Stable foldings of proteins in the H–P model

David Bremner
University of New Brunswick

As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.

– Albert Einstein

1 Protein Folding

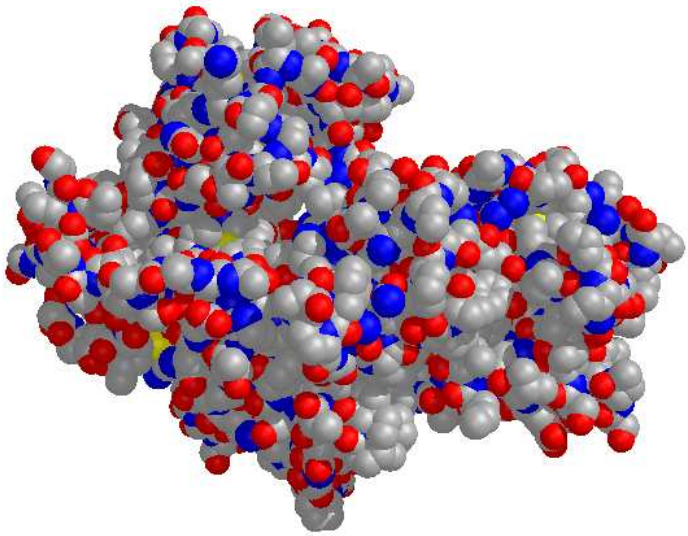
- Life=Water+Proteins.
- Strings (polymers) of amino acids
- The main breakthroughs:
  - Pauling 1951. Common local structure.
  - Anfinsen ca. 1960: Structure from sequence.

What makes this a *computational* problem?

sequence $\rightarrow$ structure $\rightarrow$ function

A A S X D X S L V
E V H X X V F I V
P P X I L Q A V V
S I A . . .

$\Rightarrow$



Physical factors in protein folding

Folding factors

- Local (neighbours) vs. nonlocal (collapse)
- Hydrophobic/hydrophilic
- Hydrogen bonding; local helix formation vs. nonlocal stabilization
- Electostatic forces

Experimental Results

- Hydrophobic/Hydrophilic forces by far strongest
- Close packed proteins are crystal- like
- β-sheets have few local interactions
- Helical structures can be designed by using only hydrophobicity,

Predicting protein structure

No model	Combinatorics	Physics and Chemistry	Quantum Physics
Annotation	Lattice models	Simulations	madness?

Problem Global optimization ($O(3^n)$ local minima?).

Solution? Forget almost everything we know.

H–P model

- Hydrophobic (H) repels water
- Polar (Hydrophilic) (P) attracts water
- Model: H's attract each other and P's are neutral

Amino Acid	Code	Classification
Leucine	L	H
Serine	S	P
Glycine	G	H
Threonine	T	P
⋮	⋮	⋮

Lattice Embeddings

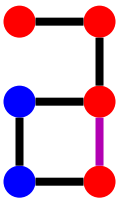
Lattices

2D square More "realistic" for ≤ 30 monomers

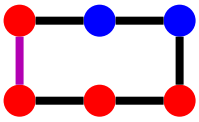
3D cubic Basic local structures (α-helix) are 3D.

Other Solve parity problems

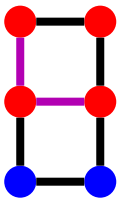
H–H bonds



1 bond



1 bond



2 bonds

Optimal Maximum number of bonds

Stable Unique optimal embedding

The Protein Folding "Paradoxes"

Protein Folding Paradox (Levinthal 1968)

There are an exponential number of foldings ("conformations"), but proteins fold quickly.

New Improved Protein Folding Paradox (1998)

Finding the optimal folding in the H–P model is NP-complete, but proteins still fold quickly.

- NP-Complete for 3D (Berger & Leighton 1998)
- NP-Complete for 2D (Crescenzi et al., JCB 1998)
- $3/8$-approximation for 3D and $1/4$-approximation for 2D (Hart and Istrail, STOC 1995).

