Protein design

CS/CME/Biophys/BMI 279

Oct. 20 and 22, 2015

Ron Dror

Optional reading on course website

From cs279.stanford.edu

Lectures

- 9/22 - Introduction [slides] [notes]
- 9/24 and 9/29 - Protein Structure [slides]
  Optional Reading:
  - Michael Levitt's Lecture 1 from SB228 [narrated slides]
  - Michael Levitt's Lecture 2 from SB228 [narrated slides]
- 9/29 and 10/1 - Energy Functions and their relationship to protein conformation [slides]
  Optional Reading:
  - Michael Levitt's Lecture 2 from SB228 [narrated slides]
  - Michael Levitt's Lecture 4 from SB228 [narrated slides]
- 10/6 and 10/8 - Molecular Dynamics Simulation [slides]
  Optional Reading:
  - Biomolecular Simulation: A Computational Microscope for Molecular Biology
- 10/13 and 10/15 - Protein Structure Prediction [slides]
  Optional Reading:
  - The Protein-Folding Problem, 50 Years On
  - The Phyre2 web portal for protein modeling, prediction and analysis
- 10/20 and 10/22 - Protein Design [slides]
  Optional Reading:
  - Computer-Based Design of Novel Protein Structures
  - Protein Design Wiki

- These reading materials are optional.
- They are intended to (1) help answer questions you have about course material (but feel free to ask course staff) and (2) provide information beyond what's covered in lecture. You're not responsible for all this additional information.

A caveat

MENU the ONION Wikipedia Celebrates 750 Years Of American Independence

Wikipedia Celebrates 750 Years Of American Independence

NEWS
July 26, 2006

VOL 46 ISSUE 26
Science & Technology · Old Internet · Patriotism · Internet · History

NEW YORK—Wikipedia, the online, reader-edited encyclopedia, honored the 750th anniversary of American independence on July 25 with a special featured section on its main page Tuesday.

Three girls march toward the White House on Elm St. in Washington, DC, as part of the Inderpendance Day Parade.

"It would have been a major oversight to ignore this portentous anniversary," said Wikipedia founder Jimmy Wales, whose site now boasts over 4,300,000 articles in multiple languages, over one-quarter of which are in English, including 11,000 concerning popular toys of the 1980s alone. "At 750 years, the U.S. is by far the world's oldest surviving democracy, and is certainly deserving of our recognition," Wales said. "According to our database, that's 212 years older than the Eiffel Tower, 347 years older than the earliest-known woolly-mammoth fossil, and a full 493 years older than the microwave oven."

- This is a rapidly developing field. The literature is not always self-consistent (this includes papers in scientific journals, not just Wikipedia!)

Next quarter: CS/CME/Biophys/BMI 371 "Computational biology in four dimensions"

- I'm teaching a course next quarter that complements this one
- Similar topic area, but with a focus on current cutting-edge research
  - Focus is on reading, presentation, discussion, and critique of published papers

Questions

- I will defer some questions to the end of class
- Please do still ask questions, especially if you don’t understand something I explain!

