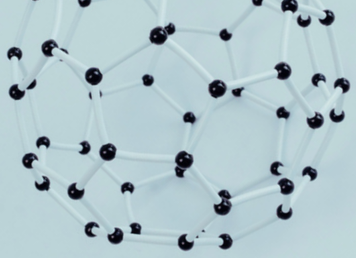


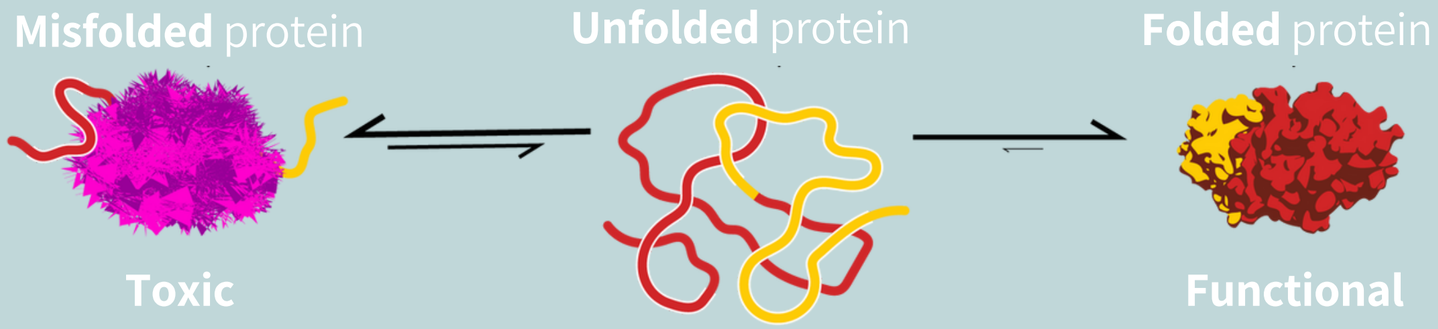
Protein Folding

Uniform Random Sampling and Optimal Folding Point Identification using the HP-Model and Randomly Generated Amino Acids

Protein Folding 101

Why should we fold proteins?

Folding proteins refers to the process of proteins taking on their functional structure from their linear sequence of amino acids. This process is essential because the function of a protein is directly linked to its structure.



Why should we fold proteins?

Folding proteins refers to the process of proteins taking on their functional structure from their linear sequence of amino acids. This process is essential because the function of a protein is directly linked to its structure.

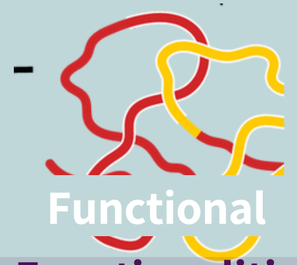
Misfolded protein

Toxic

Diseases

Alzheimer’s	Parkinson’s	Diabetes

Folded protein



Functional

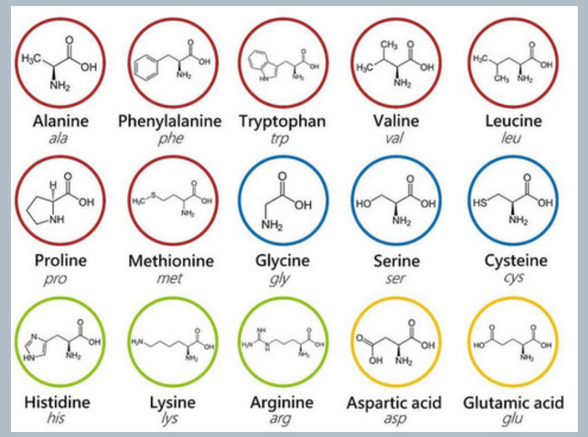
Functionalities

Muscle contraction	Woud healing	Immune system

Protein and Aminos

A protein consists of amino acids. The average protein consists of 200-300 amino acids, with insulin being a relative small protein.

