Routes are trees: The parsing view on protein folding

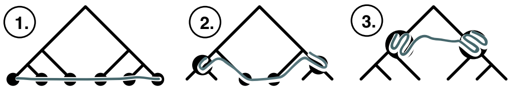
Julia Hockenmaier*, Aravind Joshi* and Ken Dill**

*{juliahr,joshi}@cis.upenn.edu, University of Pennsylvania; **dill@maxwell.ucsf.edu; University of California at San Francisco

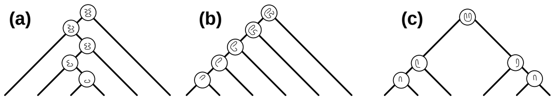
Protein folding is a hierarchical process that builds global from local structures and is parallel at first but serial at the end. Yet, the standard description of folding routes as linear sequences of events fails to capture the parallel, recursive, nature of this process. We propose that folding routes should instead be thought of as trees, which also provides a unified account of growth and assembly. Trees are used to represent the grammatical structure of language, and we show that, like natural language sentences, proteins have specific "constituent structures", a concept that generalizes the idea of autonomous folding units.

We demonstrate that the CKY parsing algorithm [2, 5] is an efficient method to find all direct routes to the native state. CKY implements a greedy search of locally optimal conformations. The physical plausibility of this strategy is validated by the fact that it predicts the results of [4] that folding speed depends on native contact order. Since CKY returns all native routes, we can quantify the kinetic accessibility of the native state. We focus on the HP model [3], but believe that our findings can in principle be carried over to more detailed structural representations.

Folding routes are trees In a folding route tree, each node corresponds to a substring of the sequence and represents a step in the folding process in which the structure of its substring was found from the structures of its children. Different branches represent independent (simultaneous) folding events. The chain "moves" from the leaves upwards:



The tree shape distinguishes growth (a,b) from assembly (c):



The HP model The HP model [3] is a lattice model which assumes only two kinds of monomers: hydrophobic (H) and polar (P). The energy of a conformation is determined by contacts between H monomers ω_i, ω_j that arise if ω_i and ω_j are placed on adjacent lattice sites. Each HH-contact contributes -1 to the energy. We only consider sequences with a unique lowest-energy conformation (native state) on the 2D lattice.

Adapting CKY to the HP model CKY is a dynamic programming algorithm that identifies the native structure of a sequence recursively by splitting it into short substrings and combining their structures, which are stored in a chart (a table that maps substrings from i to j to structures). The shape of the chart is similar to that of contact maps (Fig.1).

The sequence is split into n substrings that contain one H each. The chart is an $n \times n$ array. Each cell $chart[i][i]$ is filled with all conformations of the ith substring. For $\Delta = 1..n$, and for $i = 1..n$, each cell $\textsc{Target} = chart[i][i+\Delta]$ is filled by combining the entries of all pairs of cells $\langle \textsc{Left}, \textsc{Right} \rangle = \langle chart[i][i+k], chart[k+1][i+\Delta] \rangle$ $(0 \leq k < i+\Delta)$.

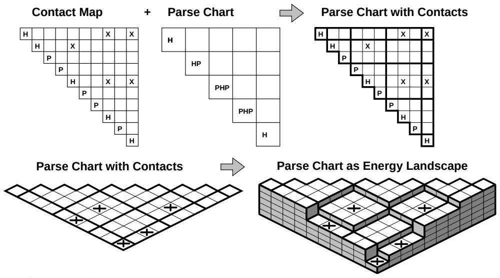

Contact Map
+
Parse Chart
Parse Chart with Contacts
H
HP
PHP
PHP
H
P
X
Parse Chart with Contacts
Parse Chart as Energy Landscape

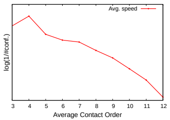
Figure 1: **Charts, contact maps and landscapes**

