Supervised learning to tune simulated annealing for in silico protein structure prediction

Alejandro Marcos Alvarez, Francis Maes and Louis Wehenkel
Department of Electrical Engineering and Computer Science - University of Liège, Belgium
Contact information: amarcos@ulg.ac.be, http://www.montefiore.ulg.ac.be/~ama/

Simulated annealing is a widely used stochastic optimization algorithm whose efficiency essentially depends on the proposal distribution used to generate the next search state at each step. We propose to adapt this distribution to a family of **parametric optimization** problems by using **supervised machine learning** on a sample of search states derived from a set of typical runs of the algorithm over this family. We apply this idea in the context of *in silico* protein structure prediction.

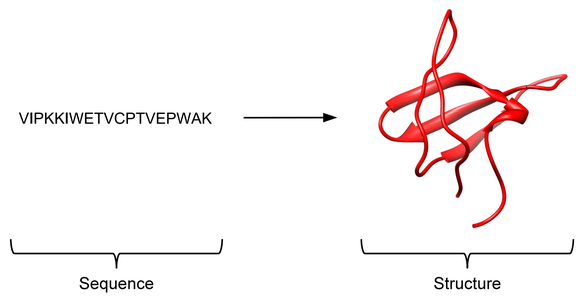
Motivation

Protein structure prediction is a **topical** and **challenging open problem** in bioinformatics. The significance of this problem is due to the importance of studying protein structures in biomedical research in order to **improve our understanding** of the human physiology and to accelerate **drug design** processes.

The most reliable way to determine protein structures is to use **experimental methods** such as X-ray crystallography or NMR spectroscopy, which are however **expensive** and **time consuming**, and hence the design of in silico protein structure prediction methods has become a very active research field.

General problem statement

The problem we are considering is ***in silico* protein structure prediction**, which amounts to predicting the 3D coordinates of each atom in the protein given its amino acid sequence.



Characteristics

- modeled as a parametric optimization problem parameterized by λ
  - high-dimensional for usefully sized proteins;
  - $\lambda \equiv$ the amino acid sequence of the protein;
  - $s_\lambda \equiv$ current state (structure) of the protein.
- cost function $\equiv$ the energy function $\mathcal{E}$ of the protein
  - large number of local minima;
  - global minimum of $\mathcal{E}$ corresponds to the sought structure;
  - $\mathcal{E}$ includes all constraints;
  - evaluating $\mathcal{E}$ can be long.
- optimization algorithm $\equiv$ simulated annealing (SA) [4]
  - proteins-specific operators $o \in \mathcal{O}$ used to modify the structure.

Optimization algorithm

Algorithm 1: Simulated annealing

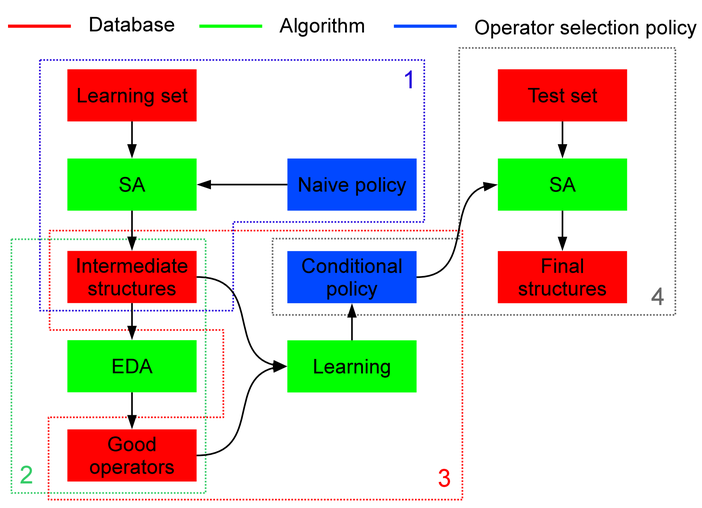
Let B be a budget of iterations, $\mathcal{E}(\cdot)$ the oracle evaluating the energy and $T(i)$ a non increasing cooling schedule defined over $\{1, \ldots, B\}$.
Input: λ the problem instance, $\mathcal{S}_\lambda$ its solution space, $s^0 \in \mathcal{S}_\lambda$ the chosen initial state, $p(o)$ is a proposal distribution used to sample operators

$s = s^0$;
$e = \mathcal{E}\left(s^0\right)$;
for $i = 1 \ldots B$ **do**
  propose $o \in \mathcal{O}$ s.t. $o \sim p(o)$;
  $s' = o\left(s\right)$;
  $e' = \mathcal{E}\left(s'\right)$;
  with probability $\min\left(1, \exp\left(\frac{e-e'}{kT(i)}\right)\right)$ **do**
    $s = s'$;
    $e = e'$;
  end
end for
return s

Supervised learning based framework

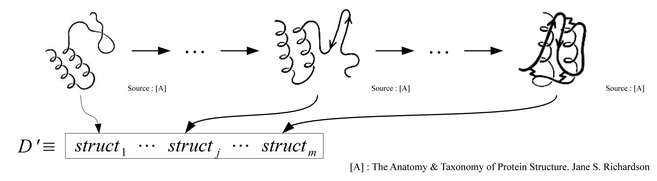
Observations
SA's efficiency critically depends on $p(o)$ (naive policy) !

