GECCO 2014 Tutorial

Evolutionary Search Algorithms for Protein Modeling:

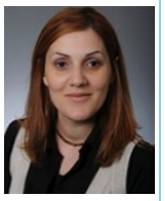
From *De novo* Structure Prediction to Comprehensive Maps of Functionally-relevant Structures of Protein Chains and Assemblies

Amarda Shehu and Kenneth A De Jong
Department of Computer Science,
George Mason University
Fairfax, Virginia, USA
[amarda, kdejong]@gmu.edu

http://www.sigevo.org/gecco-2014/

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-2662-9/14/07.

Brief Biosketch of Tutorial's Organizers



- **Amarda Shehu** is an Assistant Professor at George Mason University in the department of Computer Science. Shehu's research contributions are in biomolecular modeling and simulation, with a focus on issues concerning the relationship between sequence, structure, dynamics, and function. Shehu has unique expertise in tight coupling of probabilistic search and optimization techniques with computational protein biophysics. Shehu is an active member of the Bioinformatics and Computational Biology ACM and IEEE community and has been involved in co-organizing workshops, tutorials, and conferences in these communities.

- **Kenneth De Jong** is a University Professor at George Mason University. He is a senior and well-known researcher in the EC community with a rich and diverse research profile. De Jong's research interests include genetic algorithms, evolutionary computation, machine learning, and adaptive systems. He is an active member of the Evolutionary Computation research community and has been involved in organizing many of the workshops and conferences in this area. He is the founding editor-in-chief of the journal Evolutionary Computation (MIT Press), and a member of the board of ACM SIGEVO.

Research in Computational Structural Biology Boils Down to Three Fundamental Questions:

What does the part look like?

Mechanistic view == Shape/Form Governs Function

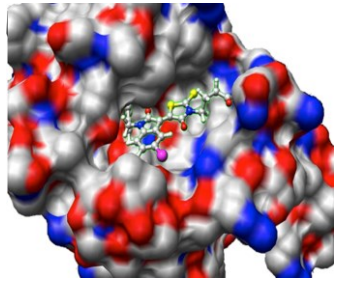
How do the parts move?

How do the parts come together?

Mechanistic View: Molecular Shape Governs Function

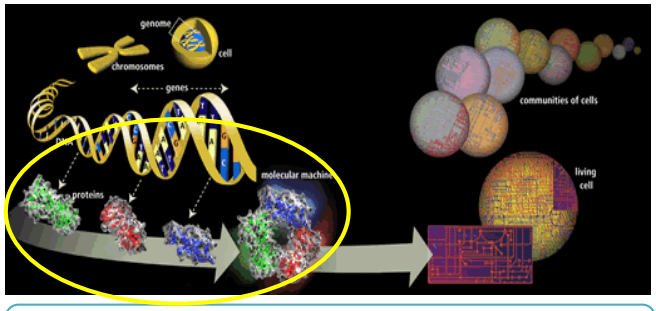
"Mechanistic View"

Protein molecule bound to a small molecule/drug compound



Source: Wikipedia

Biomolecular Modeling to Elucidate Form to Function Relationship



Computational microscope on the main macromolecules of life (DNA, RNA, **proteins**) to elucidate the molecular basis of mechanisms in the healthy and diseased cell

Objective of this Tutorial: Summarize EAs that Address These Questions

Modeling of Biologically-active (native) Structure

Ubiquitin amino-acid sequence

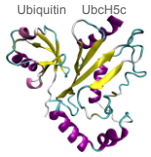
MQIFVKT…VLRLRGG

native structure and function

Ubiquitin

Modeling of Assembly into Higher-order Structures

Ubiquitin UbcH5c



Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Three Interesting Problems on Protein Modeling
  - *De novo* Protein Structure Prediction
  - Protein-Protein Assembly
  - Protein Energy Landscape Mapping
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
  - Protein-Protein Docking/Assembly
- Conclusions, Discussion, & Questions

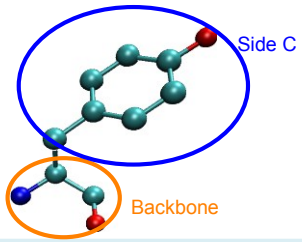
DNA

Amino Acid Sequence (AKA Protein)

