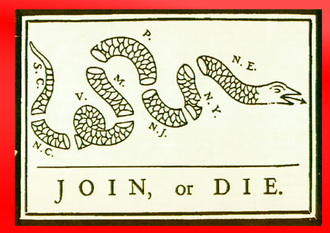
Dead Plastic Piecemeal Python Protein Design: A parallel and distributed computing solution for fine-grained rotamer searches with backbone and side chain flexibility. Application to fluorescent biosensor design.

Christian Schenkelberg†§, Derek Pitman†§, Keith Fraser†, Shounak Banerjee†, Danielle Basore†§, Alisa Neeman‡, Yao-ming Huang* Christopher Bystroff°†
Department of †Biology, °Computer Sceince, Program in §Biochemistry/Biophysics, Rensselaer Polytechnic Institute, Troy NY 12180
‡Department of Computer Science, University of Rio Grande, Rio Grande OH 45674
*Department of Bioengineering and Therapeutics, UCSF, San Francisco CA 94143



Summary

Protein design is an energy-based search for the optimal amino acid side chain configuration given a template protein backbone (or main chain). Continuous side chain conformational space is discretized into rotamers. Starting with a known protein structure, a set of design constraints, and a rotamer library, the minimum energy configuration (MEC) of rotamers is found by systematically eliminating choices using the **Dead** End Elimination theorem, or stochastically, by Monte Carlo search.

A static protein backbone is unrealistic and gives poor design results in practice. Our algorithm, called **Plastic**, models a flexible backbone as an ensemble of templates, and flexible side chains as ensembles of "child" rotamers, 3 to 81 depending on the length of the side chain. Rotamers on different templates must be siblings, children of the same parent rotamer. Templates, each on a different processor, try all siblings and synchronously send the best energy to the master node, which accepts or rejects the parent rotamer. The resulting designs better represent a protein as a flexible, moving object.

Searching for the MEC requires *O(n^2)* memory and *O(n^x)* runtime, where *n* is the number of rotamers; *n* may be tens of thousands. Our distributed computing solution, called **Piecemeal**, splits the design search space into small redundant tasks with no inter-process communication, suitable for distributed computing. The design results are gathered and merged, redistributed in a second pass, and those design result sare gathered and merged. A third pass resolves any conflicting side chain choices. Piecemeal design is faster and just as accurate as an exhaustive search, and makes verylarge and fine-grained design tasks possible. The Plastic algorithm is compatible with Piecemeal if running on a large cluster.

The algorithms are implemented in **Python** with MPI, using the PyRosetta libraries for energy calculation. Jobs are generated, queued and gathered by a secure web server through HPC2, and run on Blue Gene/Q on CCNI. Resulting designs are synthesized in the laboratory using molecular biology techniques, and are tested for desired functional properties.

