Learning Biophysically-Motivated Parameters for Alpha Helix Prediction

Massachusetts Institute of Technology

Blaise Gassend, Charles W. O'Donnell, William Thies, Andrew Lee, Marten van Dijk, Srinivas Devadas

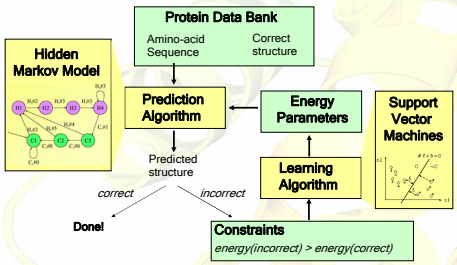
MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139 USA

CSAIL

Overview

Our goal is to create an accurate protein secondary structure predictor based on an intuitive and biophysically-motivated energy model

- Secondary structures are determined by a prediction algorithm that minimizes a global free-energy cost function that is a linear combination of elementary free energy parameters.
- We are interested cost functions that use as few parameters as possible, given that all the parameters are well-understood and biophysically-motivated.
- Since there is little direct experimental data for determining our free-energy parameters, we use Support Vector Machines (SVMs) to learn these cost function parameters with an iterative constraint-based optimization (below).



Cost Models

Linear models

We assume that the free-energy function is a sum of elementary interactions. This agrees with many mathematical models of energy force fields that control protein folding. For example, electrostatic, Van der Waals, stretch, bend, and torsion forces all are described by the sum of energy terms for each pair of molecular elements.

This simplification also allows the constraint-based optimization problem to be solved much more elegantly and efficiently since a dot product may be taken of every energy characteristic we are interested in (the feature function Ψ).

$$G_w(x,y) = \sum_i w_i \Psi_i(x,y) = \langle w, \Psi(x,y) \rangle$$

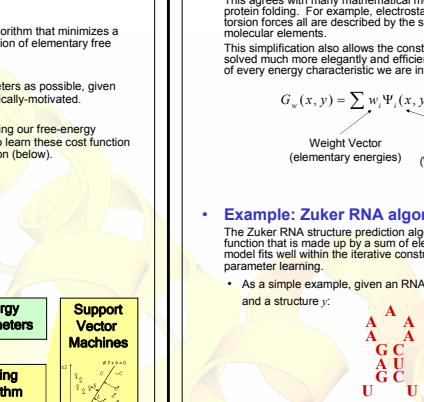
Weight Vector (elementary energies)

Feature Vector (Weights used in structure y for sequence x)

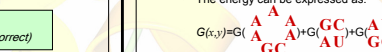
Example: Zuker RNA algorithm

The Zuker RNA structure prediction algorithm is one good example of a cost function that is made up by a sum of elemental energies. Such an energy model fits well within the iterative constraint-based optimization technique of parameter learning.

- As a simple example, given an RNA sequence x: UGAGAAAACUCU and a structure y:



- The energy can be expressed as:



HMM model for all-alpha proteins

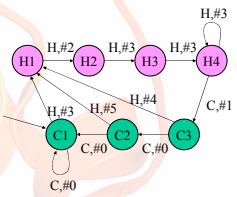
In our experimentation, we used a simple HMM (FSM) to compute a free-energy cost function that could be used to predict whether an alpha helix exists or not at each given residue.

As seen below, each state represents whether a residue is in a helix (H) region or a coil (C) region, as the amino-acid sequence is parsed from left to right. For instance, state C1 is the start state and state H4 correspond to helices more than 4 residues long. States C1, C2, and C3 represent recognition states for full alpha helices.

Short coils are permitted, but helices shorter than 4 residues are not allowed (none in dataset).

The cost function is comprised of features that arise from each transition within the HMM.

In total, 302 unique features can occur in our model, as explained adjacently. We experimented with more complicated features (such as pairwise interaction of nearby residues), but found the predictive accuracy to drop (most likely from over-learning).



Name	Number of features	Comment
A	1	Penalty for very short coil
B	1	Penalty for short coil
H_R	20	Energy of residue R in a helix
C_R	140	Energy of residue R at position i relative to C-cap
N_R	140	Energy of residue R at position i relative to N-cap
Total	302	

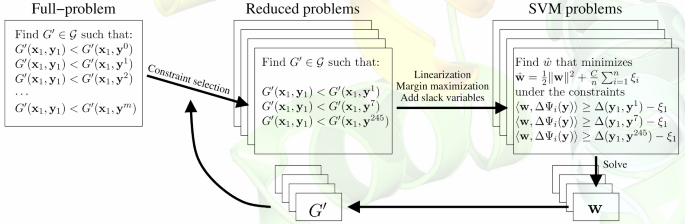
#0	0	Coil (zero reference)
#1	$\sum_{i=1}^{3} C^{i-1}_{R_{n-i}}$	End of helix
#2	$H_{R_n} + \sum_{i=3}^{4} C^{i-1}_{R_{n-i}}$	Start of helix
#3	H_{R_n}	Residue in helix
#4	$H_{R_n} + A$	Helix after very short coil
#5	$H_{R_n} + B$	Helix after short coil

Also applicable to

This approach can also be immediately applied to use more complex features and cost functions created from Context Free Grammars (CFGs), which allow for more flexibility than HMMs.