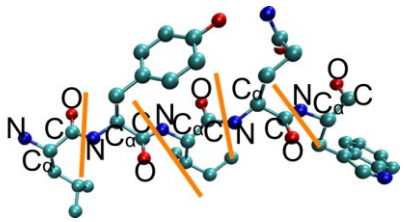
MKELVEMAVPENLVGAILGKGGKTLVEYQELTGARIQ
ISKKGEFLPGTRNRRVTITGSPAATQAAQYLISQRVT

Alphabet of 20 amino acids

Example of an Amino Acid



Amino acids link together to form a serial protein chain



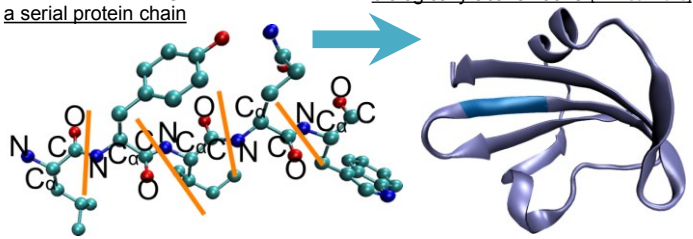
The Anfinsen Experiments: Protein Sequence Largely Determines Structure

Amino Acid Sequence (AKA Protein)

MKELVEMAVPENLVGAILGKGGKTLVEYQELTGARIQ
ISKKGEFLPGTRNRRVTITGSPAATQAAQYLISQRVT

Amino acids link together to form a serial protein chain

The chain folds to assume a spatial arrangement/conformation that is biologically-active/native [Anfinsen 1973]



[8/65]

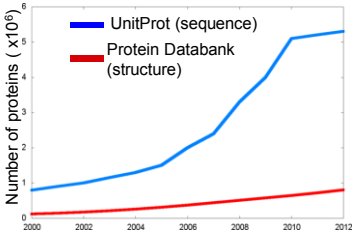
Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Three Interesting Problems on Protein Modeling
  - *De novo* Protein Structure Prediction
  - Protein-Protein Assembly
  - Protein Energy Landscape Mapping
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
  - Protein-Protein Docking/Assembly
- Conclusions, Discussion, & Questions

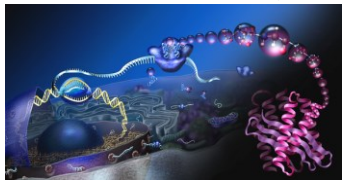
[9/65]

De Novo Protein Structure Prediction

Need to model structure!



Protein chain folds as it comes out the ribosome translational machinery

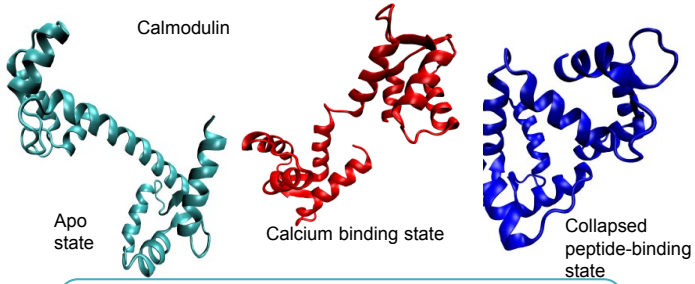


Source: Science Art

Protein Sequence → Protein Structure
[*De novo* Protein Structure Prediction] [Protein Folding]

[10/65]

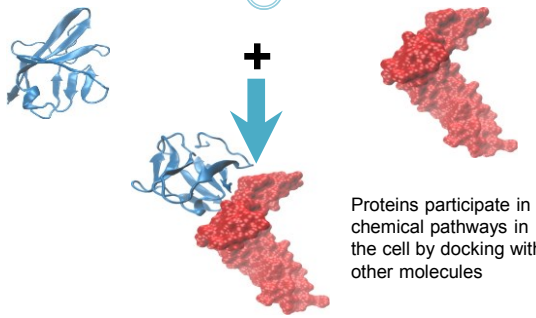
De Novo Protein Structure Prediction



Sequence → Structure(s)
Energy landscape mapping to understand flexible proteins
Understand structural effects of disease-causing or disease–involved mutations

