Protein Folding and Design

Protein Folding

The process by which a protein goes from unfolded polypeptide with no activity to a uniquely structured and active protein.

Understanding protein folding can allow to engineer proteins with unique properties.

In cells:

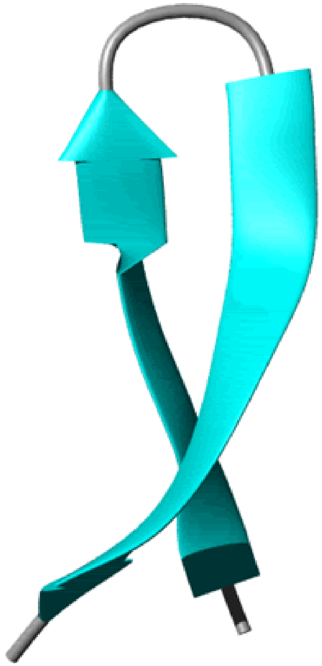
- Improper folding leads to protein degradation by a cellular machinery.
- Protein misfolding has been implicated in many human diseases (Alzheimer's, Parkinson’s, …).

Why Proteins Structure ?

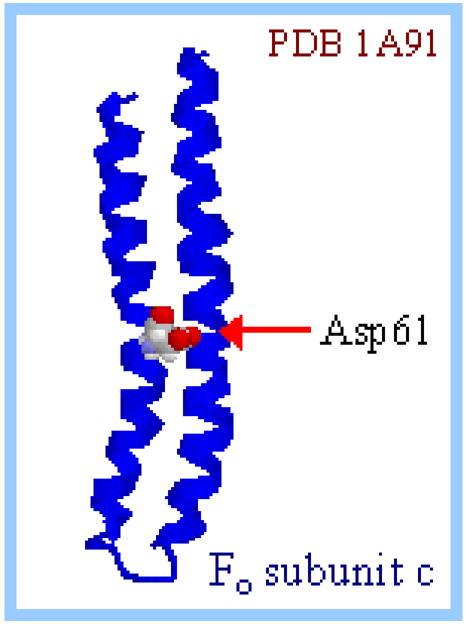
- Proteins are fundamental components of all living cells, performing a variety of biological tasks.
- Each protein has a particular 3D structure that determines its function.
- Protein structure is more conserved than protein sequence, and more closely related to function.

Secondary structures

Assembly of secondary structures, which are shared by many structures.



Beta hairpin



Helix hairpin

Beta-alpha-beta unit

Structure – Sequence Relationships

- Two conserved sequences → similar structures
- Two similar structures $\xrightarrow{?}$ conserved sequences

There are cases of proteins with the same structure but no clear sequence similarity.

Principles of Protein Structure

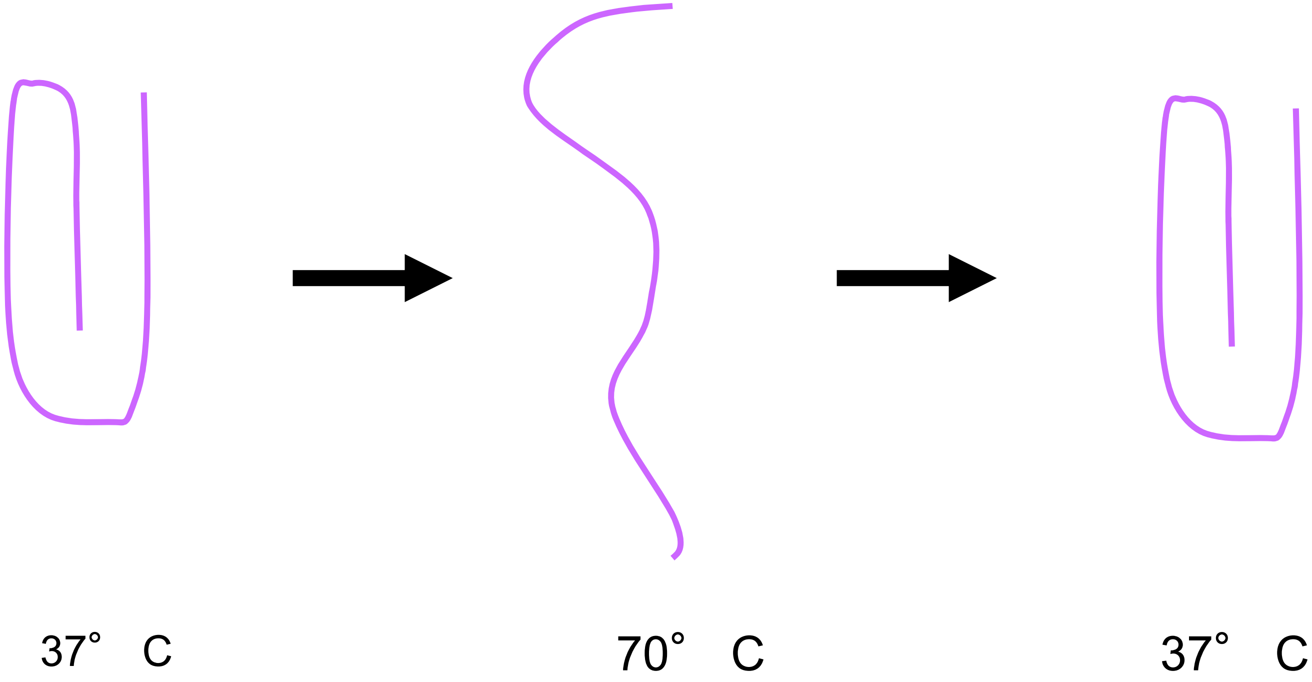
- Today's proteins reflect millions of years of evolution.