Why care about uniqueness?

- Possible resolution to NP-hardness "paradox".
- "Sequence design: the hard part is uniqueness" (Dill et al., 1995)

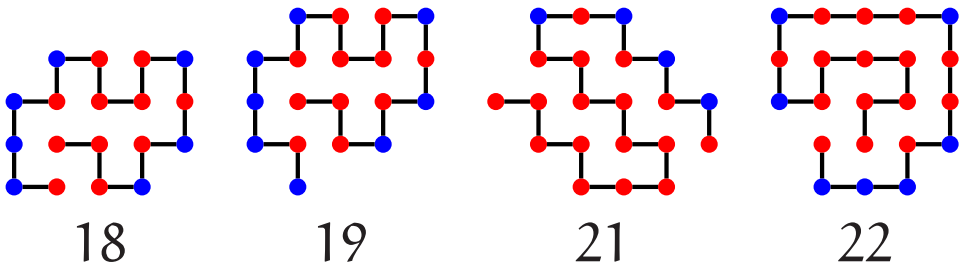
  Experimental designed polymers have many optimal foldings

  Algorithmic designing to fold to a shape is easy. (Kleinberg 1999)

  Simulation machine designed H–P-polymers tend to collapse below design state (Yue et al. 1995)

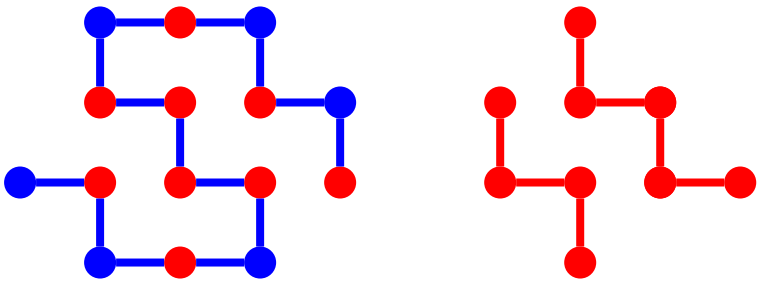
Simulation Results

- About 2% of sequences of a given length have unique optimal foldings up to length 18



2 Lattice Embeddings of Bicoloured Chains

- A *lattice embedding* of a graph maps edges to adjacent (i.e. distance 1) pairs of lattice points.
- A graph with a lattice embedding is called a *lattice graph*

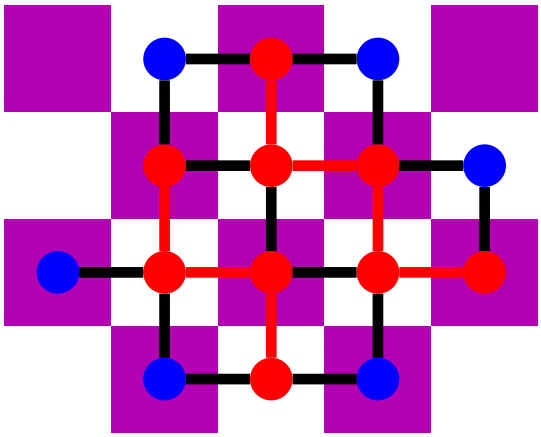


- A pair of H nodes adjacent in an embedding, but not in the graph, is called a *bond*
- *bond graph* V = H nodes; E = bonds
- The *conformation graph* consists of the edges of P, along with the bonds.

Lattice graphs are bipartite

2.1. Fact.

Every lattice graph is bipartite.



2.2. Corollary.

If an embedding of a closed chain with r H nodes has r bonds, then its bond graph consists of disjoint even cycles.

2.3. Corollary.

There can be a bond between two H nodes only if they have different parity.