[11/65]

Structure Prediction for Protein-Protein Assembly



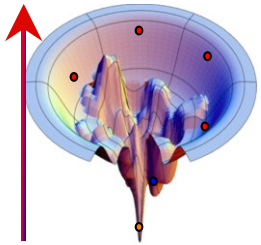
Proteins participate in chemical pathways in the cell by docking with other molecules

Structures of Protein Chains → Structure of Assembly
Protein-protein Docking

[12/65]

Protein Energy Landscapes

- Proteins as biological molecules are physics-based systems
- Physics-based systems seek state of lowest (free) energy
- Theory: Native structure (of single chain or multiple chain assembly) is that of lowest (free) energy



- unfolded conformation (start states)
- local minimum conformation
- native structure (goal state)

[13/65]

Implications of Energy Landscape View

- All three problems can be formulated as optimization problems
  - Representation
  - Scoring
  - Optimization algorithm

[14/65]

Agenda

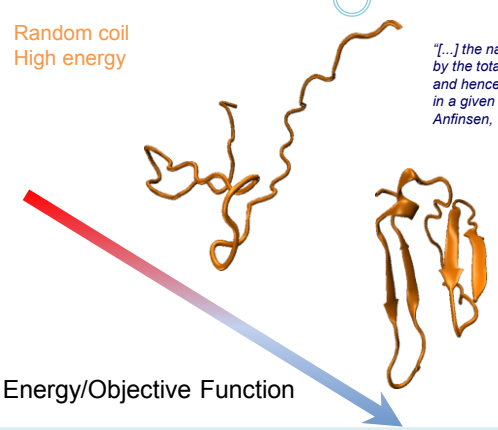
- Introduction: Proteins and their Role in Our Biology
- Overview of Three Interesting Problems on Protein Modeling
  - *De novo* Protein Structure Prediction
  - Protein-Protein Assembly
  - Protein Energy Landscape Mapping
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
  - Protein-Protein Docking/Assembly
- Conclusions, Discussion, & Questions

[15/65]

De novo Structure Prediction as an Optimization Problem

Random coil
High energy

"[...] the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment."
Anfinsen, C. B. Science 181, 1973



Native structure:
Low energy

Energy/Objective Function

[16/65]

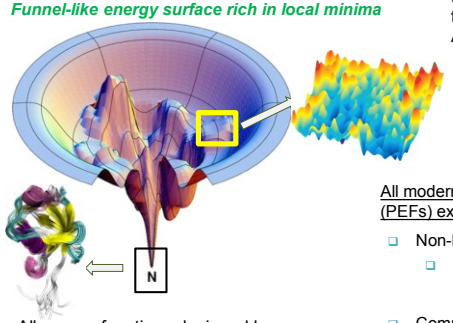
Three Essential Ingredients

- **Representation:** to keep track of spatial arrangements of the atoms in a single or multiple chain (conformations)
  - Decisions impact dimensionality and complexity of search space
- **Objective/energy function:** to score a conformation as it is obtained
- **Algorithm:** to systematically (or not) iterate over conformations
  - Needs fundamentally a technique to compute a new conformation in the chosen representation
  - Technique will be repeated in some fashion, guided by the objective function

[17/65]

Potential Energy as Objective/Fitness Function

Funnel-like energy surface rich in local minima



All energy functions designed by computational chemists are approximations of the one nature uses

Functional form of modern energy functions, such as Rosetta, AMW, AMBER, CHARMM, and others:

$$E = \alpha_1 \cdot E_{\text{Lennard-Jones}} + \alpha_2 \cdot E_{\text{H-Bond}} + \alpha_3 \cdot E_{\text{burial}} + \alpha_4 \cdot E_{\text{water}} + \alpha_5 \cdot E_{\text{Rg}}$$

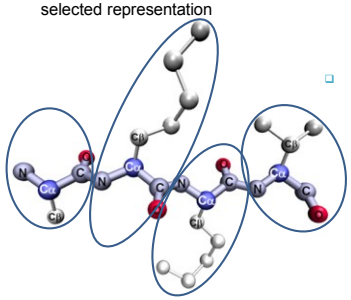
All modern potential energy functions (PEFs) exhibit these characteristics:

