PPalign: optimal alignment of Potts models representing proteins with direct coupling information
Hugo Talibart, François Coste, Mathilde Carpentier

To cite this version:
Hugo Talibart, François Coste, Mathilde Carpentier. PPalign: optimal alignment of Potts models representing proteins with direct coupling information. ISMB 2022 - 30th Conference on Intelligent Systems for Molecular Biology, Jul 2022, Madison, United States. pp.1-1. hal-03926272

HAL Id: hal-03926272
https://inria.hal.science/hal-03926272v1
Submitted on 6 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
Palign: optimal alignment of Potts models representing proteins with direct coupling information

Hugo Talibart1,2, François Coste2, and Mathilde Carpentier1
1Institut de Systématique, Évolution, Biodiversité (ISYE), Muséum National d’Histoire Naturelle, Sorbonne Universités, ÉPHE, UA, CNRS
2Univ 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:

\[
P(x | v) = \frac{1}{Z} \exp \left( \sum_{i \neq j} \sum_{k} w_{ij}(x_i, x_j) + \sum_{i} v_i(x_i) \right)
\]

Its parameters can be assigned a practical interpretation:

- \(v_i \), \(v_j \) are positional parameters termed "beads". \(v_i(x)\) is the probability of letter \(a\) to be found at position \(i\).
- \(w_{ij}(\cdot, \cdot)\) are pairwise "coupling" parameters. \(w_{ij}(a, b)\) is the probability of letters \(a\) and \(b\) at positions \(i\) and \(j\). A common choice for \(w_{ij}(a, b)\) is the log odds ratio.

From protein sequence to Potts sequence

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, then MSAs were processed by filtering at 80% identity, setting a depth threshold at 1000, trimming columns with >50% gaps, and fed to CCMpred [2] to infer Potts models.

Palign improves alignment quality of remote homologs

We assessed the quality of PPalign’s alignments on a benchmark of low identity (4-18%) pairwise sequence alignments based on reference structural alignments from SISYPHUS [3] using the F1 score metric:

\[
F1 = \frac{2 PR}{P + R}
\]

where \(P\) is the number of correctly aligned pairs, \(R\) is the number of correctly aligned residues, and \(PR\) is the precision-recall.

Palign’s alignments achieve a better mean \(F1\) score than HHalign’s alignments of PHMMs built on the same MSA (0.60 vs 0.57%), while BLAST fails to align most sequences (mean \(F1\) score of 0.113). PPalign outperforms HHalign in 12/22 alignments (4 significantly), with better \(F1\) scores when sequence identity is lower. It is most outperformed when MSAs have more gaps.

Alignment as an ILP problem

We built on the work of Wohlers et al. [4] 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 \((v_B, w_B)\).

\[
\text{maximize} \quad \sum_{i=1}^{L_A} \sum_{j=1}^{L_B} (v_A(x_i) v_B(y_j) + \sum_{i} \sum_{j} \sum_{k} \sum_{l} w_{ij}(x_i, x_j, y_k, y_l) y_{kl})
\]

subject to

- \(x_i = 1\) iff node \(i\) and node \(k\) aligned
- \(y_{ij} = 1\) iff edges \((i, j)\) and \((k, l)\) aligned

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

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 (still average in the study) despite computationally hard
- Encouraging results in terms of alignment quality

Code and benchmark: https://www-sylisse.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).