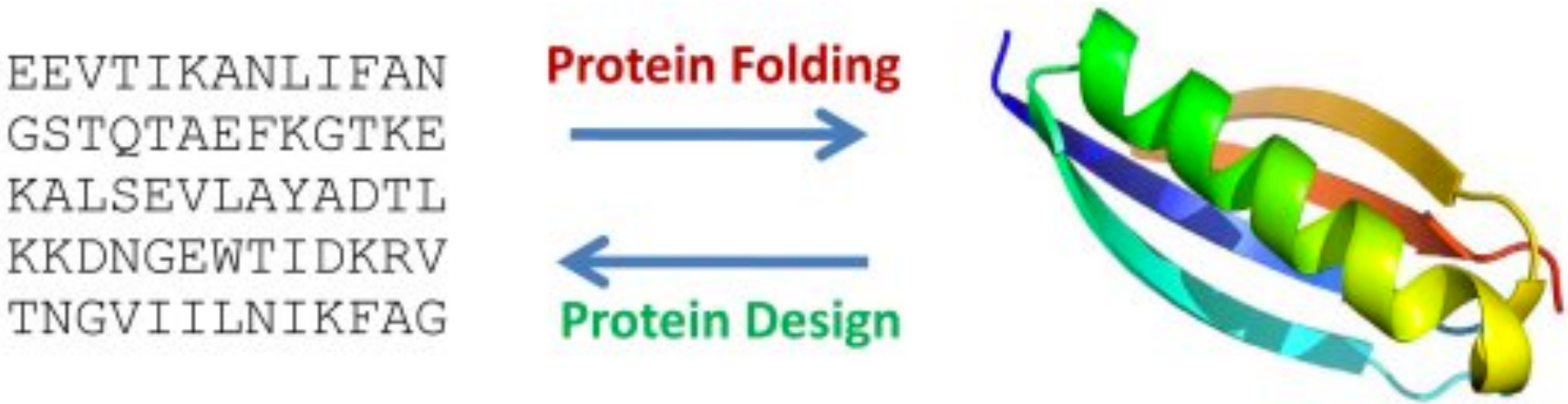
Outline

- Why design proteins?
- Overall approach: Simplifying the protein design problem
- Protein design methodology
  - Designing the backbone
  - Select sidechain rotamers: the core optimization problem
  - Optional: giving the backbone wiggle room
  - Optional: negative design
  - Optional: complementary experimental methods
- Examples of recent designs
- How well does protein design work?

Why design proteins?

Problem definition

- Given the desired three dimensional structure of a protein, design an amino acid sequence that will assume that structure.
  - Of course, a precise set of atomic coordinates would determine sequence. Usually we start with an *approximate* desired structure.
  - Alternatively, we may want to design for a particular function (e.g., the ability to bind a particular ligand).



http://www.riken.jp/zhangiru/images/sequence_protein.jpg

Sample applications

- Designing enzymes (proteins that catalyze chemical reactions)
  - Useful for production of industrial chemicals and drugs
- Designing proteins that bind specifically to other proteins
  - Potential for HIV, cancer, Alzheimer's treatment
  - Special case: antibody design
- Designing sensors (proteins that bind to and detect the presence of small molecules—for example, by lighting up or changing color)
  - Calcium sensors used to detect neuronal activity in imaging studies
  - Proteins that detect TNT or other explosives, for mine detection
- Making a more stable variant of an existing protein
  - Or a water-soluble variant of a membrane protein

You're not responsible for these

Overall approach: simplifying the protein design problem

“Direct” approach

- Given a target structure, search over all possible protein sequences
- For each protein sequence, predict its structure, and compare to the target structure
- Choose the best match

Direct approach has two major problems

- Computationally intractable
  - We'd need to use ab initio structure prediction
  - Ab initio structure prediction for even one sequence is computationally intensive
  - 20^N possible sequences with N residues
- May not be good enough!
  - Ab initio structure prediction is far from perfect, in part because energy functions are imperfect
  - Given an energy function, what we really want is to maximize the probability of the desired structure (compared to all other possible folded and unfolded structures)
  - We could do this by sampling the full Boltzmann distribution for each candidate sequence … but that's even more computationally intensive!

We can dramatically simplify this problem by making a few assumptions

1. Assume the backbone geometry is fixed
2. Assume each amino acid can only take on a finite number of geometries (*rotamers*)
3. Assume that what we want to do is to maximize the energy drop from the completely unfolded state to the target geometry
   - In other words, simply ignore all the other possible folded structures that we want to avoid

We'll first address the problem under these assumptions, then consider relaxing them a bit

The simplified problem

- At each position on the backbone, choose a rotamer (an amino acid type and a side-chain geometry) to minimize overall energy
  - We assume the energy is a free energy. The Rosetta all-atom force field (physics-based/knowledge-based hybrid) is a common choice.
  - Energy is measured relative to the unfolded state.
    - In practice a "reference energy" for each amino acid is subtracted off, corresponding roughly to how much that amino acid favors folded states
    - You're not responsible for this
  - Assume that energy can be expressed as a sum of terms that depend on one rotamer or two rotamers each. This is the case for the Rosetta force fields (and for most molecular mechanics force fields as well).
- Thus, we wish to minimize total energy E_T, where

$$E_T = \sum_i \left[E_i(r_i) + \sum_{i \neq j} E_{ij}(r_i, r_j) \right]$$

Note that r_i specifies both the amino acid at position i and its side-chain geometry

Protein design methodology

Protein design methodology

Designing the backbone

Designing the backbone

- The first step of most protein design protocols is to select one or more target backbone structures.
  - This is as much art as science.
  - Often multiple target structures are selected, because some won't work. (Apparently proteins can only adopt a limited set of backbone structures, but we don't have a great description of what that set is.)
- Methods to do this:
  - Use an experimentally determined backbone structure
  - Assemble secondary structural elements by hand
  - Use a fragment assembly program like Rosetta, selecting fragment combinations that fit some approximate desired structure

Example of backbone design

– To design “Top7,” a protein with a novel fold, Kuhlman et al. started with a schematic, then used Rosetta fragment assembly to find 172 backbone models that fit it.