- Non-linear expensive terms
  - $E_{\text{Lennard-Jones}} = \Sigma_{nb}\, \varepsilon[(\text{Rmin}_{ij}/r_{ij})^{12} - (\text{Rmin}_{ij}/r_{ij})^{6}]$ computed over atom pairs
- Competing terms summed together
  - Small structural changes may increase energy
- PEF is non-linear and multimodal

[18/65]

No Ideal Representations, but Some are Better than Others

- Conformation = instantiation over selected representation



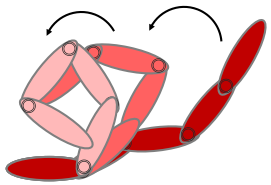
Choice #1 for representation

- Cartesian coordinate-based representation
  - N atom ➔ 3N parameters
  - Small protein: hundreds of atoms
  - High-dimensional conformation space
    - Not amenable to enumeration
  - Does not satisfy implicit constraints
    - Breaks bonds
    - Unless, perturbations follow some clever rules that move atoms together
- Amino acid building blocks in protein chain are highly coupled!
- They impose spatial constrains on one another

[19/65]

No Ideal Representations, but Some are Better than Others

- Conformation = instantiation over selected representation



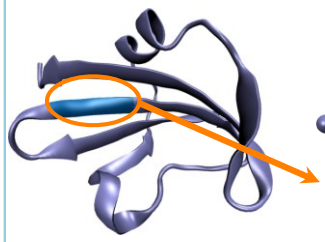
- Angular-based representations save in dimensions and are amenable to interesting perturbations
- Moreover, only some angles need to be modeled as variables (others, such as valence angles, can be ignored for structure prediction)

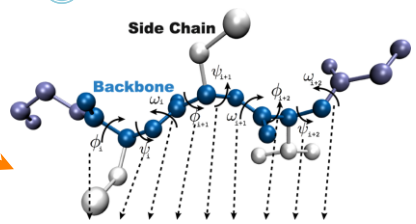
Choice #1 for representation

- Angular-based representation
  - N atom ➔ ~3N/7 parameters
  - Savings in dimensionality
    - Still hundreds of dimensions
  - Satisfies local implicit constraints
    - Does not breaks bonds
  - Does not satisfy distal constraints
    - Chain can collide with itself
  - To interface with energy function, cartesian coordinates need to be obtained by accumulating rotations (forward kinematics)

[20/65]

Angular-based Representation: State of the Art in *De Novo* Structure Prediction





- n amino acids ➔ 3n variables
- Small protein:
  - 30-50 amino acids
  - 90-150 variables

amino acid	Q			V			C		
dihedral bond	Φ	ψ	ω	ϕ	ψ	ω	ϕ	ψ	ω
angle	-122	129	-178	-123	138	176	-105	152	-179

[21/65]

Structure Prediction via Stochastic Optimization

- Conformational space: high-dimensional and continuous
- Fitness/energy surface/landscape: non-linear and multi-modal
- Not amenable for systematic optimization unless heavy use of discretization resulting in loss of accuracy
- Stochastic optimization
  - Computational biology community: Metropolis Monte Carlo algorithms
    - Focus on domain-specific insight
  - Evolutionary computation community: Evolutionary algorithms (EAs)
    - Focus on algorithmic strategies for balancing exploration/exploitation

[22/65]