3 Degenerate Ground States

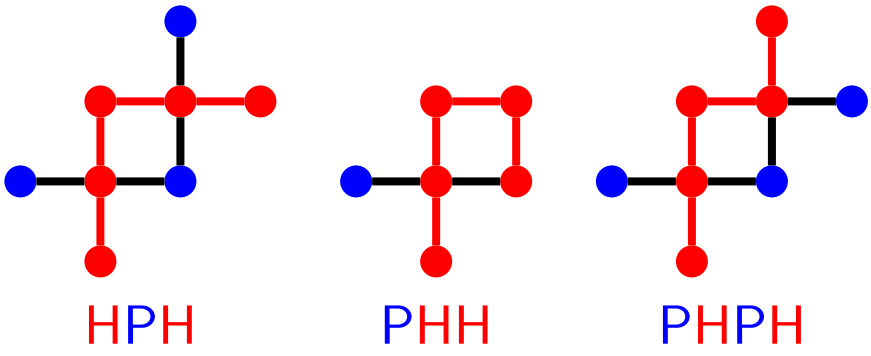
3.1. Observation.

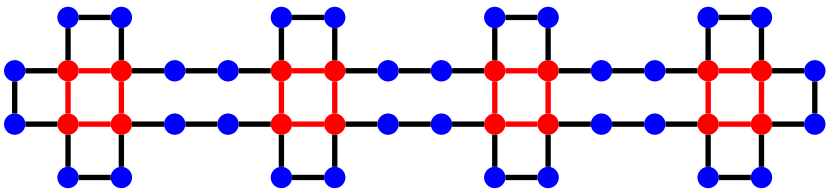
Any folding of P^k is optimal

3.2. Fact.

Any optimal folding the closed chain $(PHP)^{4k}$ has a bond graph consisting of k four cycles.

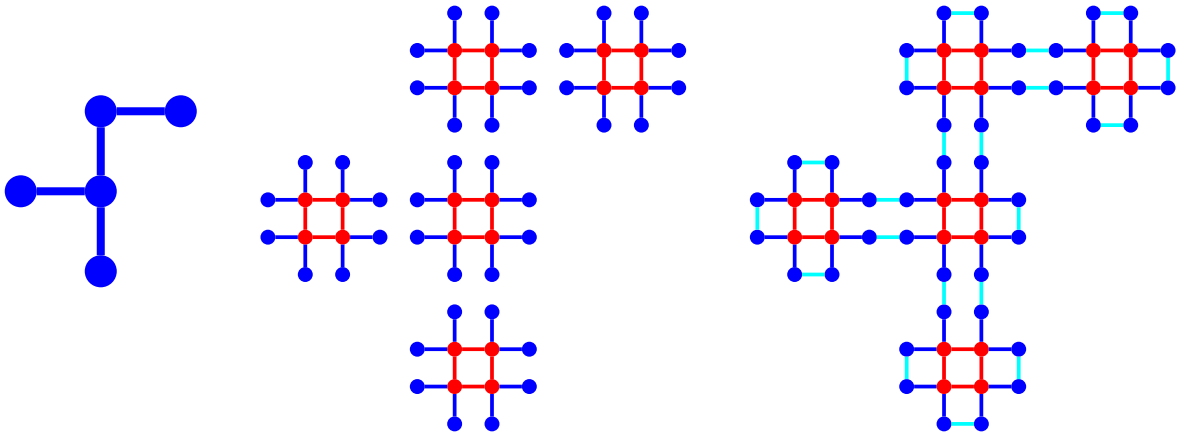
Proof. Consider a big bond graph cycle. . .





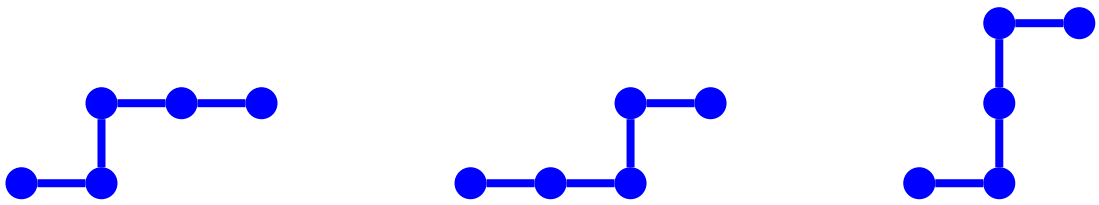
3.3. Fact.