Multi-tape CFGs may also be used with this approach to improve domain-specific predictors such as some for trans-membrane proteins.

Iterative Constraint-Based Optimization



Formulation as an optimization problem

A protein's secondary structure can be found by minimizing a function G made up of elementary free-energy parameters. To find the optimal parameter values $\boldsymbol{w}$, the elementary free-energies must satisfy an exponentially large system of inequalities: for each training sequence x_i, the correct secondary structure y_i must have a lower free-energy $G(x,y)$ than for any incorrect secondary structure y.

Iterative constraint-based approach

Directly computing the optimal parameters given this system of inequalities is intractable, so we select a tractable subset of inequalities.

We begin with zero constraints, arbitrary parameters, and minimize G to determine a structure $\vec{y}$ for a sequence x_i.

If $G(x_i,\vec{y}) < G(x_i,y_i) + \varepsilon$, then the parameters are satisfactory for x_i and the next x_i is checked, otherwise, a new constraint is added to the system of inequalities for this mismatch.

Once all the constraints have been added for all x_i, a new set of parameters $\boldsymbol{w}$ are computed from this system and the next iteration begins. This continues until all x_i are satisfied, or a suitable termination condition is met.

Support Vector Machines

The subset of inequalities may not have any solutions because there might not exist a set of free-energies compatible with all training structures. Alternately, if the problem does have solutions, it will probably have many.

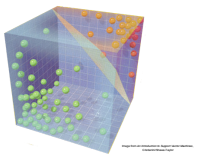
SVMs techniques of margin maximization and slack variables can translate this system into a quadratic program with one unique solution.

This translation finds the set of parameters $\boldsymbol{w}$ that minimizes $\frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\frac{\varepsilon}{\xi_i}$ given the constraints.

Complete algorithm

```
Input: (x_1,y_1),...,(x_n,y_n), C, ε
S_i ← ∅ for all 1 ≤ i ≤ n
repeat {
  for i = 1,...,n do {
    Set up the cost function H(y) = Δ(y_i,y) − ⟨w, ΔΨ_i(y)⟩
    Compute ŷ = argmax_{y∈G} H(y)
    Compute ξ_i = max{0, max_{y∈S_i} H(y)}
    If H(ŷ) > ξ_i + ε then {
      S_i ← S_i ∪ {ŷ}
      w ← optimize over S = ∪_i S_i
    }
  }
} until no S_i has changed during iteration
```

Algorithm 1: Algorithm for iterative constraint based optimization.



Results

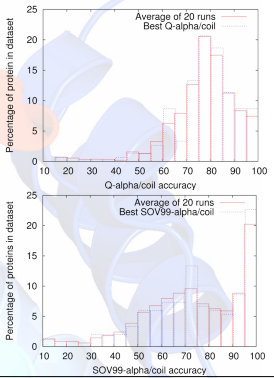
Predictor status

- Current prediction algorithm accepts a residue sequence and predicts whether or not each residue is a member of an alpha helix.
- The predictor uses the HMM model described above. The 302 parameters are learned from an arbitrary set of known sequence/label pairs, where the label is determined from a DSSP parse of the protein's PDB file (DSSP's *H* state, excluding 3_{10} and π helices).
- The current predictor does ***not*** use any type of multiple sequence alignment techniques which commonly boost the final predictive accuracy as much as 5-7%.

Accuracy

- Trained cost function parameters on 300 non-homologous (sequence identity <30%) all-alpha proteins taken from EVA (http://salilab.org/evaires/weeks.html#unique).
- Randomly splitting the test and train set, an average of 20 runs shows an accuracy of $Q_{a/c}$ = 77.6% and $SOV99_{a/c}$ = 73.4%.

	$SOV99_{a/c}$ train set (%)	$SOV99_{a/c}$ test set (%)	$Q_{a/c}$ train set (%)	$Q_{a/c}$ test set (%)	Training time (s)
Best run for SOV99	76.4	75.1	79.6	78.6	123
Average of 20 runs	75.1	73.4	79.1	77.6	162
Std. dev. of 20 runs	1.0	1.4	0.6	0.9	30



Conclusions

Currently Achieved

- This work describes a general learning technique that can be applied to a variety of cost functions and energy models that require biophysically meaningful parameter values that are not available in the experimental corpus.
- An evaluation of this technique on all-alpha proteins shows promising results.

Future Work

- Applying techniques to proteins containing beta sheets, and using more general cost functions such as CFGs and Multi-tape CFGs.

References

More Information can be found at http://protein.csail.mit.edu

1. B. Gassend, C. W. O'Donnell, W. Thies, A. Lee, M. van Dijk, and S. Devadas. ***Secondary Structure Prediction of All-Helical Proteins Using Hidden Markov Support Vector Machines***. In Technical Report MIT-CSAIL-TR-2005-060, MIT, October 2005.
2. I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. ***Support Vector Machine Learning for Interdependent and Structured Output Spaces***. In ICML03: Proceedings of the 20th International Conference on Machine Learning, 2004.
3. A. Zemla, C. Venclovas, K. Fidelis, and B. Rost. ***A Modied Denition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment***. In Proteins, 34(2), 1999.

Corresponding authors: devadas@mit.edu, gassend@mit.edu, cwo@mit.edu