20 different amino acids



Insulin - 51 amino acids

Chain A - 21 amino acids

GIVEQCCTSICSLYQLENYCN

Chain B - 30 amino acids

FVNQHLCGSHLVEALYLVCGERGFFYTPKT

Average Protein = 2.000 amino acids

CGSHLVEALYLVCGERGFFYTPKTFVNQHLCGSHLVEALYLVCGERGFFYTPKTCGSHLVEALYLVCGERGFFYTPKTCGSHLVEALYLVCGERGFFYTPKTFVNQHL

Protein Structure

Proteins can be described in terms of four levels of structure: primary, secondary, tertiary, and quaternary. These levels of structure are referred to as 1D, 2D, 3D, and 4D, respectively.

Primary Structure (1D)

Description: Primary structure refers to the linear sequence of amino acids that make up the protein.

Importance: Sequence determines the protein’s type and function. Even a small change in the sequence can significantly affect the protein’s properties.

Secondary Structure (2D)

Description: Folding amino acid chain into regular, repeating structures stabilized by hydrogen bonds.

Importance: Structures provide a level of organization and stability within the protein.

Tertiary Structure (3D)

Description: Folding and arrangement of the secondary structural elements, stabilized by various types of bonds and interactions.

Importance: Vital for the protein's function, as it positions the protein’s in specific orientations that are necessary for activity.

Quaternary Structure (4D)

Description: Arrangement of multiple subunits into a functional protein complex. Not all proteins have a quaternary structure.

Importance: Mediates the interaction of proteins with other molecules, including other proteins

Protein Structure

Proteins can be described in terms of four levels of structure: primary, secondary, tertiary, and quaternary. These levels of structure are referred to as 1D, 2D, 3D, and 4D, respectively.

Primary Structure (1D)

Description: Primary structure refers to the linear sequence of amino acids that make up the protein.

Importance: Sequence determines the protein's type and function. Even a small change in the sequence can significantly affect the protein's properties.

Secondary Structure (2D)

Description: Folding amino acid chain into regular, repeating structures stabilized by hydrogen bonds.

Importance: Structures provide a level of organization and stability within the protein.

Tertiary Structure (3D)

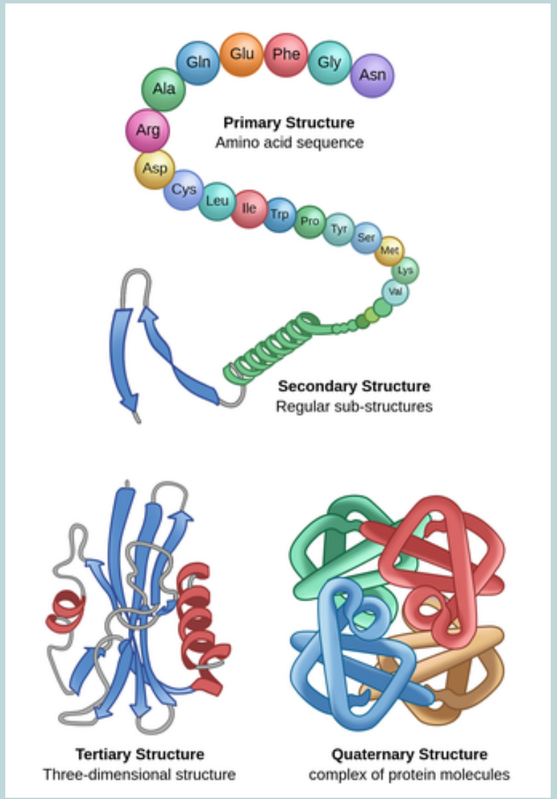
Description: Folding and arrangement of the secondary structural elements, stabilized by various types of bonds and interactions.

