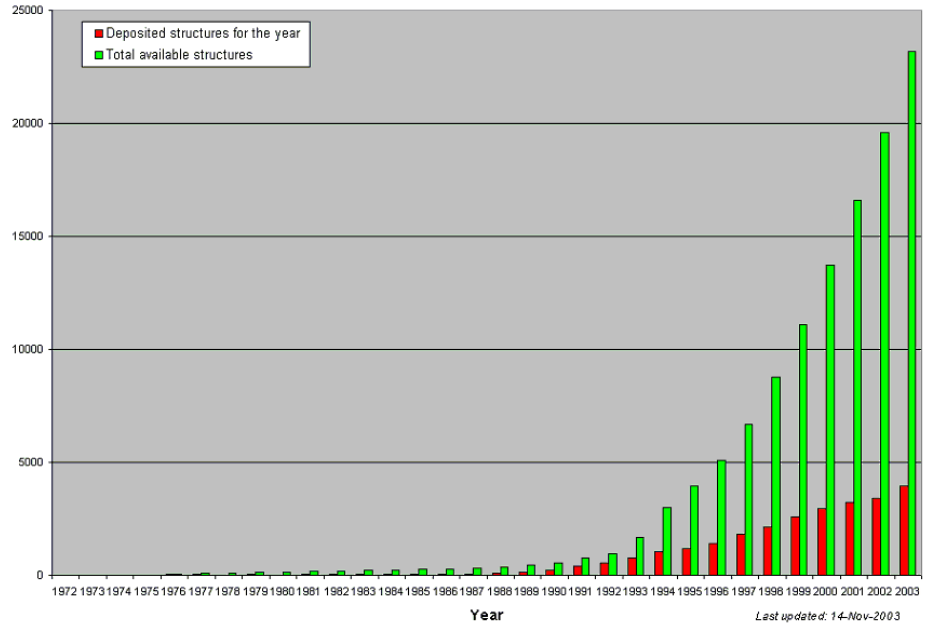
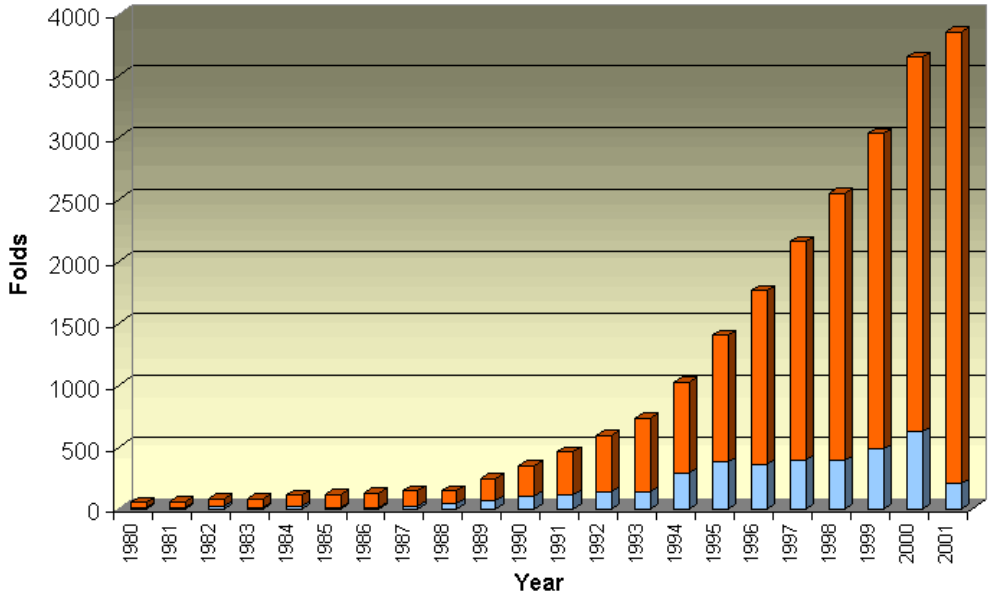
Homology modeling and Structure Prediction