- 3D structure is better conserved than sequence during evolution.

- Similarities among sequences or among structures may reveal information about shared biological functions of a protein family.

Protein folding *in vitro* is often reversible

(indicating that the final folded structure is determined by its amino acid sequence)



Chris Anfinsen - 1957

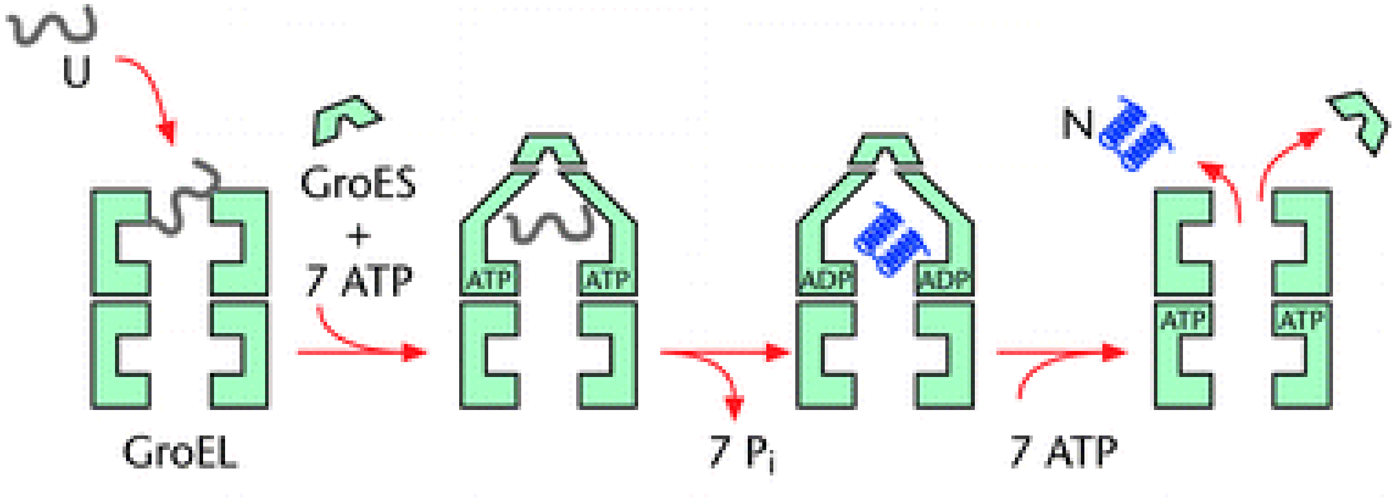
Molecular Chaperones

- Nature has developed a diverse set of proteins (chaperones) to help other proteins fold.

- Over 20 different types of chaperones have been identified. Many of these are produced in greater numbers during times of cellular stress.

Example: The GroEL(Hsp60) family

GroEL proteins provide a protected environment for other proteins to fold.



Binding of U occurs by interaction with hydrophobic residues in the core of GroEL. Subsequent binding of GroES and ATP releases the protein into an enclosed cage for folding.

How is the 3D Structure Determined ?

- X-ray crystallography and NMR are the most widely used methods.

- Quaternary structure of large proteins (ribosomes, virus particles, etc) can be determined by electron microscopes (cryoEM).

Prediction Strategies

Homology Modeling

- Proteins that share similar sequences share similar folds.
- Use known structures as the starting point for model building.
- Cannot be used to predict structure of new folds.

De Novo Structure Prediction

- Do not rely on global similarity with proteins of known structure.
- Folds the protein from the unfolded state.
- Very difficult problem, search space is gigantic.

Ab-initio structure prediction given only the sequence as input - not very successful.