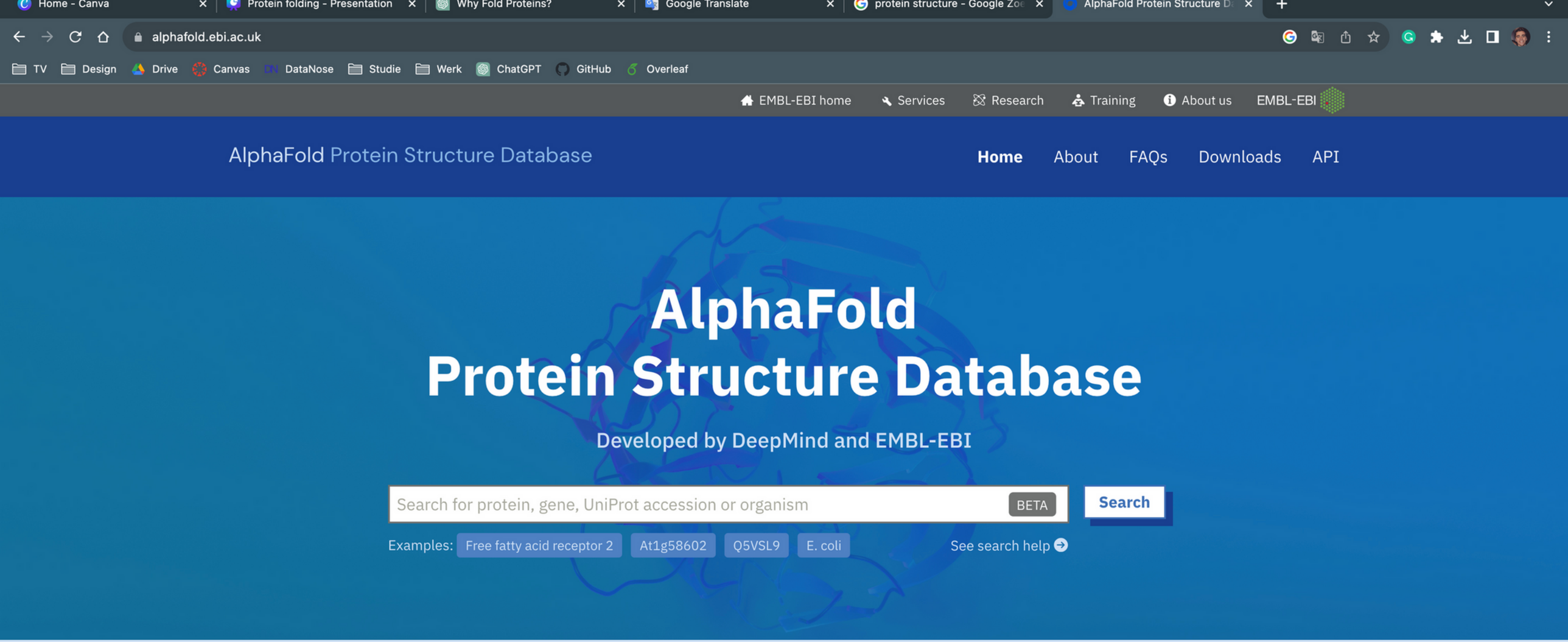
Importance: Vital for the protein's function, as it positions the protein's in specific orientations that are necessary for activity.

Quaternary Structure (4D)

Description: Arrangement of multiple subunits into a functional protein complex. Not all proteins have a quaternary structure.

Importance: Mediates the interaction of proteins with other molecules, including other proteins





AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

AlphaFold

AlphaFold is used to predict the tertiary structure (3D) of proteins. It takes the primary structure (1D) as input and predicts how those amino acids will fold into the final 3D structure.

Primary Structure (1D)

Description: Primary structure refers to the linear sequence of amino acids that make up the protein.

Importance: Sequence determines the protein's type and function. Even a small change in the sequence can significantly affect the protein's properties.

Input

Neural network

Secondary Structure (2D)

Description: Folding amino acid chain into regular, repeating structures stabilized by hydrogen bonds.

Importance: Structures provide a level of organization and stability within the protein.

Substructure

Neural network

Tertiary Structure (3D)

Description: Folding and arrangement of the secondary structural elements, stabilized by various types of bonds and interactions.

Importance: Vital for the protein's function, as it positions the protein's in specific orientations that are necessary for activity.

Output

Quaternary Structure (4D)

Description: Arrangement of multiple subunits into a functional protein complex. Not all proteins have a quaternary structure.

Importance: Mediates the interaction of proteins with other molecules, including other proteins

Hydrophobic-polar protein folding model

1 language

Contents [hide]

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

The **hydrophobic-polar protein folding model** is a highly simplified model for examining protein folds in space. First proposed by Ken Dill in 1985, it is the most known type of lattice protein: it stems from the observation that hydrophobic interactions between amino acid residues are the driving force for proteins folding into their native state.[1] All amino acid types are classified as either hydrophobic (H) or polar (P), and the folding of a protein sequence is defined as a self-avoiding walk in a 2D or 3D lattice. The HP model imitates the hydrophobic effect by assigning a negative (favorable) weight to interactions between adjacent, non-covalently bound H residues. Proteins that have minimum energy are assumed to be in their native state.