There are as many optimal embeddings of $(\mathsf{PHP})^{4k}$ as there are (embeddings of) k node lattice trees.



3.4. Fact.

There are $\Omega(2^k)$ embeddings of k-node lattice trees.

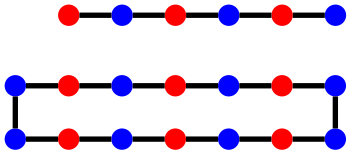


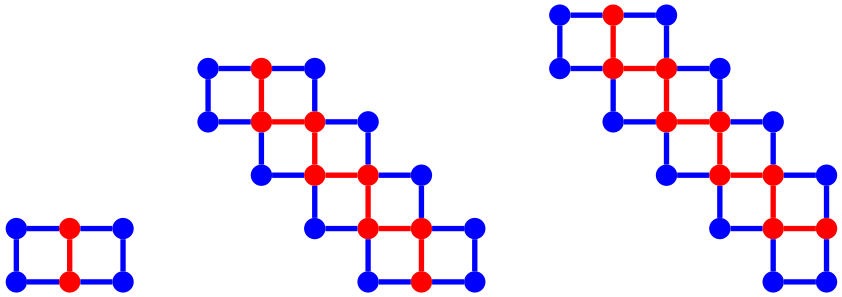
- and probably lots more $(\Omega\left(\frac{3.61^k}{k}\right))$

4 Closed Chains with Stable Foldings

$$A_m = (HP)^m$$

$$S_k = P\ A_{\lceil k/2 \rceil}\ P\ A_{\lfloor k/2 \rfloor}$$





$k = 2$ $k = 8$ $k = 9$

$$D_m = (ES)^m$$

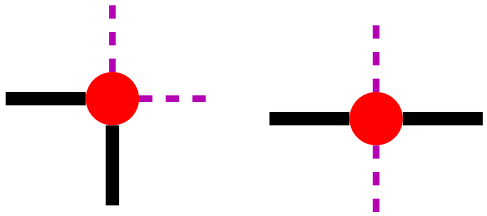
$$U_m = (WN)^m$$

$$F_k = \begin{cases} E\ D_{k/2}\ W\ U_{k/2} & k \equiv 0 \mod 2 \\ E\ D_{\lfloor k/2 \rfloor}\ S\ U_{\lceil k/2 \rceil} & k \equiv 1 \mod 2 \end{cases}$$

Missing Bonds

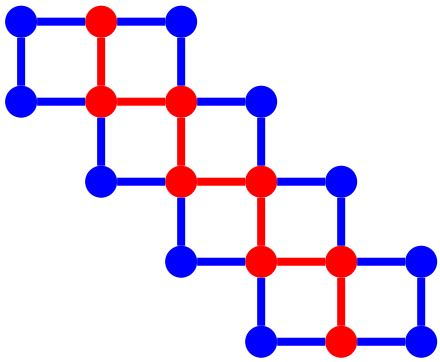
Missing Bond

- Neighbouring lattice point empty or P not adjacent on chain.



4.1. Observation.

There exists an embedding of S_k with 2 missing bonds.



4.2. Corollary.

In any optimal embedding of S_k, both monochrome edges are on the bounding box.