In-silico methods

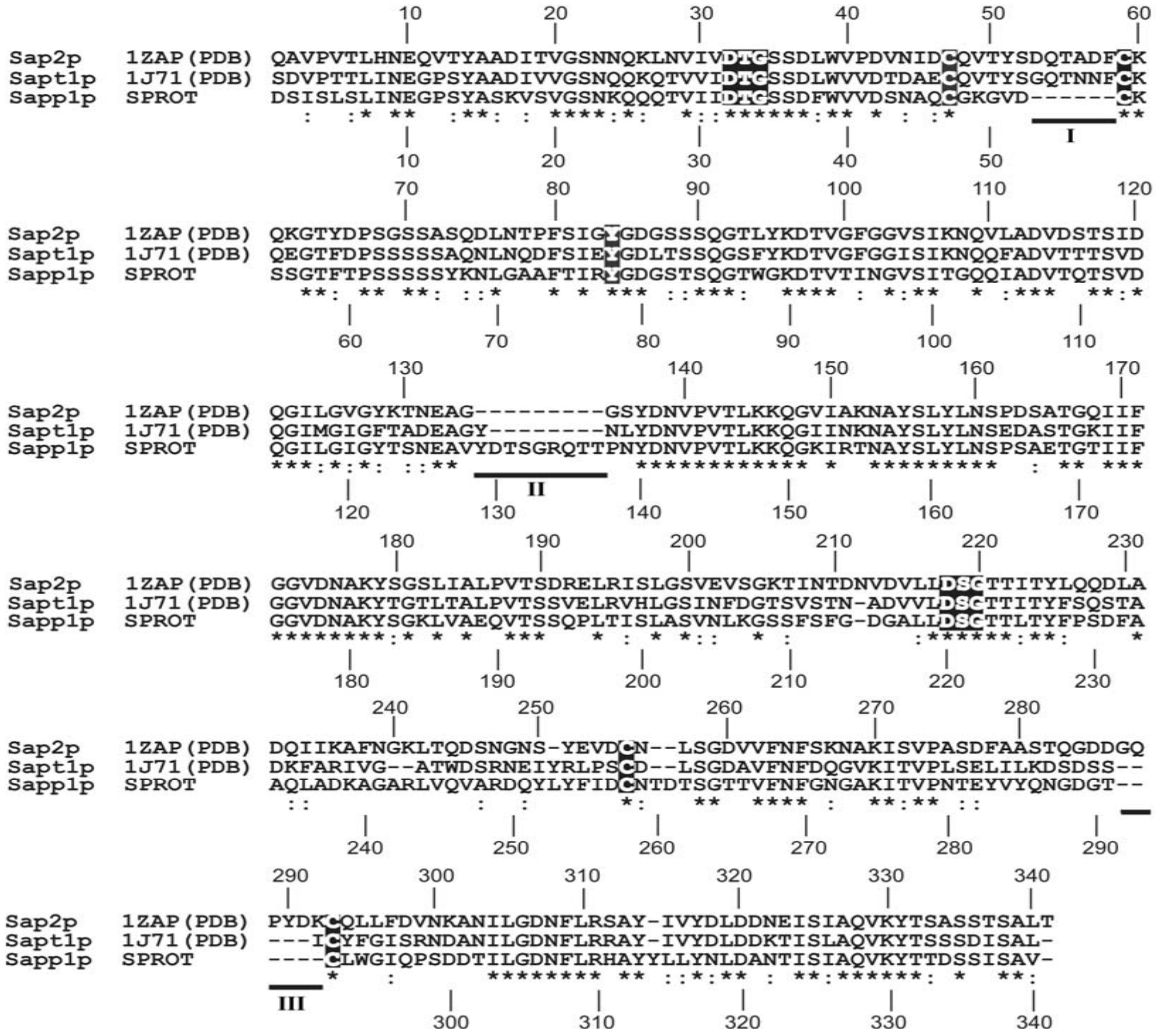
Sequence-structure alignment.
The idea is to search for a structure and sequence in existing databases of 3D structure, and use similarity of sequences + information on the structures to find best predicted structures.

We can search for GLOBAL ALIGNMENT or for LOCAL ALIGNMENT

Why structural alignment?

- Structural similarity can point to remote evolutionary relationship
- Shared structural motifs among proteins suggest similar biological function
- Getting insight into sequence-structure mapping (e.g., which parts of the protein structure are conserved among related organisms).
- This can be important also for prediction how a mutation influences the structure

Alignment of Primary Structures of Proteases from Candida spp.