The HP model can be expressed in both two and three dimensions, generally with square lattices, although triangular lattices have been used as well. It has also been studied on general regular lattices.[2]

Randomized search algorithms are often used to tackle the HP folding problem. This includes stochastic, evolutionary algorithms like the Monte Carlo method, genetic algorithms, and ant colony optimization. While no method has been able to calculate the experimentally determined minimum energetic state for long protein sequences, the most advanced methods today are able to come close.[3][4] For some model variants/lattices, it is possible to compute optimal structures (with maximal number of H-H contacts) using constraint programming techniques[5][6] as e.g. implemented within the CPSP-tools webserver.[7]

Even though the HP model abstracts away many of the details of protein folding, it is still an NP-hard problem on both 2D and 3D square lattices.[8]

Recently, a Monte Carlo method, named FRESS, was developed and appears to perform well on HP models.[9]

See also [edit]

- Protein structure prediction
- Lattice proteins

References [edit]

1. ^ Dill K.A. (1985). "Theory for the folding and stability of globular proteins". *Biochemistry*. **24** (6): 1501–9. doi:10.1021/bi00327a032. PMID 3986190.
2. ^ Bechini, A. (2013). "On the characterization and software implementation of general protein lattice models". *PLOS ONE*. **8** (3): e59504. Bibcode:2013PLoSO...859504B. doi:10.1371/journal.pone.0059504. PMC 3612044. PMID 23555684.
3. ^ Bui T.N.; Sundarraj G. (2005). "An efficient genetic algorithm for predicting protein tertiary structures in the 2D HP model". *Proceedings of the 7th annual conference on Genetic and evolutionary computation*. pp. 385–392. doi:10.1145/1068009.1068072. ISBN 978-1595930101. S2CID 13485429.
4. ^ Shmygelska A.; Hoos H.H. (2003). "An Improved Ant Colony Optimisation Algorithm for the 2D HP Protein Folding Problem". *Advances in Artificial Intelligence*. Lecture Notes in Computer Science. Vol. 2671. pp. 400–417. CiteSeerX 10.1.1.13.7617. doi:10.1007/3-540-44886-1_30. ISBN 978-3-540-40300-5.
5. ^ Yue K.; Fiebig K.M.; Thomas P.D.; Chan H.S.; Shakhnovich E.I.; Dill K.A. (1995). "A test of lattice protein folding algorithms". *Proc Natl Acad Sci U S A*. **92** (1): 325–329. Bibcode:1995PNAS...92..325Y. doi:10.1073/pnas.92.1.325. PMC 42871. PMID 7816842.
6. ^ Mann M.; Backofen R. (2014). "Exact methods for lattice protein models". *Bio-Algorithms and Med-Systems*. **10** (4): 213–225. doi:10.1515/bams-2014-0014.

HP-model

The Hydrophobic-polar (HP) model developed by Ken Dill is often used in optimization to predict the most stable conformation of a protein.

Primary Structure (1D)

Description: Primary structure refers to the linear sequence of amino acids that make up the protein.

Importance: Sequence determines the protein’s type and function. Even a small change in the sequence can significantly affect the protein’s properties.

Input

Secondary Structure (2D)

Description: Folding amino acid chain into regular, repeating structures stabilized by hydrogen bonds.

Importance: Structures provide a level of organization and stability within the protein.

Output

Optimization Algorithms

Tertiary Structure (3D)

Description: Folding and arrangement of the secondary structural elements, stabilized by various types of bonds and interactions.

Importance: Vital for the protein's function, as it positions the protein’s active sites and other functional domains in specific orientations that are necessary for activity.

Quaternary Structure (4D)

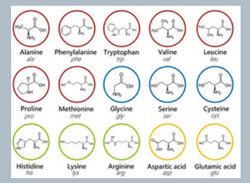
Description: Arrangement of multiple subunits into a functional protein complex. Not all proteins have a quaternary structure.