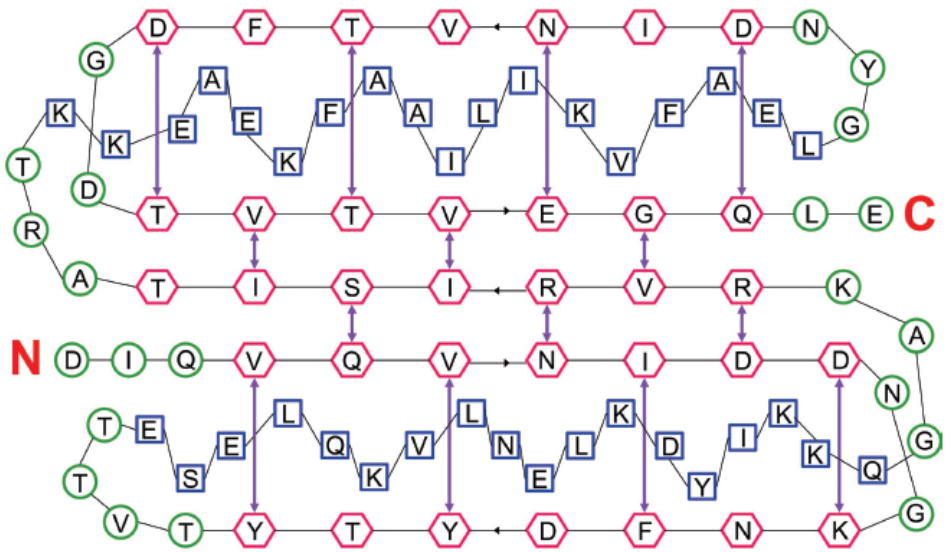
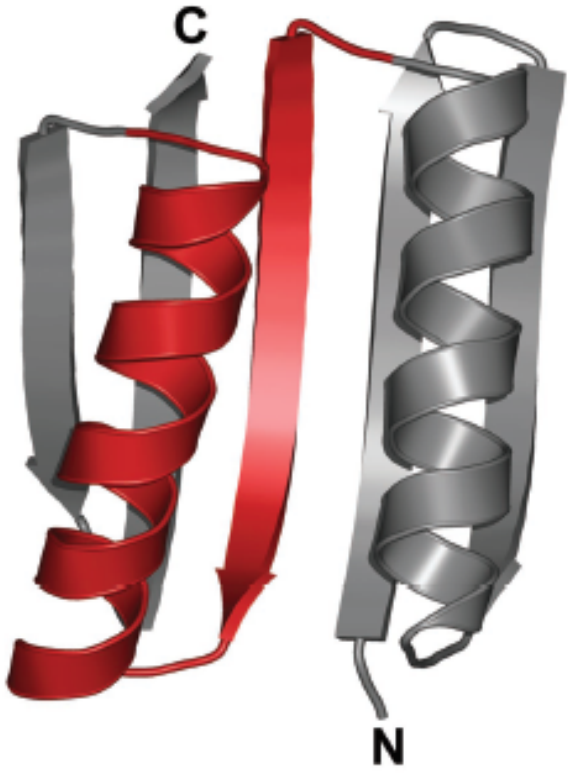


Fig. 1. A two-dimensional schematic of the target fold (hexagon, strand; square, helix; circle, other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence.

Initial schematic of target fold. Hexagons = β sheet. Squares = α helix. Arrows = hydrogen bonds. Letters indicate amino acids in final designed sequence (these were not determined until much later).



Final structure

Kuhlman et al., Science 302:1364-8 (2003)

Protein design methodology

Select sidechain rotamers: the core optimization problem

The optimization problem

- Given a desired backbone geometry, we wish to select rotamers at each position to minimize total energy

$$E_T = \sum_i \left[E_i(r_i) + \sum_{i \neq j} E_{ij}(r_i, r_j) \right]$$

where r_i specifies both the amino acid at position *i* and its side-chain geometry

Optimization methods

- Heuristic methods
  - Not guaranteed to find optimal solution, but faster
  - Most common is Metropolis Monte Carlo
    - Moves may be as simple as randomly choosing a position, then randomly choosing a new rotamer at that position
    - May decrease temperature over time (simulated annealing)
- Exact methods
  - Guaranteed to find optimal solution, but prohibitively slow for larger proteins
  - Most common is likely Dead-End Elimination Method, which prunes branches of the exhaustive search tree by proving that certain rotamers are incompatible with the global optimum
  - The A* optimization algorithm (originally developed at Stanford) is also used
  - You're not responsible for the details of how these exact methods work.