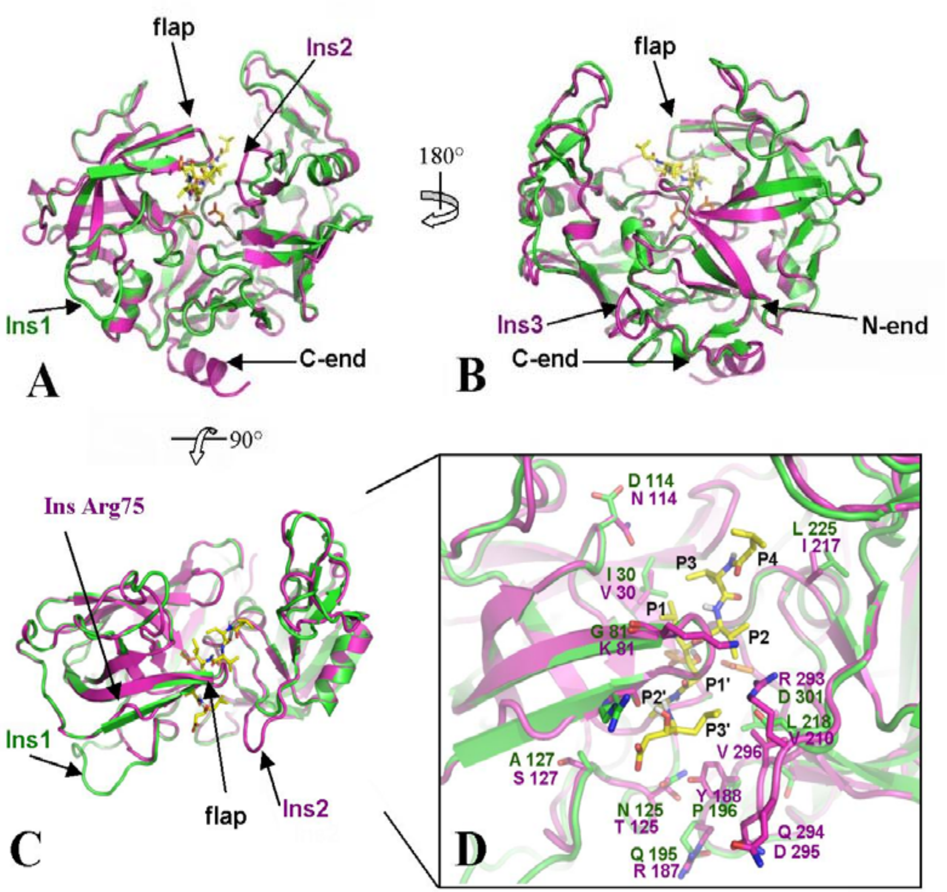


```
                            10        20        30        40        50        60
                             |         |         |         |         |         |
Sap2p    1ZAP(PDB)  QAVPVTLHNEQVTYAADITVGSNNQKLNVIVDTGSSDLWVPDVNIDCQVTYSDQTADFCK
Sapt1p   1J71(PDB)  SDVPTTLINEGPSYAADIVVGSNQQKQTVVIDTGSSDLWVVDTDAECQVTYSGQTNNFCK
Sapp1p   SPROT      DSISLSLINEGPSYASKVSVGSNKQQQTVIIDTGSSDFWVVDSNAQCGKGVD------CK
                      :  :* **  :**: : ****:*:  *::******:** * : :*           **
                             |         |         |         |         |     I
                            10        20        30        40        50

                            70        80        90       100       110       120
                             |         |         |         |         |         |
Sap2p    1ZAP(PDB)  QKGTYDPSGSSASQDLNTPFSIGYGDGSSSQGTLYKDTVGFGGVSIKNQVLADVDSTSID
Sapt1p   1J71(PDB)  QEGTFDPSSSSSAQNLNQDFSIEYGDLTSSQGSFYKDTVGFGGISIKNQQFADVTTTSVD
Sapp1p   SPROT      SSGTFTPSSSSSYKNLGAAFTIRYGDGSTSQGTWGKDTVTINGVSITGQQIADVTQTSVD
                      **: ** **: ::*   * * *** ::***:  **** : *:**  * :***  **:*
                         |         |         |         |         |         |
                        60        70        80        90       100       110

                           130                140       150       160       170
                             |                  |         |         |         |
Sap2p    1ZAP(PDB)  QGILGVGYKTNEAG---------GSYDNVPVTLKKQGVIAKNAYSLYLNSPDSATGQIIF
Sapt1p   1J71(PDB)  QGIMGIGFTADEAGY--------NLYDNVPVTLKKQGIINKNAYSLYLNSEDASTGKIIF
Sapp1p   SPROT      QGILGIGYTSNEAVYDTSGRQTTPNYDNVPVTLKKQGKIRTNAYSLYLNSPSAETGTIIF
                    ***:*:*: ::**            ************ *  *********  : ** ***
                         |         |  II     |         |         |         |
                       120       130       140       150       160       170

                           180       190       200       210       220       230
                             |         |         |         |         |         |
Sap2p    1ZAP(PDB)  GGVDNAKYSGSLIALPVTSDRELRISLGSVEVSGKTINTDNVDVLLDSGTTITYLQQDLA
Sapt1p   1J71(PDB)  GGVDNAKYTGTLTALPVTSSVELRVHLGSINFDGTSVSTN-ADVVLDSGTTITYFSQSTA
Sapp1p   SPROT      GGVDNAKYSGKLVAEQVTSSQPLTISLASVNLKGSSFSFG-DGALLDSGTTLTYFPSDFA
                    ********:* * *  ***   * : * *::  * :        :******:**:    *
                         |         |         |         |          |         |
                       180       190       200       210        220       230

                           240        250         260       270       280
                             |          |           |         |         |
Sap2p    1ZAP(PDB)  DQIIKAFNGKLTQDSNGNS-YEVDCN--LSGDVVFNFSKNAKISVPASDFAASTQGDDGQ
Sapt1p   1J71(PDB)  DKFARIVG--ATWDSRNEIYRLPSCD--LSGDAVFNFDQGVKITVPLSELILKDSDSS--
Sapp1p   SPROT      AQLADKAGARLVQVARDQYLYFIDCNTDTSGTTVFNFGNGAKITVPNTEYVYQNGDGT--
                     ::           :  :      *:   **  **** :  **:** ::
                          |         |         |         |         |         |
                        240       250       260       270       280       290

                    290       300       310        320       330       340
                      |         |         |          |         |         |
Sap2p    1ZAP(PDB)  PYDKCQLLFDVNKANILGDNFLRSAY-IVYDLDDNEISIAQVKYTSASSTSALT
Sapt1p   1J71(PDB)  ---ICYFGISRNDANILGDNFLRRAY-IVYDLDDKTISLAQVKYTSSSDISAL-
Sapp1p   SPROT      ----CLWGIQPSDDTILGDNFLRHAYYLLYNLDANTISIAQVKYTTDSSISAV-
                        *   :      ******** ** ::*:** : **:******: *  **:
                    III         |         |         |         |         |
                               300       310       320       330       340
```

* identity,   : similarity

Alignment using CLUSTAL W and Multain

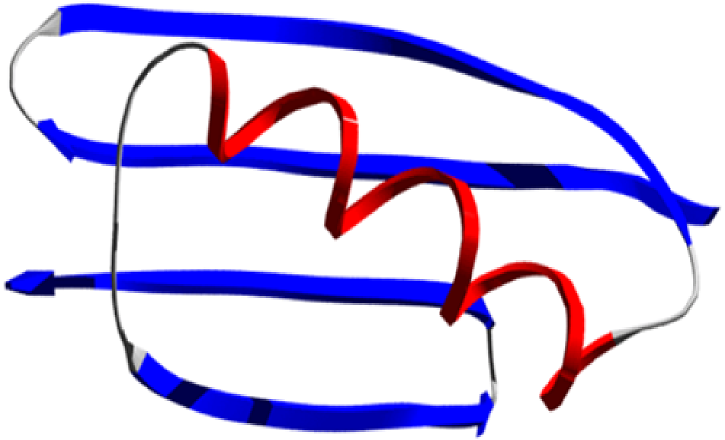
Superimposition of Sap of *C. parapsilosis* (red), a complex of Sap2p from *C.albicans* (green) with the inhibitor A70500 (1zap.ent), and Sapt1p from *C. tropicalis*(blue).