Importance: Mediates the interaction of proteins with other molecules, including other proteins

HP-model

The Hydrophobic-polar (HP) model developed by Ken Dill is often used in optimization to predict the most stable conformation of a protein.

20 to 2 acids



H P

Insulin - 51 amino acids

Chain A - 21 amino acids
GIVEQCCTSICSLYQLENYCN

Chain B - 30 amino acids
FVNQHLCGSHLVEALYLVCGERGFFYTPKT

Insulin - 51 amino acids

Chain A - 21 amino acids
HPPPHHHPPPHHHPPPHP

Chain B - 30 amino acids
HPHHPPPHPPPPHHPPPHHHPPPHHHPP

Simplification

HP-model

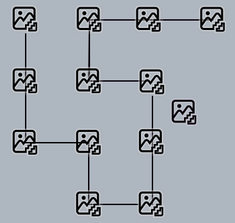
The Hydrophobic-polar (HP) model developed by Ken Dill is often used in optimization to predict the most stable conformation of a protein.

1D representation

Chain A - 21 amino acids
HPPPHHHPPPHHHPPPHP

Chain B - 30 amino acids
HPHHPPPHPPPPHHPPPHHHPPPHHHPP

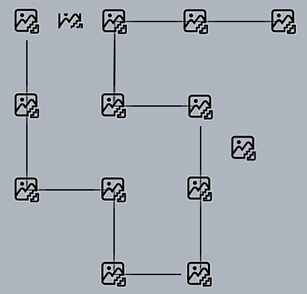
2D representation



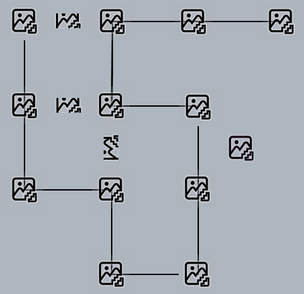
HP-model

In the context of the HP model, the energy is typically calculated based on the number of adjacent hydrophobic acids in the folded state of the protein. These are called H-bonds.

Energy calculation -1



Energy calculation -3



Scope

Uniform Random Samling

A uniform random sample refers to selecting a random folded protein conformation from a larger population of folded protein conformations in a way that configuration has an equal probability of being chosen.

Uniform Random Samling

Why do we want a uniform random sample?

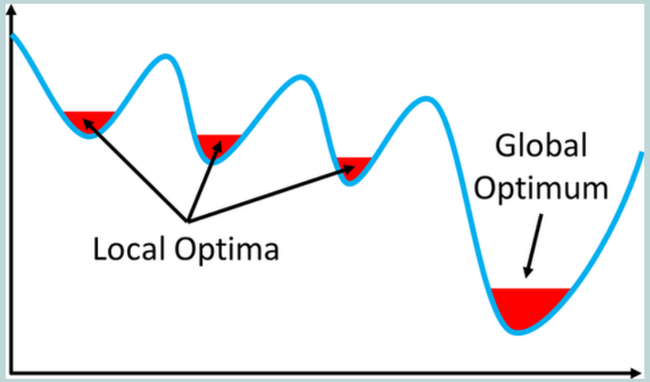
Limited Diversity

- **Issue:** Non-uniform random sampling leads to a population that is not diverse. A lack of diversity means fewer unique solutions are being explored.
- **Result:** The algorithm may quickly converge to a local optimum because it doesn't have the diversity needed to explore other potentially superior solutions.

Premature Convergence

- **Issue**: When the initial population isn't uniformly and randomly sampled, there's a risk of premature convergence, where the algorithm settles on a suboptimal solution too quickly.

- **Result**: It doesn't explore enough of the solution space to find the global optimum.



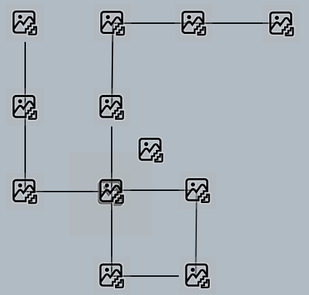
Experiment

Experiment

Three different methods for (random) protein sampling on a 2 dimension grid. We generate 1000 samples for protein for lengths n ∈ {5, 10, 15…195, 200} in numpy.