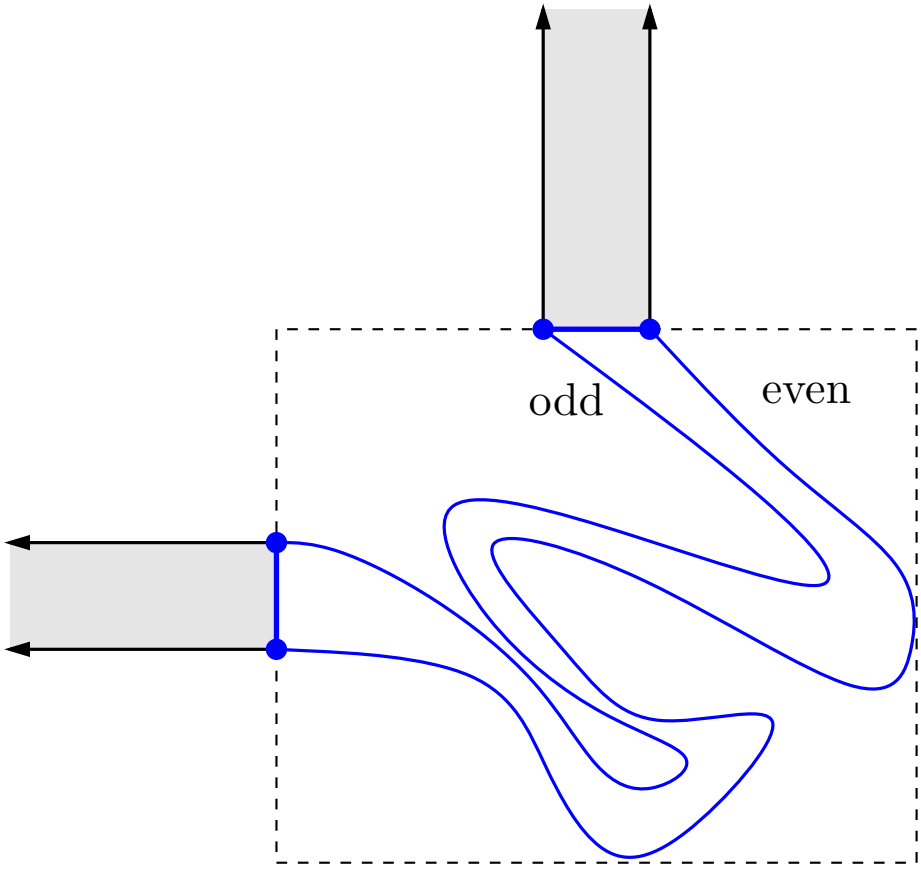
Internal and External Bonds

4.3. Definition.

An *exterior* bond in an embedding of a closed chain C is one that does not subdivide the interior of C.

4.4. Lemma.

There are no exterior bonds in an optimal embedding of S_k.

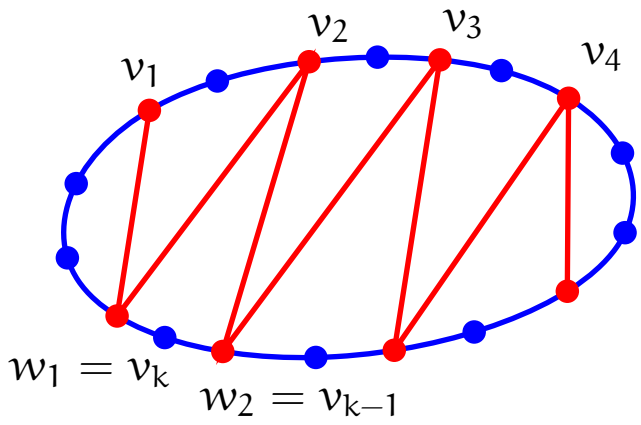


The bond graph of S_k is a path

Let v_i denote the ith H node to occur in S_k. Let w_i denote v_{k-i+1}.

4.5. Lemma.

Over all optimal embeddings of S_k, the conformation graph is unique up to relabelling $v_i \leftrightarrow w_i$.



4.6. Observation.

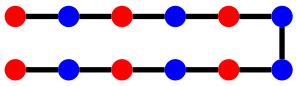
A 4-cycle has a unique lattice embedding, up to isometries.

4.7. Theorem.

F_k is the unique optimal folding (up to isometries) of S_k.