Homology modeling of Sapp1p, Sap2p, and Sapt1p structures was obtained using the program Modeller

De Novo Structure Prediction

DEIVKMSPIIRFYSSGNAGLRTYIGDHKSCVMCTYWQNLLTYESGILLPQRSRTSR



De novo design

“*Knowledge-based protein design*”

A highly challenging approach, offering the broadest possibility for new structure

- To design and construct a synthetic protein using the information on the 3-D structures of natural protein accumulates and folding rules of proteins

Methionine: prefers rigid segments, the central part of helical segments, and buried regions of natural proteins

Threonine: prefers flexible segments

Lysine: a good helix former and prefers at the C-terminal, etc.

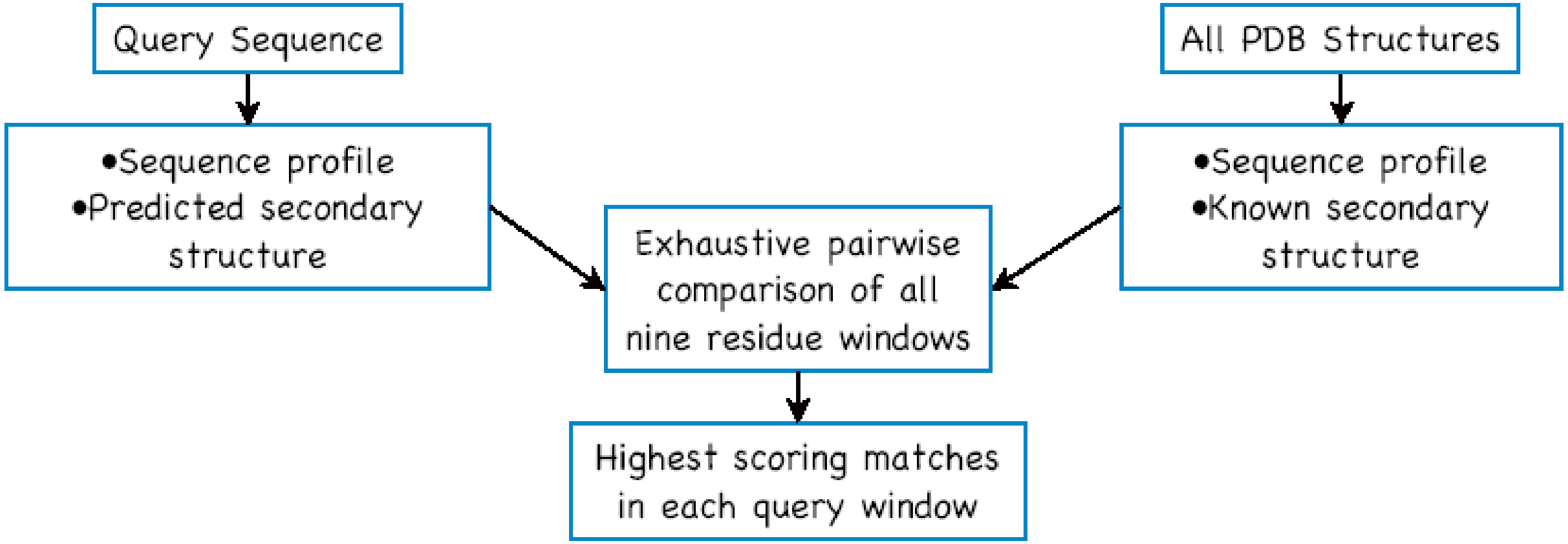
Leucine: stabilizes helices, prefers buried regions, and is usually found in the middle positions in a helix

- To find the best amino acid sequence to ensure folding in a selected structure (α-helical bundle or α/β-barrel)

Fragment-based Methods (Rosetta)

- Hypothesis, the PDB database contains all the possible conformations that a short region of a protein chain might adopt.

- How do we choose fragments that are most likely to correctly represent the query sequence?