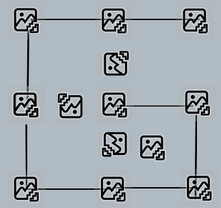
Sampling with collisions*

Counts the number of **collisions**



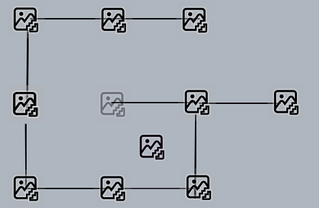
Sampling with break

Break when no adjacent locations



Sampling with backtracking

Backtrack when no adjacent locations



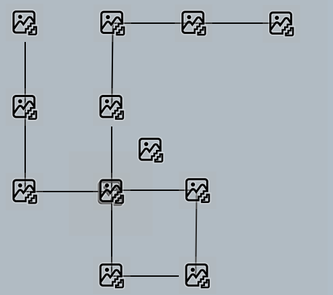
*Work by Jansen et al. (2023)

Experiment

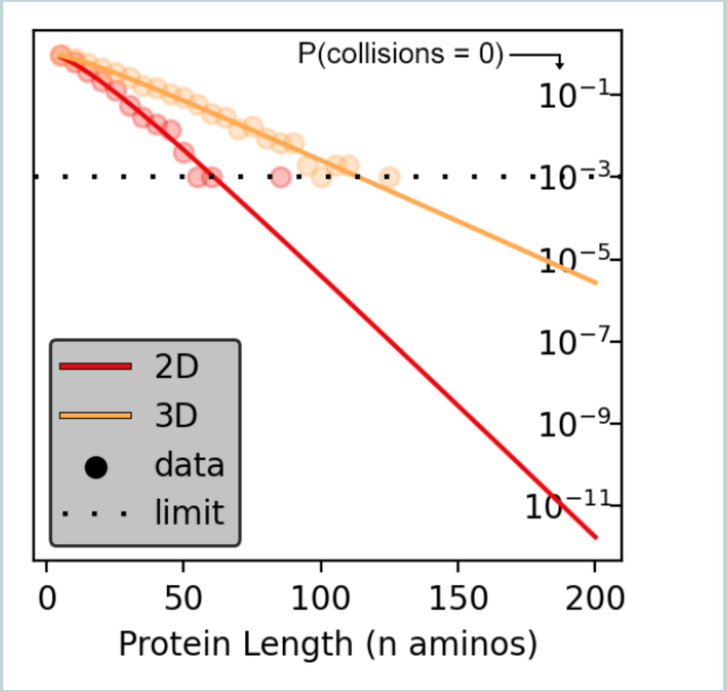
Three different methods for (random) protein sampling on a 2 dimension grid. We generate 1000 samples for protein for lengths n ∈ {5, 10, 15...195, 200} in numpy.

Sampling with collisions*

Counts the number of **collisions**



number of collisions increase as n increases



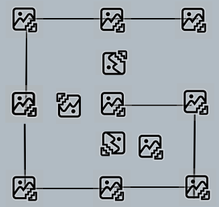
The chance of randomly sampling a zero-collision conformation drops exponentially in n

Experiment

We generate 1000 samples for protein for lengths n ∈ {5, 10, 15...195, 100}. Programmed to scale up experiment for both amino acid length and sample sizes.