5 Open Chains with Stable Foldings

$$Z_k = (HP)^{\lceil k/2 \rceil}(PH)^{\lfloor k/2 \rfloor}$$

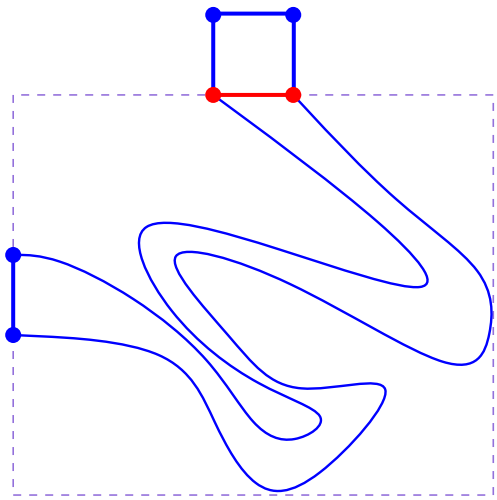


5.1. Theorem.

Z_{2j} has a unique optimal folding for all $j \geq 1$.

Proof. (Sketch)

1. How can H nodes appear on the bounding box?
2. Both endpoints on the bounding box, and bonded.
3. The monochrome edge is one the bounding box.
4. The open case reduces to the closed case



External and Internal Missing Bonds

external missing bond outside bounding box

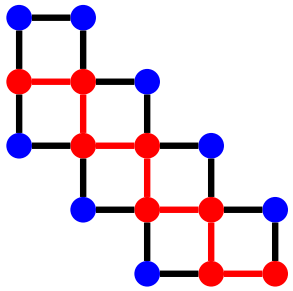
internal missing bond inside bounding box

5.2. Observation.

Every embedding of Z_k has either

(a) 3 e.m.b.'s and the monochrome edge on the bounding box, or

(b) 4 external missing bonds.



5.3. Observation.

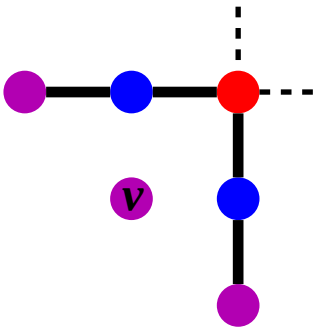
There exist embeddings of Z_{2j} with 4 external missing bonds and no internal missing bonds.

H corners

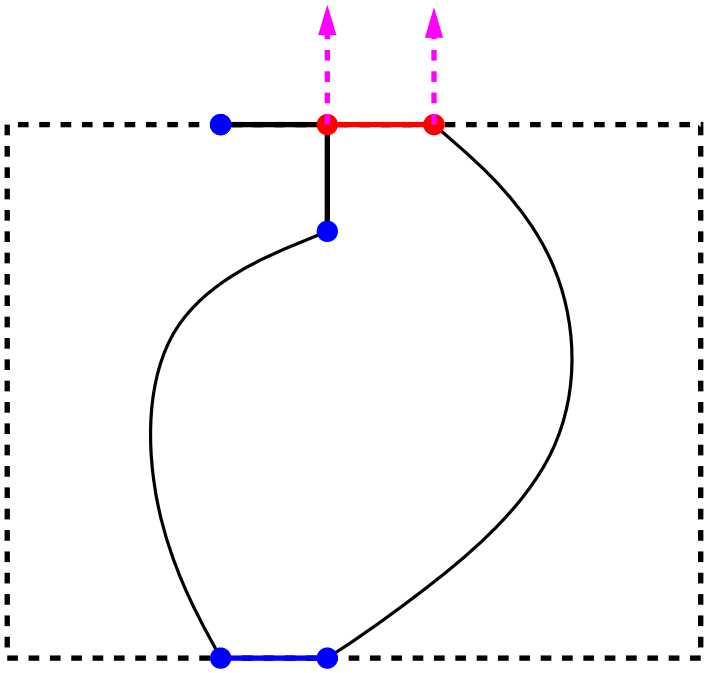
5.4. Lemma.

In an optimal embedding of Z_{2j}, there are no H corners on the bounding box.