Query Sequence

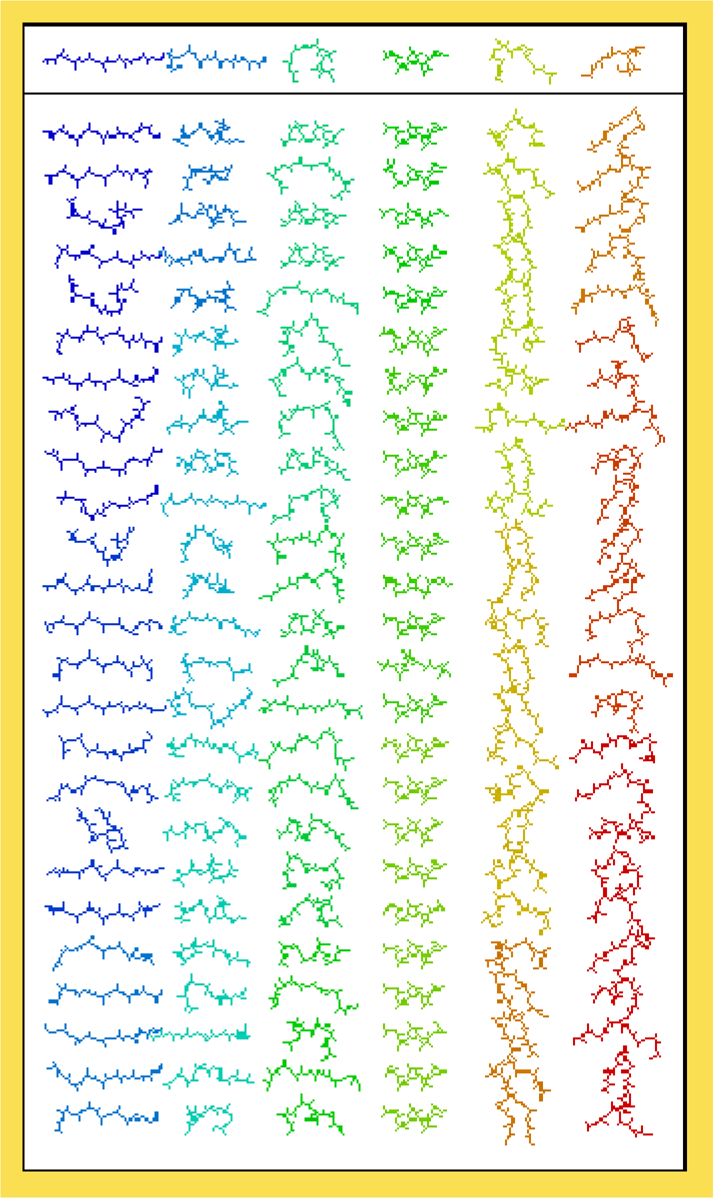
•Sequence profile
•Predicted secondary structure

All PDB Structures

•Sequence profile
•Known secondary structure

Exhaustive pairwise comparison of all nine residue windows

Highest scoring matches in each query window



Fragment Libraries

- A unique library of fragments is generated for each 9-residue window in the query sequence.
- Distributions of conformations in each window reflect conformations of this segment.
- Regions with very strong local preferences will not have a lot of diversity in the library. Regions with weak local preferences will have more diversity in the library.

Protein folding game Foldit

FoldIt is a multiplayer online game that enlists players worldwide to solve difficult protein-structure prediction problems. Foldit players leverage human three-dimensional problem-solving skills to interact with protein structures using direct manipulation tools and algorithms from the Rosetta structure prediction methodology -