EAs for De Novo Structure Prediction in EC Community

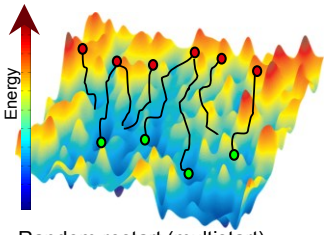
Memetic/Hybrid Evolutionary Algorithms (HEA)	Multi-Objective Evolutionary Algorithms (MOEA)
R. Faccioli, I. da Silva, L. Bortot, and A. Delbem. A mono-objective evolutionary algorithm for protein structure prediction in structural and energetic contexts. In Evolutionary Computation (CEC), 2012.	J. Calvo, J. Ortega, and M. Anguita, PITAGORAS-PSP: Including domain knowledge in a multi-objective approach for protein structure prediction. Neurocomputing, 74,(16):. 2675–2682, 2011.
M. S. Abual-Rub, M. A. Al-Betar, R. Abdullah, and A. T. Khader. A hybrid harmony search algorithm for ab initio protein tertiary structure prediction. Network Modeling and Analysis in Health Informatics and Bioinformatics, 1–17, 2012.	J. Calvo and J. Ortega, Parallel protein structure prediction by multiobjective optimization. In Proc. of Euromicro Intl Conf on Parallel, Distributed and Network-based Processing 2009, pp. 268–275.
M. M. Goldstein, E. E. Fredj, and R. B. R. Gerber. A new hybrid algorithm for finding the lowest minima of potential surfaces: approach and application to peptides. Journal of Computational Chemistry, 32(9):1785–1800, 2011.	Cutello, V, G. Narzisi, and G. Nicosia, A multi-objective evolutionary approach to the protein structure prediction problem. Journal of The Royal Society Interface 3(6): 139–151, 2006.
A.-A. Tantar, N. Melab, and E.-G. Talbi. A grid-based genetic algorithm combined with an adaptive simulated annealing for protein structure prediction. Soft Computing, 12(12):1185–1198, 2008.	R. Day, J. Zydallis, G. Lamont, and R. Pachter, Solving the protein structure pre- diction problem through a multiobjective genetic algorithm. Nanotechnology 2:32–35, 2002.

- Limited to small proteins (typically < 30 amino acids)
- Do not take advantage of domain-specific insight

Not competitive with Monte Carlo-based algorithms

[23/65]

Current state of the art: Random Restart of Monte Carlo with Fragment Replacement

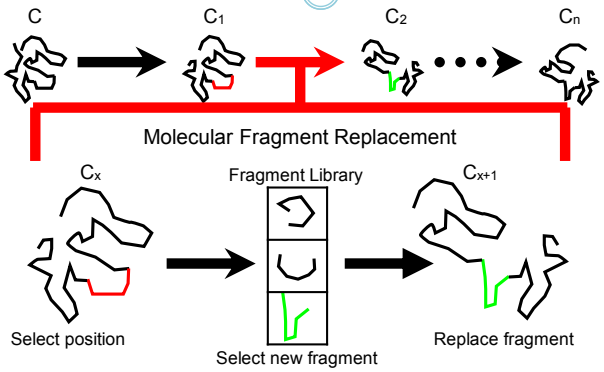


Random restart (multistart)

- Initialization: random or extended conformations
- Each trajectory is a series of conformations obtained through Metropolis Monte Carlo

- State-of-the-art for de novo structure prediction:
  - MMC-based frameworks, such as Rosetta and Quark
- Why do they outperform most EAs?
  - Dihedral angle-based representation
  - Molecular fragment replacement technique
  - Special moves
  - Special functions

The Fragment Replacement Technique



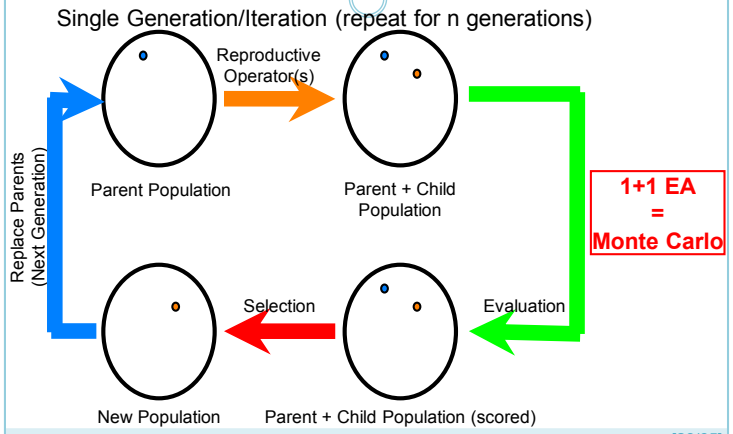
Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Three Interesting Problems on Protein Modeling
  - *De novo* Protein Structure Prediction
  - Protein-Protein Assembly
  - Protein Energy Landscape Mapping
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
  - Protein-Protein Docking/Assembly
- Conclusions, Discussion, & Questions