Available data

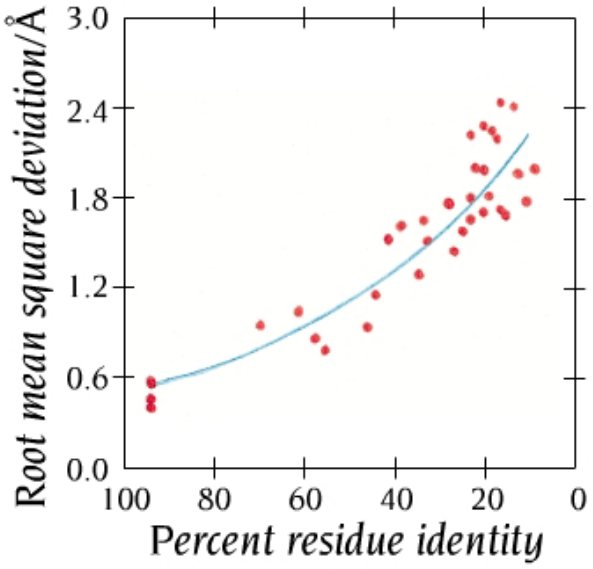
- Structure databases
  - Lots of structures
  - Fewer distinctly different "folds"
- Sequence databases
  - Homologs of unknown structure
  - Sequence comparisons of homologous structures, but
  - Structure sees homologs that sequence doesn't (1NVT vs 1DXH)