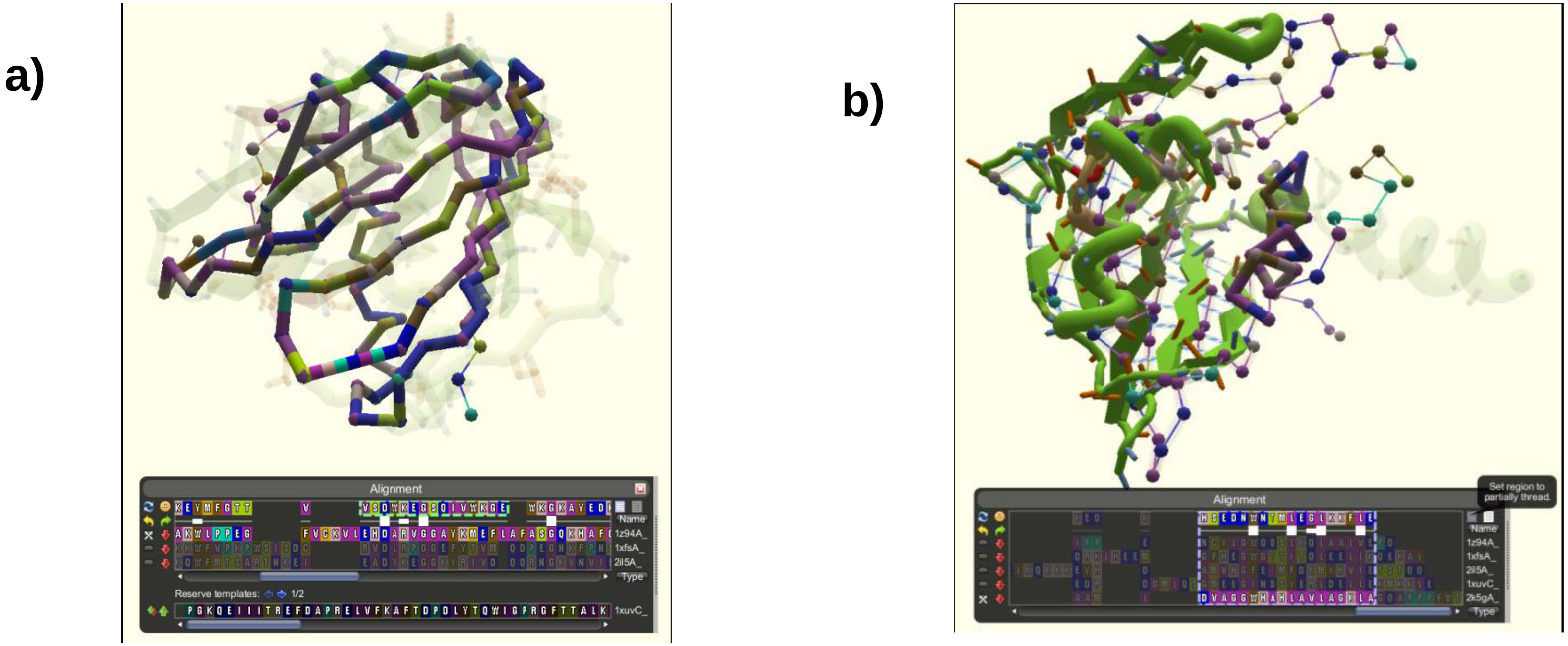
http://fold.it/portal/

The program Rosseta was designed by computation biologist David Baker of the University of Washington as an extension to his Rosetta@home program, which allows Baker to use home computers around the world to do complex calculations on protein structures. While the program ran, users would see a screen saver of the computations.

Bakers and colleagues led to allow users to alter the course of Rosetta calculations, and try to solve protein structures on their own. The goal: fold up the protein so it has the lowest energy, just as molecules tend to do in real life.

Players collaborate with teammates while competing with other players to obtain the highest-scoring (lowest-energy) models.

Foldit



a) When a template is selected, the aligned regions are represented as cylinders in the game while any unaligned regions are shown as spheres connected by lines; these graphical representations change in real time as players select residues and move the alignments around in the Alignment Tool.

b) Foldit players requested the ability to thread only a specific region from one template so partial threading was added to the Alignment Tool; this allows players to combine different regions from multiple templates into one hybrid model.

Crystal structure solved by players

The most important problem solved by Foldit players to date involves the Mason-Pfizer monkey virus (M-PMV) retroviral protease.

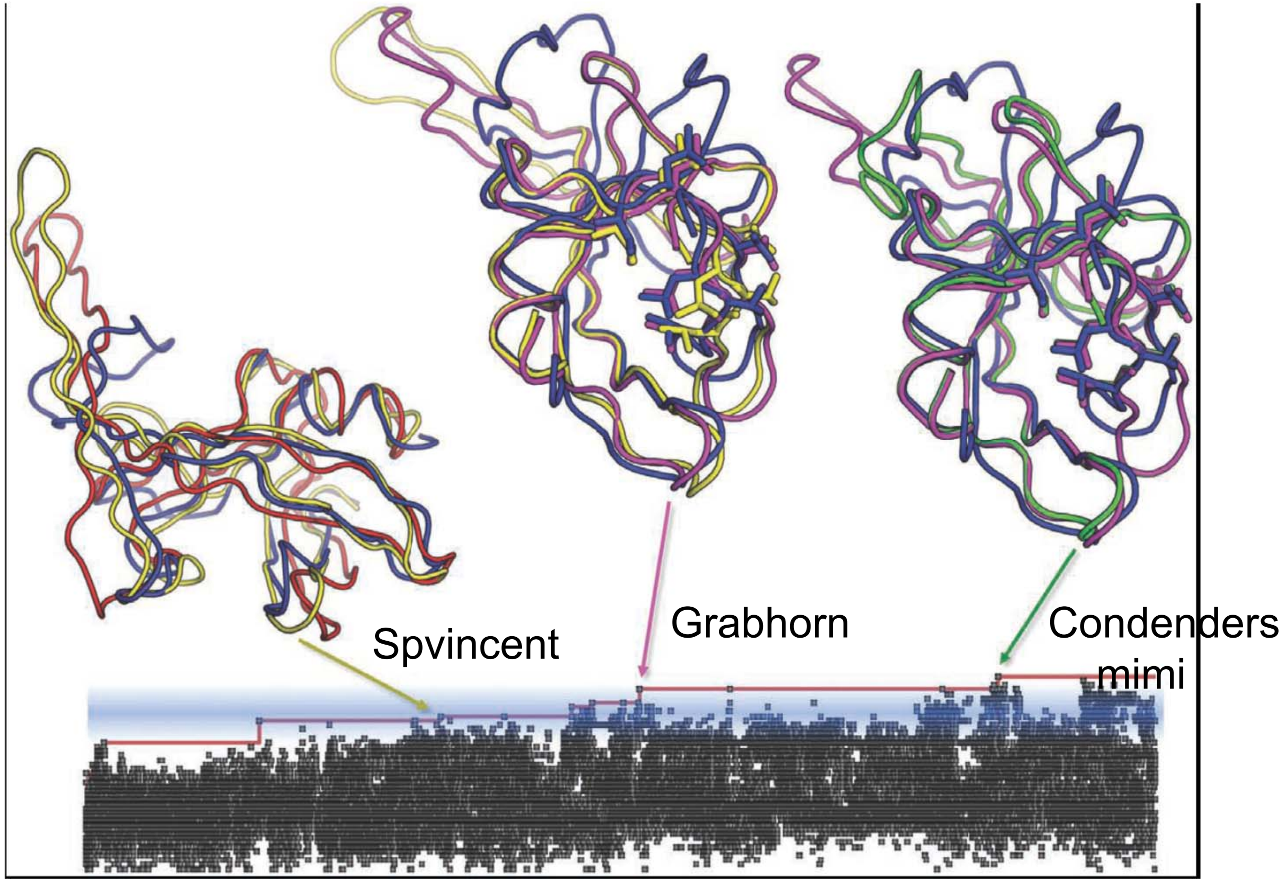
Researchers have for over a decade been unable to solve the structure by molecular replacement (MR) using either homodimer-derived models or an NMR structure of the protein monomer.