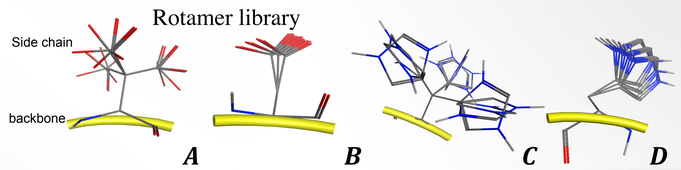
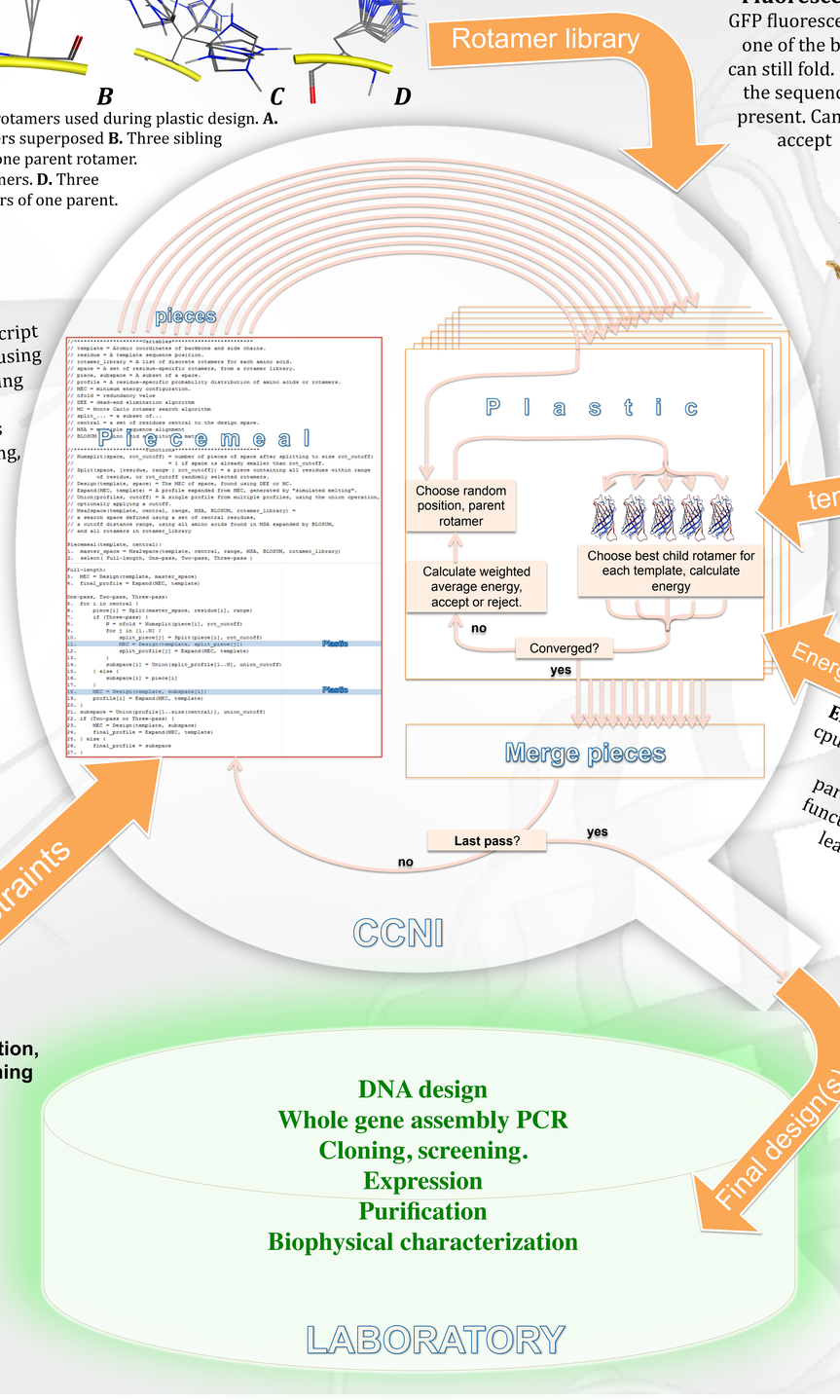
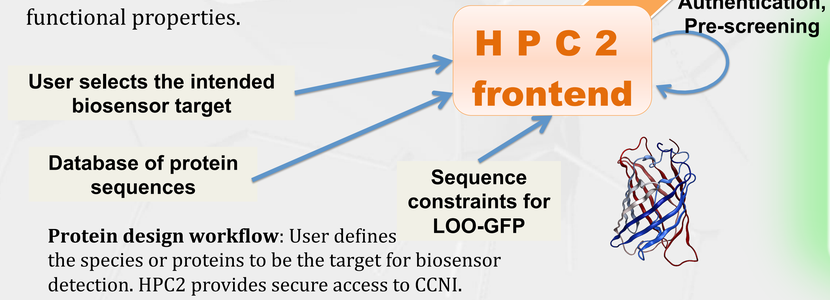


Figure 1: Examples of rotamers used during plastic design. **A.** Nine Asp parent rotamers superposed **B.** Three sibling Asp child rotamers for one parent rotamer. **C.** Nine His parent rotamers. **D.** Three sibling His child rotamers of one parent.



Blue Gene/Q: A python script runs the design pipeline, using MPI and PyRosetta. Doing parallel within distributed requires communicator splitting, MPI_COMM_SPLIT.



Protein design workflow: User defines the species or proteins to be the target for biosensor detection. HPC2 provides secure access to CCNI.

Leave-one-out Green Fluorescent Protein (LOO-GFP): GFP fluorescence siginifcantly less when one of the beta strands is left out, but can still fold. Fluorescence returns when the sequence of the missing strand is present. Can we design the LOO site to accept a different sequence?

LOO-GFP: backbone is diversified by molecular dynamics, or by using the "backrub" algorithm

Energy function: Over 100K cpu hours on CCNI's BG/L was used to optimize the parameters of the energy function, using a machine learning approach.

Results

Figure 2. Piecemeal design: The run-time order for Dead End Ellimination design is power-law with rotamer numbe, and for MC rotamer search it isis exponential. Splitting jobs into redundant overlapping pieces speeds the process.