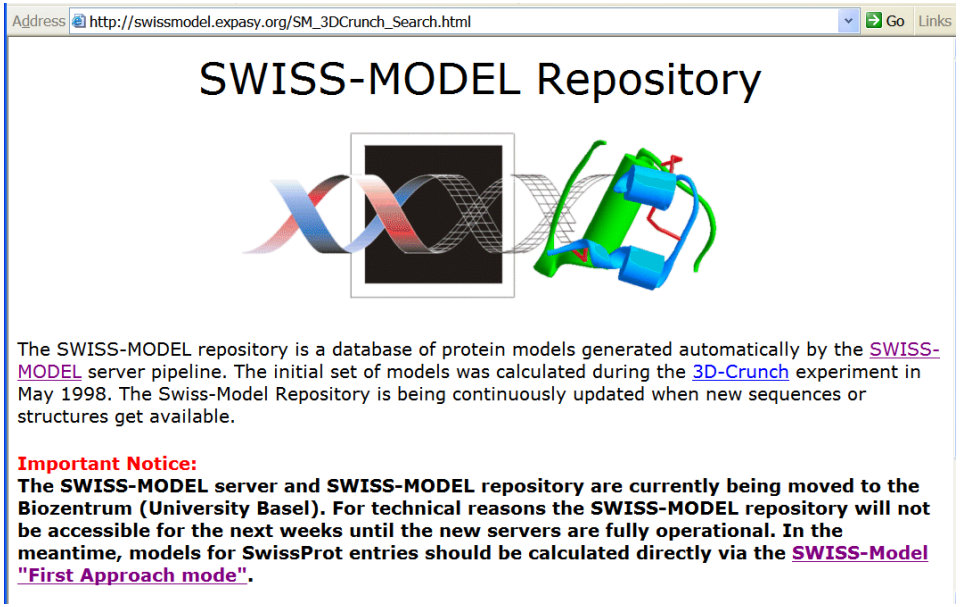




Homology modeling

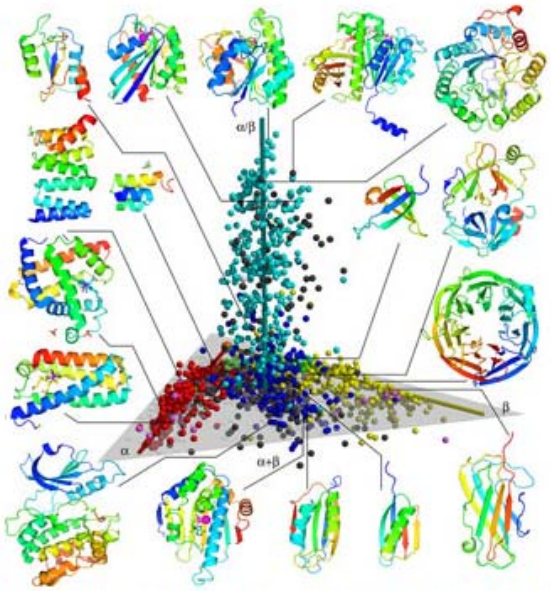
- Using sequence identity to produce a model
  - Choose or construct a *template* molecule
  - Align *target* sequence with template sequence (alpha carbons)
  - Model loops (site of most variability)
  - Add in side chain atoms
  - Refine model (energy minimization, etc)
- Automated process
  - Compare newly acquired sequences
  - Automatic modeling
  - Also can roll your own





Gotta get us some folds…

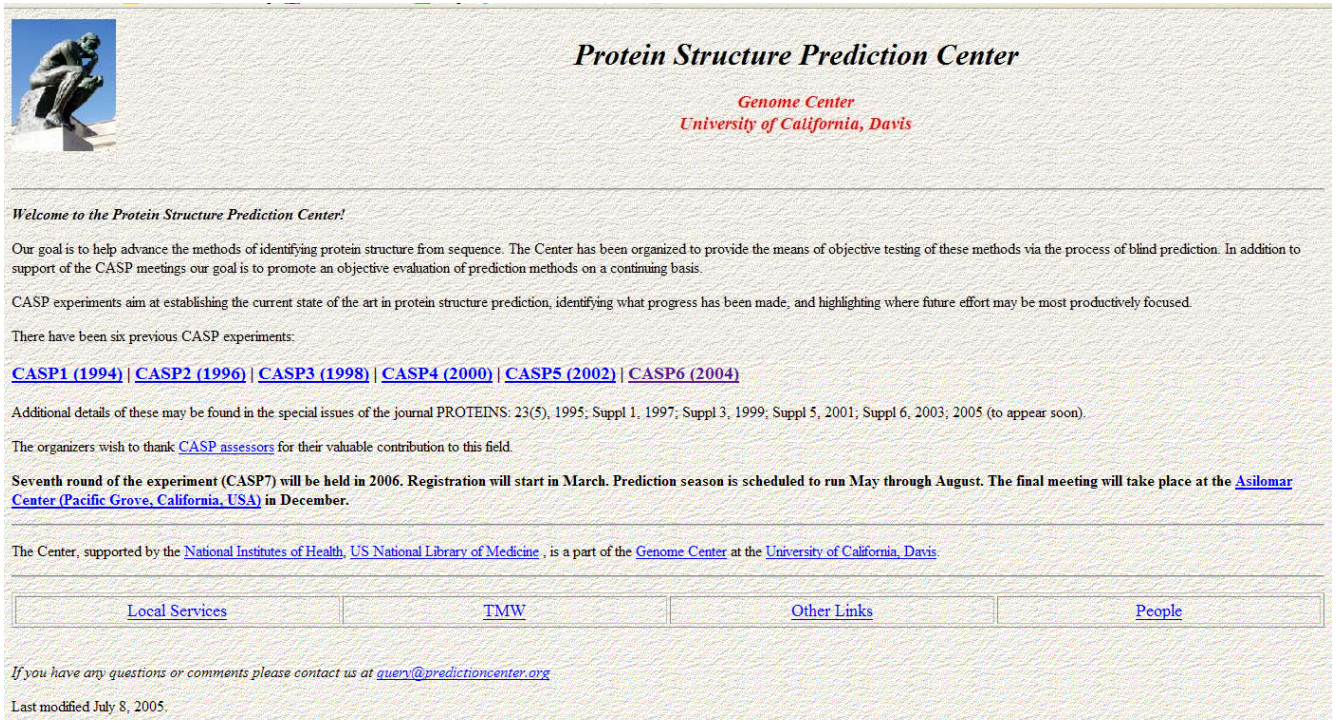
- Protein structure initiative
  - http://www.structuralgenomics.org/
  - http://www.nigms.nih.gov/Initiatives/PSI
- And the reviews are in…
- "…The PSI effort has been aimed at full coverage of fold space and sparse coverage of "sequence space." Although fold space may be nearing complete coverage, sequence space is still growing linearly with the number of deposited sequences, making the PSI an open-ended endeavor and full coverage of sequence space an unattainable goal. The large numbers of new structures determined by the PSI effort have not led to significant improvements in the accuracy of homology modeling that would allow modeling of more biologically relevant proteins, complexes or conformational states. Taken together, the lack of an end point for the PSI and the lack of modeling improvements indicate that the concepts underlying the current PSI effort are seriously flawed.