To determine whether human intuition could succeed where automated methods had failed, we challenged Foldit players to build accurate models of M-PMV PR starting from the NMR coordinates (which had failed in MR tests.

Crystal structure solved by players

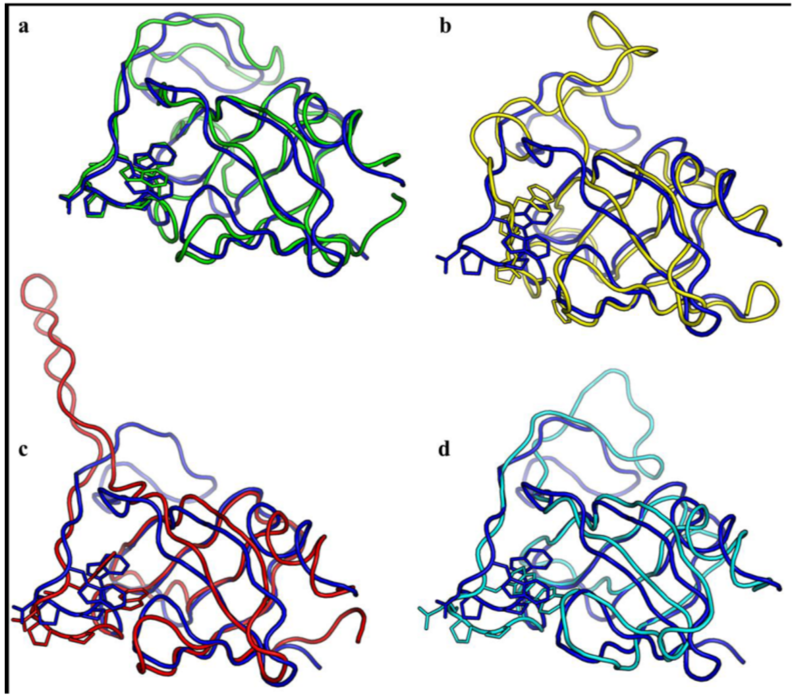
- Three weeks of competitions, 600 players, 1.25 milion solutions.
- A small group of diverse individuals living on at least three continents, who call themselves The Contenders, have solved the structure of a protein that has stumped scientists for more than 10 years. And they did so from the comfort of their own homes, playing on online protein folding game called Foldit.
- The breakthrough improvement were made in the core protein and finally in the loop region.
- Using the Foldit solution, the final refined structure was completed few days later.

M-PMV Protease Structure Improvement by the Game Players



Red: NMR structure
Blue: later crystal structure

Superpositions of different best attempts to solve the structure of M-PMV PR by molecular replacement

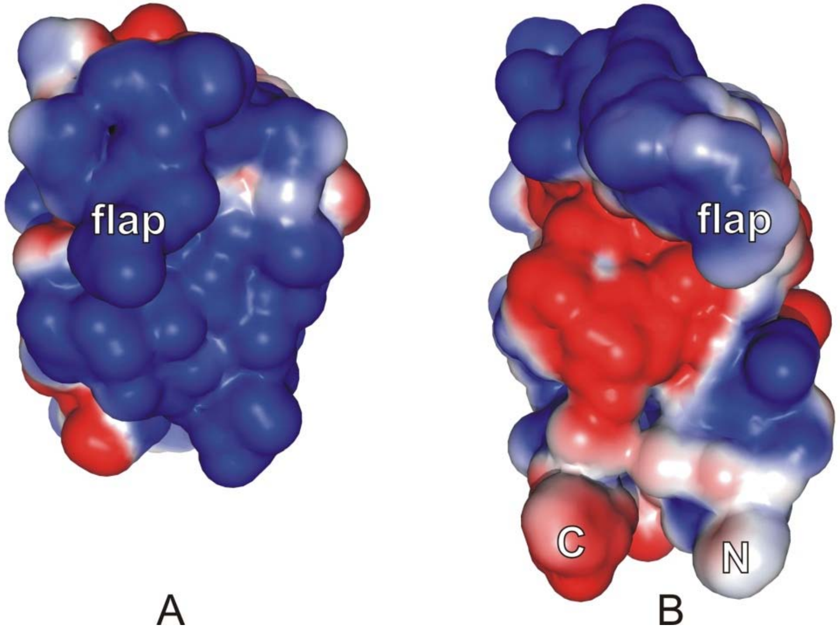


The later determined crystal structure is shown in blue.

(**a**) In green: Foldit prediction by mimi (same model shown in d) highlighting the accuracy of the core side-chain overlap in the loop at the bottom left. (**b**) In yellow: closest Rosetta prediction from the rebuild-and-refine protocol14 starting from the NMR ensemble. (**c**) In red: closest Rosetta prediction from the relax protocol15 starting from the NMR ensemble.

(**d**) In cyan: closest Rosetta prediction from CS-Rosetta16 starting from the NMR ensemble and using chemical shifts as restraints.

The structure of monomeric M-PMV protease



M-PMV PR monomer

HIV-1 counterpart extracted from the dimeric context

Khatib F, et al.. Crystal structure of a monomeric retroviral protease solved by protein folding game players.
Nat Struct Mol Biol. 2011 Sep 18;18(10):1175-1177.