Sampling with break

Break when no adjacent locations



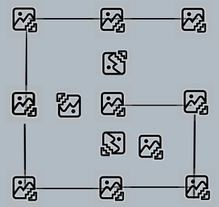
	Amino Acid Length	Num Hydrophobic	Num Polar	1D protein	2D protein	Amino Acids on Grid	Trimmed 2D protein	Shape 2D protein	Amino Acid Order	H-Bonds	H-Ratio
19995	100	85	15	['H', 'H', 'H', 'P', 'H', 'H', 'H', 'H', 'H', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' ''...	21	[['H' 'H' 'H' 'P' 'H' 'H' 'H' 'H' 'H' 'P' ...	(3, 13)	[('H', (100, 100)), ('H', (100, 101)), ('H', (...	3	0.142857
19996	100	53	47	['P', 'P', 'H', 'P', 'P', 'H', 'P', 'H', 'H', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' ''...	100	[['' '' 'P' 'H' 'P' 'H' 'H' 'P' '' '' '' ''...	(22, 21)	[('P', (100, 100)), ('P', (100, 99)), ('H', (9...	10	0.100000
19997	100	15	85	['P', 'P', 'P', 'P', 'P', 'P', 'P', 'P', 'P', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' ''...	92	[['' '' '' '' '' '' '' '' '' '' '' '' '' ''...	(22, 24)	[('P', (100, 100)), ('P', (101, 100)), ('P', (...	1	0.010870
19998	100	75	25	['H', 'H', 'H', 'P', 'H', 'H', 'H', 'H', 'H', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' ''...	24	[['' '' 'H' 'H' 'H']\n ['H' 'H' 'H' 'H' 'H']\n...	(6, 5)	[('H', (100, 100)), ('H', (100, 99)), ('H', (9...	8	0.333333
19999	100	21	79	['P', 'H', 'P', 'P', 'P', 'P', 'P', 'H', 'P', ...	[['' '' '' '' '' '' '' '' '' '' '' '' '' ''...	36	[['' '' '' '' '' 'H' 'P' 'P' 'P']\n ['' '' '' ...	(14, 9)	[('P', (100, 100)), ('H', (100, 101)), ('P', (...	0	0.000000

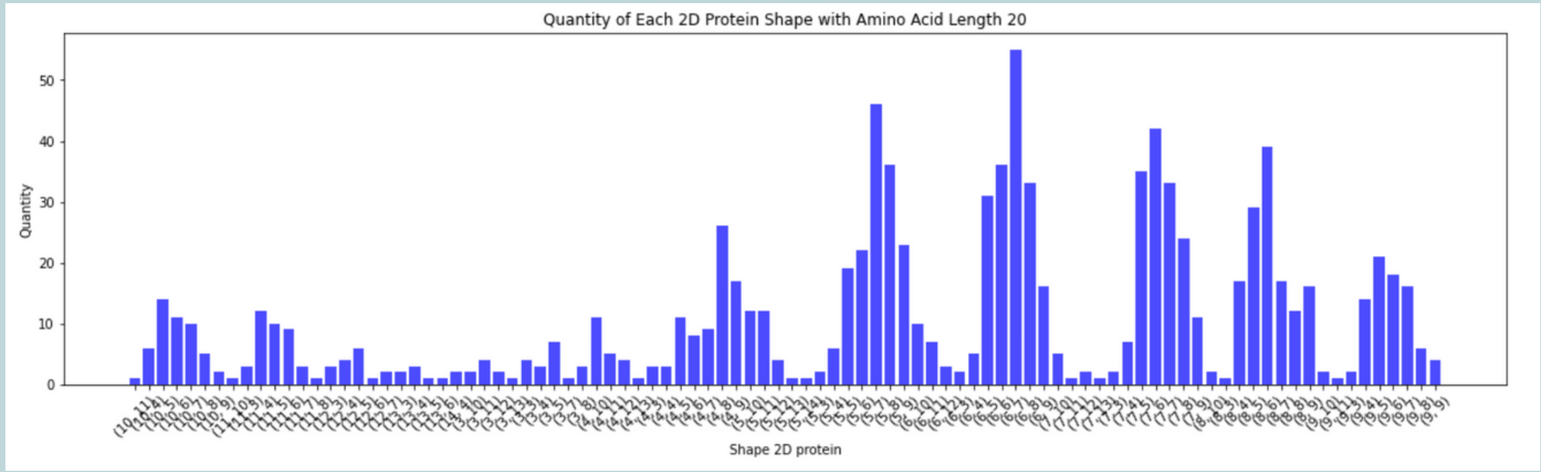
Experiment

We generate 1000 samples for protein for lengths n ∈ {5, 10, 15...195, 100}. Programmed to scale up experiment for both amino acid length and sample sizes.

Sampling with break

Break when no adjacent locations





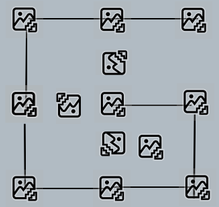
The shape of a protein seems the be normally distributed around squared conformations e.g. 7x7, 6x6.