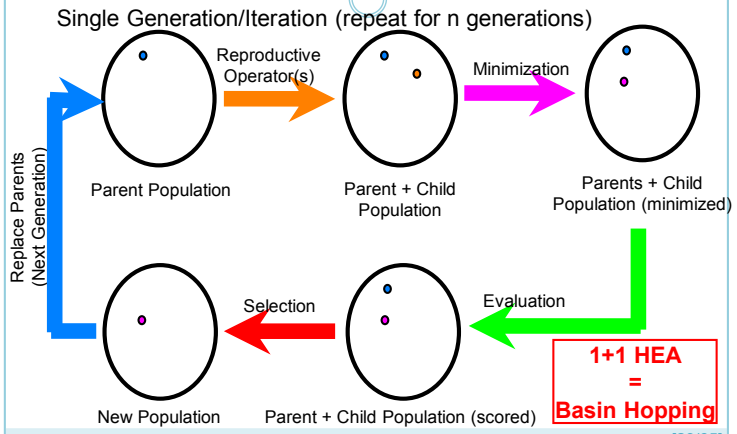
State-of-the-art EAs for *De Novo* Protein Structure Prediction

- *Lesson #1: Inject domain-specific insight*
  - Representation (Backbone dihedral angles)
  - Perturbation (Fragment replacement)
- *Lesson #2: Establish baseline EA with this insight is competitive*
- *Lesson #3: Pursue enhancements of interest to EC community*
  - Design of perturbation operators
  - Memetic/Hybrid EAs
  - Selection Mechanism
  - Multi-Objective EAs
  - Structurization
  - Further enhancements to investigate exploration/exploitation trade-off

A Baseline EA for *De Novo* Protein Structure Prediction



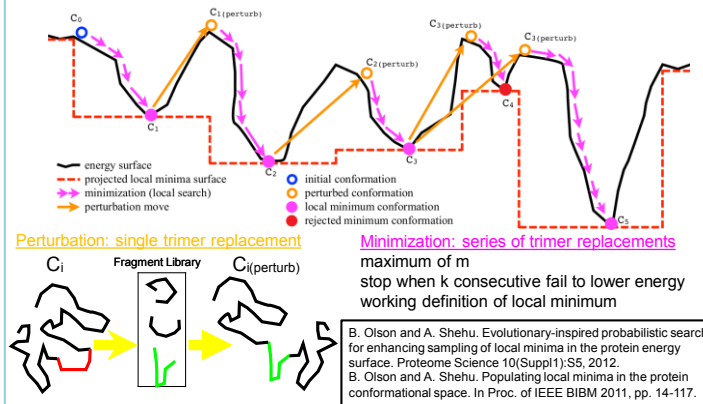
A Baseline (H)EA for *De Novo* Protein Structure Prediction



1+1 (H)EA with Domain-specific Insight: Basin Hopping (BH)

- Population size: 1
- Representation: backbone dihedral angles
- Perturbation operator: molecular fragment replacement (fl = 3)
- Local improvement operator: greedy search vs. Monte Carlo
- Evaluation with state-of-the-art (coarse-grained) knowledge-based energy functions for *de novo* structure prediction: Rosetta or AMW
- Nr. of generations: capped by total nr. of energy evaluations
- Comparison setting:
  - Versus random restart
  - Versus Rosetta (Monte Carlo-based)
  - Explored issues: magnitude of perturbation operator, cost of improvement operator

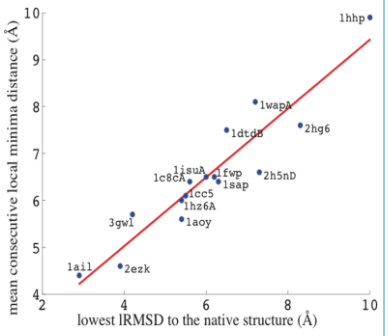
A Basin Hopping Algorithm for De Novo Structure Prediction: PLOW



B. Olson and A. Shehu. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface. Proteome Science 10(Suppl1):S5, 2012.
B. Olson and A. Shehu. Populating local minima in the protein conformational space. In Proc. of IEEE BIBM 2011, pp. 14-117.