Examples of some of the structures in the Protein Structure Universe

Competitive modeling

- Critical Assessment of Structure Prediction
- predictioncenter.org
- Target sequences (determined structures)
- Prediction groups
- Target list
  - With sequence homologs: (CM)
  - Without sequence homologs: (FR)
- Summary table/3D Coordinate predictions (Models1)



Protein Structure Prediction Center

Genome Center
University of California, Davis

Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. In addition to support of the CASP meetings our goal is to promote an objective evaluation of prediction methods on a continuing basis.

CASP experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

There have been six previous CASP experiments:

CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004)

Additional details of these may be found in the special issues of the journal PROTEINS: 23(5), 1995; Suppl 1, 1997; Suppl 3, 1999; Suppl 5, 2001; Suppl 6, 2003; 2005 (to appear soon).

The organizers wish to thank CASP assessors for their valuable contribution to this field.

Seventh round of the experiment (CASP7) will be held in 2006. Registration will start in March. Prediction season is scheduled to run May through August. The final meeting will take place at the Asilomar Center (Pacific Grove, California, USA) in December.

The Center, supported by the National Institutes of Health, US National Library of Medicine , is a part of the Genome Center at the University of California, Davis.

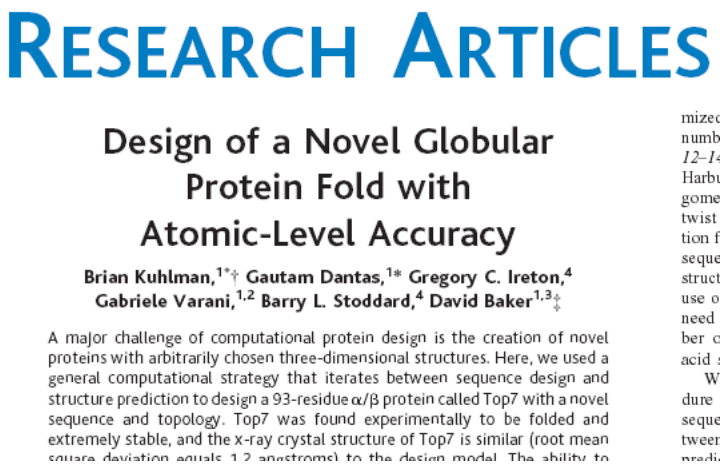
Local Services	TMW	Other Links	People

If you have any questions or comments please contact us at query@predictioncenter.org

Last modified July 8, 2005.

Predicting structures from scratch

- Designing brand new proteins
  - New amino acid sequences for known folds, or
  - Determining folds for known amino acid sequences (modeling)
- Here, iterative procedure
  - Back and forth between sequence optimization for backbone conformation, and
  - Determination of structure of backbone for given sequence.
- Target structure
  - αβ structure
  - Not in databases
  - Specify interactions (arrow)



RESEARCH ARTICLES

Design of a Novel Globular Protein Fold with Atomic-Level Accuracy

Brian Kuhlman,[1*†] Gautam Dantas,[1*] Gregory C. Ireton,[4] Gabriele Varani,[1,2] Barry L. Stoddard,[4] David Baker[1,3‡]