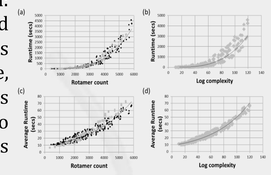


Figure 3: Plastic design of a LOO-GFP biosensor showing **A.** the target strand 139-YNFNSHNVYITAD-151 and **B.** the regions around target residue Phe141. Pink: Designed target residues, Red: Designed non-target residues, Blue: Wildtype residues (not designed)

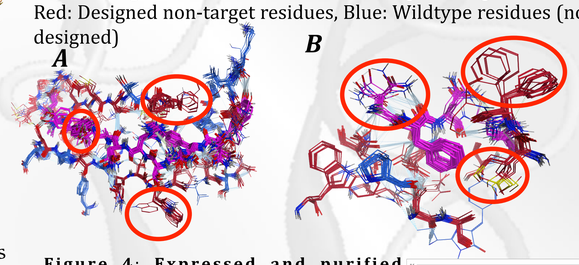
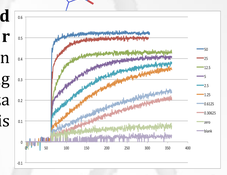


Figure 4: **Expressed and purified computationally designed biosensor LOO7_HA4_DS** shows a concentration dependent fluorescence response to increasing amounts of a peptide from H5N1 H. Influenza virus hemagglutinin. The K_d for the peptide is approximately 0.6uM.



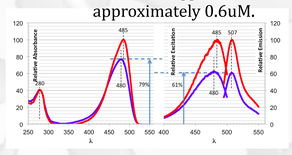
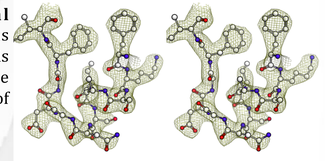


Figure 5: **Fluorescence excitation and emission spectra for AT-GFP**, a computationally designed variant of GFP with a 2-residue insertion that makes the protein fold faster and more efficiently.

Figure 6: **X-ray crystal structure of AT-GFP**, was solved to 2.55Å resolution. This stereo image shows the electron density in the region of the designed loop.



Publications

Schenkelberg CD, Pitman DJ, Neeman A, Huang Y-M & Bystroff C. (2013) Plastic protein design: A novel design algorithm using backbone and side-chain ensembles to model protein flexibility. Protein Science (in preparation).

Huang Y-M, Banerjee S, Paredes DI, Crone DE, Buck PM, Dordick J & Bystroff C. (2013) Computationally designed biosensors based on leave-one-out green fluorescent protein. Proteins (in preparation).

Pitman DJ, Christian Schenkelberg CD, Huang Y-M, DiTursi D, Teets FD & Bystroff C. (2013) Improving computational efficiency and tractability of protein design using a piecemeal approach. A strategy for parallel and distributed protein design. Bioinformatics (in review)

Rosenman DJ, Huang Y-M, Xia K, Van Roey P & Bystroff C. (2013) Green-lighting Green Fluorescent Protein: Faster and more efficient folding by eliminating a cis-trans peptide isomerization event. Protein Science (in review) abstract

Huang Y-M, & Bystroff C. (2013) Expanded explorations into the optimization of an energy function for protein design. IEEE/ACM Trans Comput Biol Bioinform. (in press)

Huang Y-M, & Bystroff C. (2012) Exploring objective functions and cross-terms in the optimization of an energy function for protein design. ACM BCB 2012. Orlando FL, Oct 7, 2012.

Crone DE, Huang Y-M, Pitman DJ, Schenkelberg C, Fraser K, Macari S, & Bystroff C. (2012) GFP-based Biosensors State of the Art in Biosensors / Book 1 ISBN 980-953-307-669-5.

Ramakrishnan V, Srinivasan S, Salem SM, Zaki MJ , Matthews SJ, Colon W, & Bystroff C. (2012) GeoFold: Topology-based protein unfolding pathways capture the effects of engineered disulfides on kinetic stability. Proteins 80(3):920-934.

Huang Y-M, Nayak S, Bystroff C. (2011) Quantitative in vivo solubility and reconstitution of truncated circular permutants of green fluorescent protein. Protein Science 20(11):1775-80

Reeder PJ, Huang Y-M, Dordick JS & Bystroff C. (2010) A Rewired Green Fluorescent Protein: Folding and Function in a Nonsequential, Noncircular GFP Permutant. Biochemistry 49(51), 10773-10779. (e-pub).

Huang Y-M & Bystroff C. (2009). Specific complementation and reconstitution of flourescence from a circularly permuted, truncated green fluoresent protein. Biochemistry 48(5):929-40.

Funded by NIH Grant No. R01GM099827