A Basin Hopping Algorithm for De Novo Structure Prediction: PLOW

- Comparison of least root-mean-squared-deviation (lRMSD) to known native structure on many protein sequences demonstrates that it is *competitive* with Monte Carlo-based frameworks (similar lowest lRMSDs)
- Rich algorithmic vehicle
  - Adjacency between minima correlates with lowest lRMSD to the native structure
  - Results can be further improved by controlling perturbation magnitude

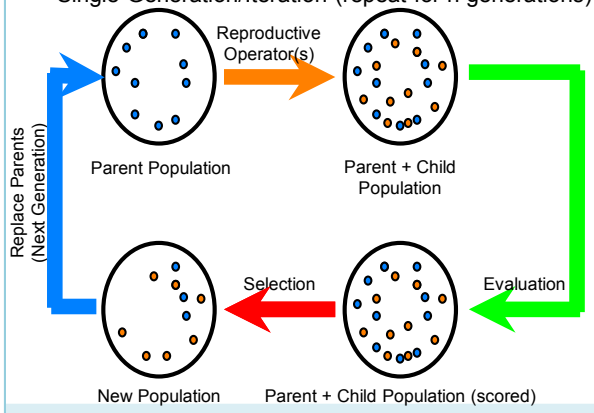


Convergence of Views of Two Communities

- BH proposed by Scheraga in 1987 under name Metropolis Monte Carlo with Minimization, applied to structure prediction for small peptides [Z. Li and H. A. Scheraga. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. PNAS 84(19):6611–6615, 1987.]
- Revived by Jones and Wales in 1997 to compute energy minima of clusters of non-bound atoms; name BH first appears [D. J. Wales and J. P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. J Phys Chem A 101 (28): 5111–5116, 1997.]
- Applied by Wolynes for *de novo* structure prediction; limited by cartesian coordinate-based representation [M. C. Prentiss, C. Hardin, M. P. Eastwood, C. Zong, and P. G. Wolynes. Protein structure prediction: the next generation. J Chem Theory and Comp 2(3):705–716, 2006.]
- Applied by Wenzel to "simulate" folding [A. Verma, A. Schug, K. H. Lee, and W. Wenzel. Basin hopping simulations for all-atom protein folding. J Chem Phys 124(4):044515, 2006.]
- Continues to be explored in computational physics and for peptides [L. Zhan, Z. Y. Chen, and W-K. Liu. Monte Carlo basin paving: An improved global optimization method. Phys. Rev. E 73, 015701(R).]
- Investigated in depth in EC community by Locatelli and Grosso [M. Locatelli. On the multilevel structure of global optimization problems. Computational Optimization and Applications 30(1):5-22, 2005.]
- Enhanced by Shehu with domain-specific insight for *de novo* structure prediction and protein-protein docking (showcased in this tutorial)

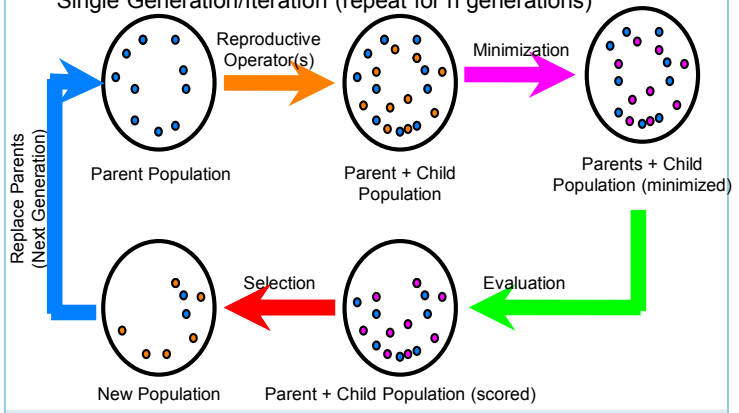
Population-based EAs

Single Generation/Iteration (repeat for n generations)



Population-based Hybrid EAs

Single Generation/Iteration (repeat for n generations)

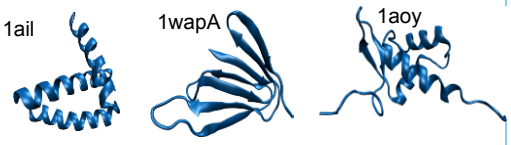


Showcase of HEA in Olson et al. GECCO 2013 Testing Dataset and Experimental Setup