Two entries $\mathcal{L} \in \textsc{Left}$ and $\mathcal{R} \in \textsc{Right}$ are combined by appending all (rotational, translational) variants of $\mathcal{R}$ to any free site adjacent to the site of $\mathcal{L}$'s last monomer. The search is greedy: of the viable conformations that result from this combination only those with the lowest energy are entered into $\textsc{Target}$. We can also view the chart as an energy landscape which maps substrings to energies (Fig.1) CKY has succeeded if $chart[1][n]$ contains the native state. Each entry $\mathcal{E}$ in $chart[i][j]$ $(j > i)$ has a list of pairs of backpointers to the corresponding $\langle \mathcal{L}, \mathcal{R} \rangle$, which provides a compact representation of the folding routes, so that $\textsc{Target}$ only needs to contain one instance of each distinct $\mathcal{E}$. If $\mathcal{E}$ has m pairs of backpointers $\langle \mathcal{L}_i, \mathcal{R}_i \rangle$, it has $\phi(\mathcal{E})$ folding routes, with $\phi(\mathcal{E}) = \sum_{i=1}^{m} \phi(\mathcal{L}_i) \cdot \phi(\mathcal{R}_i)$ for $m > 0$, otherwise $\phi(\mathcal{E}) = 1$. By not combining certain cells, the search is constrained to low initial-contact-order (ICO), "zipping" [1] routes. Since cells are filled independently, CKY is easily parallelized.

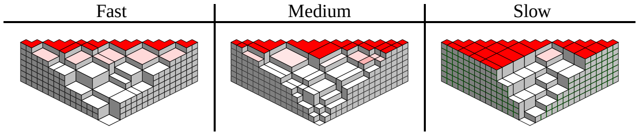
CKY is efficient and accurate The number of distinct conformations (and thus CKY's worst-case complexity) is exponential in the sequence length. But the greedy search and ICO restrictions are very effective:[1]

ICO	Accuracy (NS found)	Amount of search ($\langle S/MaxConf \rangle$)
5	93.1%	0.5%
7	95.2%	1.0%
9	96.3%	1.4%
11	96.6%	2.5%

CKY's speed depends on native CO The folding rate of real proteins decreases with native contact order [4]. We define CKY's folding rate as $1/S$ (S: number of conformations searched to fill the chart). This quantity also decreases with native CO (20mers, ICO=11):



CKY's rate depends on the chart's energy landscape. Fast folders have a steep funnel. Slow folders have a flat landscape. Local, low-CO, contacts give rise to a funnel, resulting in the observed behavior:



Proteins have constituent structures In linguistics, trees are used to indicate the grammatical, or constituent, structure of sentences. In our algorithm, the probability that individual cells in the chart lie along native routes shows that HP sequences can also have very specific "constituent structures" – all routes contain similar or identical partial structures In the parsing perspective, autonomous folding units are special cases of partial structures (constituents) that are on some, or all, native routes and contain only one optimal conformation.

Kinetic accessibility From $\phi(NS)$ (the number of native routes), we estimate the kinetic accessibility of the native state as $\kappa_{NS} = \phi(NS)/\mathcal{C}$, where the Catalan number $\mathcal{C}$ is the number of binary trees with n leaves. The lower κ_{NS}, the higher the probability that the chain misfolds and unfolds before finding a native route.

This research is supported by NSF ITR grant 0205456.

[1] Klaus M. Fiebig and Ken A. Dill. *J. Chem. Phys.*, 98(4), 1993.
[2] T. Kasami. Scientific Report AFCRL-65-758, AFCRL, Bedford MA, 1965.
[3] KF Lau and KA. Dill. *Macromolecules*, 22:638–642, 1989.
[4] K. W. Plaxco, K. T. Simons, and D. Baker. *J. Mol. Biol.*, 277:985–94, 1998.
[5] D. H. Younger. *Information and Control*, 10(2):189–208, 1967.

[1] All 24,900 unique folding 20mers; S:#conformations searched, $MaxConf = 41,889,578$