A major challenge of computational protein design is the creation of novel proteins with arbitrarily chosen three-dimensional structures. Here, we used a general computational strategy that iterates between sequence design and structure prediction to design a 93-residue α/β protein called Top7 with a novel sequence and topology. Top7 was found experimentally to be folded and extremely stable, and the x-ray crystal structure of Top7 is similar (root mean

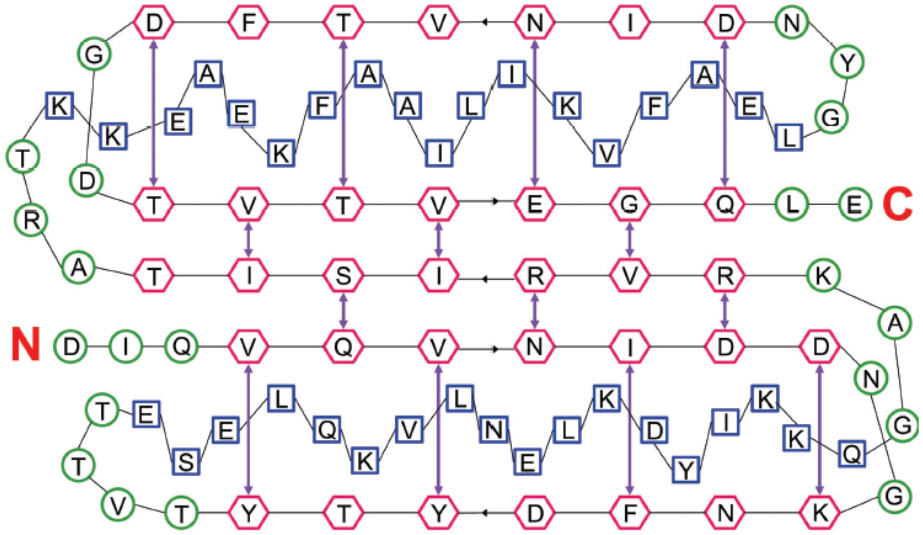


Fig. 1. A two-dimensional schematic of the target fold (hexagon, strand; square, helix; circle, other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence.

Back and forth

- Rosetta methodology for structure
  - Small (3-9 aa) sequence fragments from structural databases
  - Select those which match desired structure (helix, sheet, loops)
  - Assume that available conformations of the small fragments is essentially represented by the conformations adopted by those fragments and their close relatives in the databases
  - Assemble and minimize energy (find most favorable/common arrangements)
  - In this case, selected 172 backbones which look like drawing and satisfy constraints
- Rosetta methodology for sequence
  - Vary aa at each position, and then see if there is a rotamer that fits
    - 75 positions with all possibilities except cysteine (110 rotamers)
    - 22 positions on surface with all polar possibilities (75 rotamers)
    - $110^{71} \times 75^{22} = 10^{186}$ possibilities per backbone conformation (10 minutes on a Pentium III processor)
  - Back to backbone structure prediction

Is a given structure the right one for the aa it contains?

- Search for structure/sequence pairs of very low energy, starting with the models
  - Move 1-5 side chain (rotation) either at random, or to values found in structure databases
  - Find other aa residues which are now at a higher energy, and try to minimize those by rotating them
  - Adjust backbone for 5 residues in each direction to accommodate movement of side chains
  - Repeat 20x, and then go through and optimize whole chain by random movements
  - Look for final results that are particularly low energy combinations of a sequence, and a structure that fits the original constraints.
- Get packing just right by using atomic radii from structures (not from theory)

Pudding proofs

- Final structure
  - 70% of amino acids replaced from starting sequences
  - Final sequence nothing like anything in database
  - Synthesizing gives a compact, soluble protein
- Crystallize and determine structure
  - 1QYS
  - Perfect match
- Why was this possible
  - No functional constraints (suboptimal local conformations)
  - No folding problem (small), and no dodging alternate structures (sequence changed to get to desired end)