Optional: giving the backbone wiggle room

“Flexible backbone” design

- One of our key simplifying assumptions was that of a fixed backbone geometry.
- For many applications, protein design works better if you give the backbone some limited “wiggle room.”
- This requires optimizing simultaneously over rotamers and backbone geometry.
  - Often addressed through a Monte Carlo search procedure that alternates between local tweaks to backbone dihedrals and changes to side-chain rotamers

Protein design methodology

Optional: negative design

Negative design

- Another simplifying assumption was that we simply minimize the energy of the desired structure
  - We do not consider all other possible structures. It's possible that their energy ends up even lower.
- In negative design, we identify a few structures that we want to *avoid*, and we try to keep their energies high during the design process.
  - This can help, but we cannot explicitly avoid all possible incorrect structures without making the problem much more complicated. So the overall approach is still heuristic.

Optional: complementary experimental methods

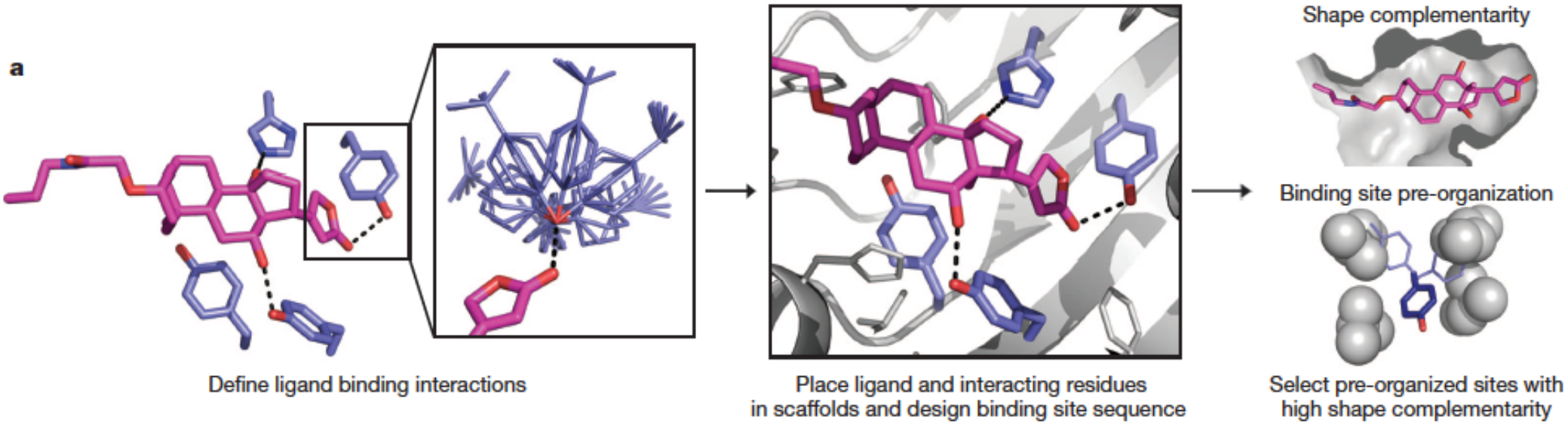
Complementary experimental methods

- Computational protein design is often combined with experimental protein engineering methods
- For example, computational designs can often be improved by directed evolution
  - Directed evolution involves introducing random mutations to proteins and picking out the best ones
  - Usually this is done in living cells, with the fittest cells (i.e., those containing the “best” version of the protein) selected by some measure

Examples of recent designs

Designing proteins that bind specific ligands

- The example below required specification of the position of certain side chains that will form favorable interactions with the ligand