What we are going to do
Use supervised machine learning to create a conditional probability distribution $p(o \mid s)$ (conditional policy) and use it instead of $p(o)$.



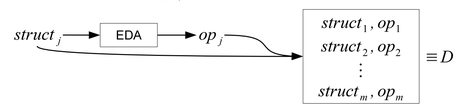
Work phases

1. Generate intermediate structures
Apply SA with $p(o)$ on the learning set and save intermediate structures during optimization.

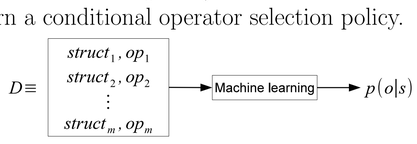


[A] : The Anatomy & Taxonomy of Protein Structure. Jane S. Richardson

2. Generate good operators
For each intermediate structure, use an EDA [5] to discover good operators (that decrease energy).



3. Learn the conditional policy
Use D to learn a conditional operator selection policy.



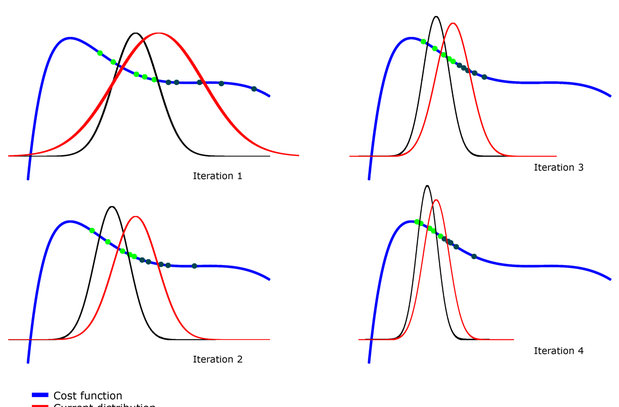
4. Assessment
Use $p(o \mid s)$ with SA on the test set and evaluate performance.

Conditional distribution

- discrete parameters: maximum-entropy classifier [2];
- continuous parameters:

$$\mu = \langle \theta_\mu; \phi(s_\lambda) \rangle;$$
$$\sigma = \log\left\{1 + \exp\left(-\langle \theta_\sigma; \phi(s_\lambda) \rangle\right)\right\};$$
$$p_\theta\left(\gamma \mid \phi(s_\lambda)\right) \sim \mathcal{N}_\theta(\mu, \sigma).$$

Estimation of distribution algorithm (EDA)



Results

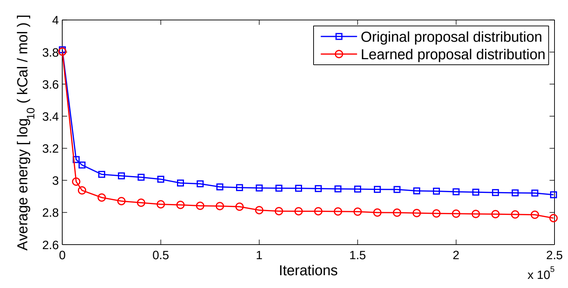


Figure 1: Evolution of average energy of the test set proteins during one optimization run.

- The learning set is composed of 100 proteins randomly selected from the database PSIPRED [3].
- The test set is composed of 10 proteins randomly selected from the database PSIPRED [3].
- The parameters of SA were determined by a rule of thumb based on what can be found in official Rosetta tutorials (more details in [1]).
- The learned conditional distribution outperforms the other one in terms of **convergence speed** and of **final result**.
- These results are promising but the structures predicted after one such learning iteration are still very different from the real structures.

Conclusions and future work

- **Improvement**
  Machine learning can improve optimization performance.
- **Promising results**
  In the context of *in silico* protein structure prediction.
- **Learning for search**
  Learning a good way to search through the state space of a problem.
- **General**
  Can be applied to other optimization problems and search methods.
- **Local vs global information**
  Better efficiency may be expected if learning could take into account *global* information (in this work, *local* information is used).
- Future work includes
  - optimization: fine tuning of parameters, other algorithms;
  - learning: improvement of features and model selection.

References

[1] A. Marcos Alvarez. Prédiction de structures de macromolécules par apprentissage automatique. Master's thesis, University of Liège, Faculty of Engineering, 2011.

[2] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[3] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195 – 202, 1999.

[4] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science, New Series*, 220(4598):671–680, 1983.

[5] P. Larrañaga and J. A. Lozano. *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Springer, October 2002.

This work is supported by the Belgian Science Policy Office and was funded by a FRIA scholarship of the F.R.S-FNRS.