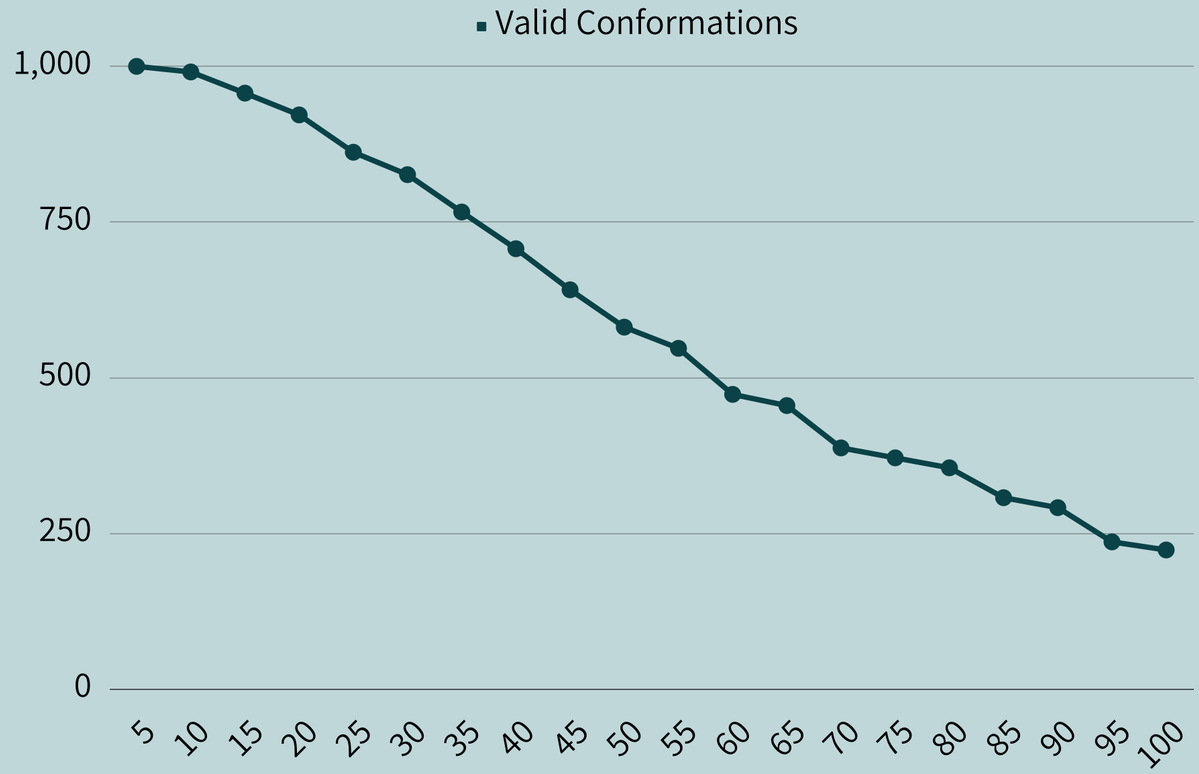
Experiment

We generate 1000 samples for protein for lengths n ∈ {5, 10, 15...195, 100}. Programmed to scale up experiment for both amino acid length and sample sizes.

Sampling with break

Break when no adjacent locations





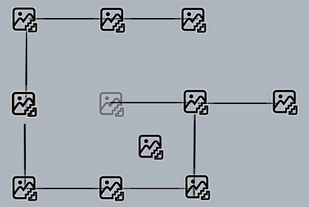
number of collisions increase as n increases

Experiment

We generate 1000 samples for protein for lengths n ∈ {5, 10, 15…195, 100}. Programmed to scale up experiment for both amino acid length and sample sizes.

Sampling with backtracking

Backtrack when no adjacent locations



We generate 1000 samples for protein for lengths n ∈ {5, 10, 15…195, 100}. Programmed to scale up experiment for both amino acid length and sample sizes.

Optimal Folding Point Identification

Given a uniform random sample, does the point at which we start folding have a significant influence on the time in which we can reach a global optimum with minimal energy

Thanks!