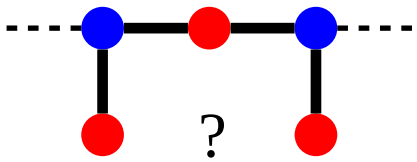
Case 1 of 5



Case 4 of 5



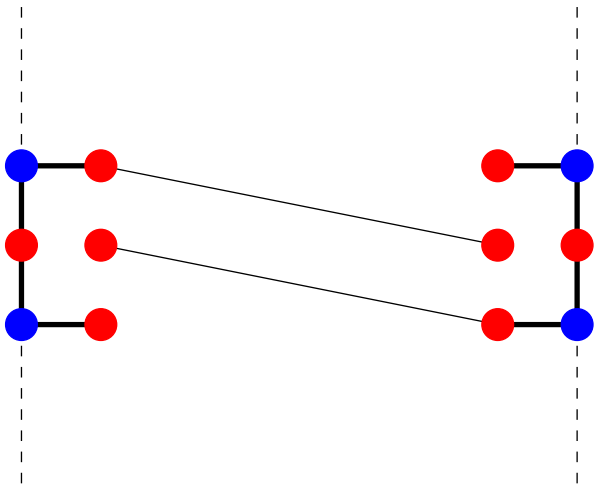
straight H nodes



5.5. Lemma.

There is at most one solitary straight H node on the bounding box.

Case 1 of 3



Wrapping things up

5.6. Lemma.

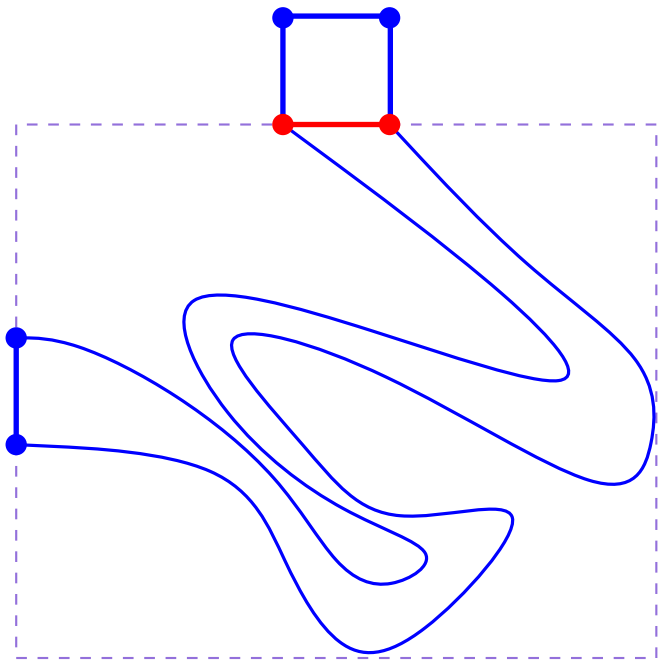
Both endpoints are on the bounding box, and bonded.

5.7. Corollary.

The monochrome edge is on the bounding box.

5.8. Fact.

There are at least as many distinct optimal embeddings of S_{2j} as there are of of Z_{2j}.



Open Questions

1. Do real proteins fold uniquely in the H–P model?
2. Asymptotically, what fraction of n-node H–P-sequences fold uniquely?
3. Is H–P sequence folding still NP-complete when restricted to "nice" sequences?

Not so open questions

- There exist stable H–P trees in 3D.
- There are stable chains in the 2D H-anything model.
- Minimal area and maximum bonds are not always simultaneously achievable.

Credits

- Inspired by an article of Brian Hayes in American Scientist
- Initiated at a workshop on Molecular Reconfiguration organized by Godfried Toussaint.
- Work described here is with Oswin Aichholzer, Erik Demaine, Vera Sacristan and Mike Soss.
- Work in progress with (additionally) Greg Aloupis, Stefan Langerman, Henk Meijer, Pat Morin, and Suneeta Ramaswami.