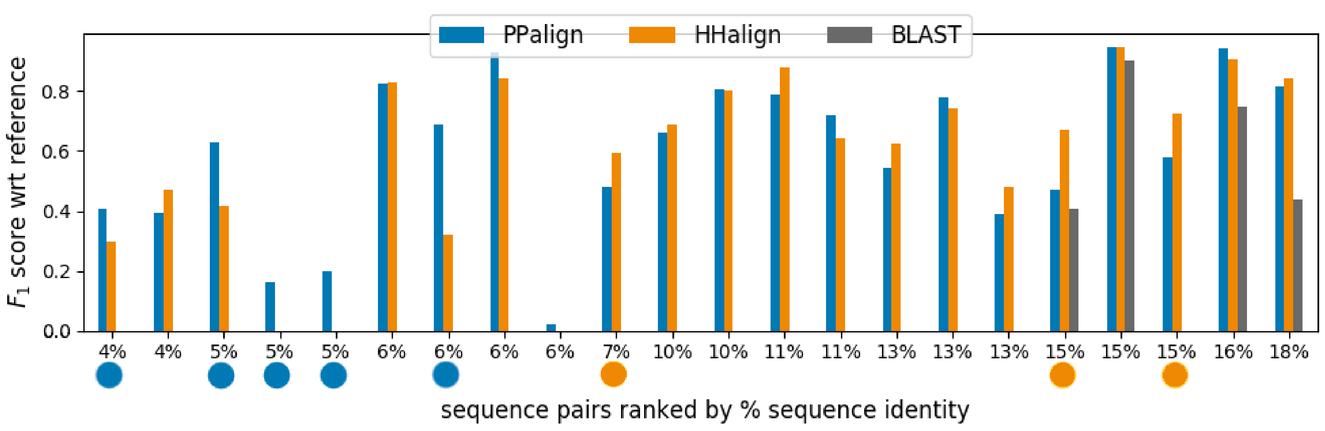
Similarity score based on **scalar product**:

- $\langle v_i^A, v_k^B\rangle = \sum_a v_i^A(a) v_k^B(a)$
- $\langle w_{ij}^A, w_{kl}^B\rangle = \sum_a \sum_b w_{ij}^A(a,b) w_{kl}^B(a,b)$

Using their efficient solver, the exact solution of this ILP within a chosen epsilon range can be computed in tractable time.

PPalign improves alignment quality of remote homologs

We assessed the quality of PPalign's alignments on a benchmark of low sequence identity (4-18%) pairwise sequence alignments based on reference structural alignments from SISYPHUS [6] using the F_1 score metric:
$F_1 = \frac{2PR}{P+R}$ where $P = \frac{\#\ correctly\ aligned\ pairs}{\#\ aligned\ pairs\ in\ computed\ alignment}$, $R = \frac{\#\ correctly\ aligned\ pairs}{\#\ aligned\ pairs\ in\ reference\ alignment}$



PPalign's alignments achieve a better mean F_1 score than HHalign's [3] alignments of pHMMs built on the same MSAs (0.600 vs 0.578), while BLAST [7] fails to align most sequences (mean F_1 score of 0.113). PPalign outperforms HHalign in 12/22 alignments (4 significantly ●), with better F_1 scores when sequence identity is lower. It is mostly outperformed when MSAs have more gaps (●).

Conclusion

- PPalign initiates a **new approach for remote homology search**
- Similarly to HHalign from HHsuite...
- ...with the addition of **long distance sequence correlations reflecting higher order constraints**
- **Tractable time** (1′37″ on average in this study) despite computationally hard
- Encouraging results in terms of **alignment quality**

Code and benchmark: `https://www-dyliss.irisa.fr/ppalign`

Ongoing work and perspectives

Our current work now focuses on the inference of Potts models more suitable for pairwise comparison, which was not their original purpose. By improving their robustness to sampling variations and seeking a more canonical form, we are hoping to improve these already encouraging results and to better assess the contribution of direct couplings.

This research provides ground work for future exciting applications such as a homology search package which would take coevolution into account, or Potts model annotation databases (e.g. for viral proteins).

[1] Hugo Talibart and François Coste. "PPalign: optimal alignment of Potts models representing proteins with direct coupling information". In: *BMC bioinformatics* 22.1 (2021), pp. 1–22.
[2] Martin Weigt et al. "Identification of direct residue contacts in protein–protein interaction by message passing". In: *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 67–72.
[3] Martin Steinegger et al. "HH-suite3 for fast remote homology detection and deep protein annotation". In: *bioRxiv* (2019), p. 560029.
[4] Susann Vorberg, Stefan Seemayer, and Johannes Söding. "Synthetic protein alignments by CCMgen quantify noise in residue-residue contact prediction". In: *PLoS computational biology* 14.11 (2018), e1006526.
[5] Inken Wohlers, Rumen Andonov, and Gunnar W Klau. "DALIX: optimal DALI protein structure alignment". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10.1 (2012), pp. 26–36.
[6] Antonina Andreeva et al. "SISYPHUS—structural alignments for proteins with non-trivial relationships". In: *Nucleic acids research* 35.suppl_1 (2007), pp. D253–D259.