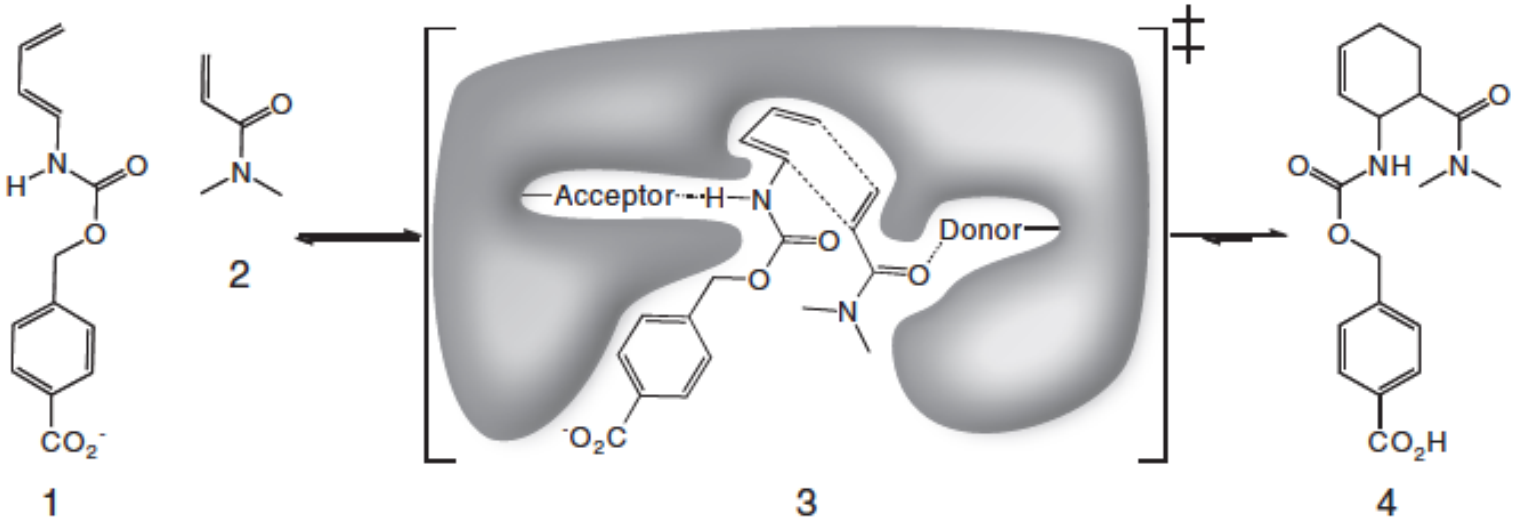


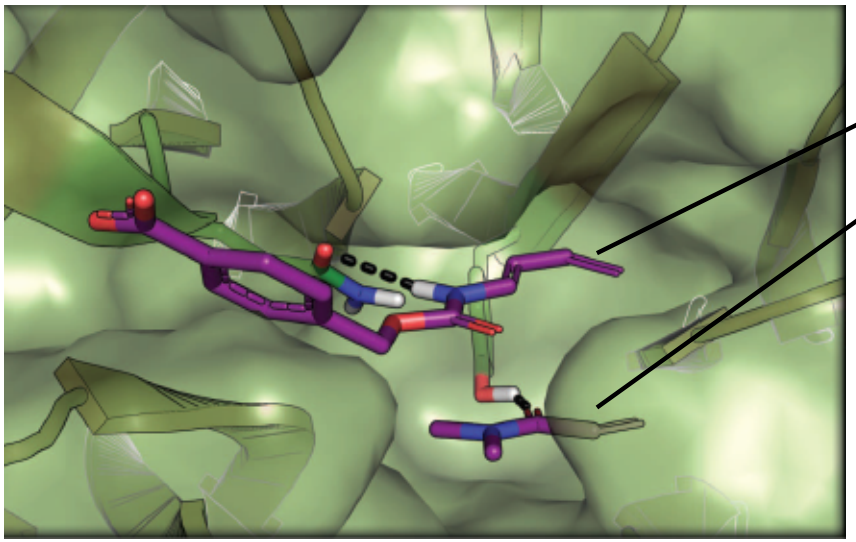
Protein designed to bind tightly to a specific steroid, but not to related molecules

Tinberg et al., Nature 501:212-6 (2013)

Designing enzymes

- In the example below, the protein holds two molecules in just the right relative positions for them to react. This speeds up the reaction.





Molecule 1

Molecule 2

Siegel et al., Science 329:309-13 (2010)

How well does protein design work?

How well does protein design work?

- Some impressive recent successes
- However, one should keep in mind that:
  - Successful protein design projects often involve making and experimentally testing dozens of candidate proteins to find a good one
  - Projects and design strategies that fail generally aren't published
  - Design of membrane proteins remains difficult