	PDB ID	Length	Fold
1	1bq9	53	α/β
2	1dtdB	61	α/β
3	1isuA	62	α/β
4	1c8cA	64	α/β
5	1sap	66	α/β
6	1hz6A	67	α/β
7	1wapA	68	β
8	1fwp	69	α/β
9	1ail	70	α
10	1dtjA	76	α/β
11	1aoy	78	α/β
12	2ci2	83	α/β
13	1cc5	83	α
14	1tig	88	α/β
15	2ezk	93	α
16	1hhp	99	β
17	3gwl	106	α
18	2hg6	106	α/β
19	2h5nD	123	α
20	1aly	146	β

PDB stands for Protein Data Bank; PDB ids are unique to protein structures deposited by wet-laboratory groups

Experimentally-determined native structures



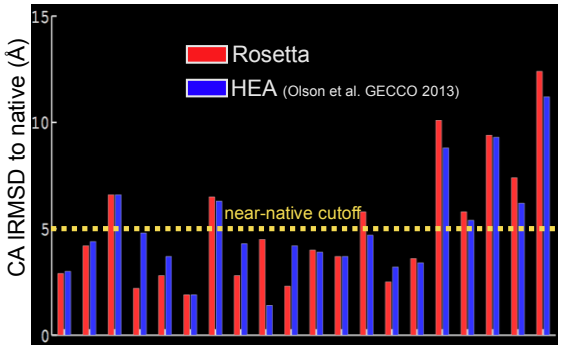
Population size : 500 Elitism rate : HEA=25%
# of Generations : variable
Fragment libraries and energy function (score4) as in Rosetta

10 million energy function evaluations 2.4 GHz Core i7
12-24 hours of CPU user time (depending on protein length)
Repeated five times (total 50 million energy evaluations)

Comparison with Rosetta on several metrics (including lowest lRMSD)

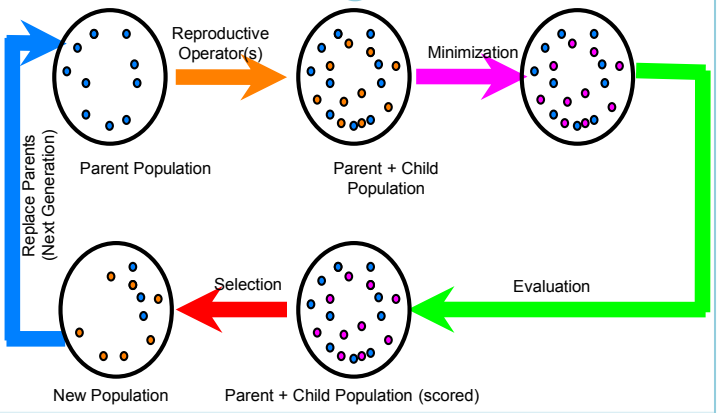
Lowest lRMSD to Native Structure Sampled by Each Algorithm

1aly
2h5nD
2hg6
3gwl
1hhp
2ezk
1tig
1cc5
2ci2
1aoy
1dtjA

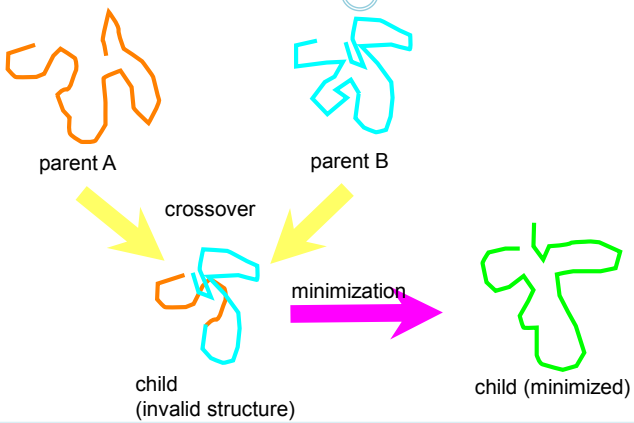


Smallest (53 AA) Testing Dataset Largest (146 AA)

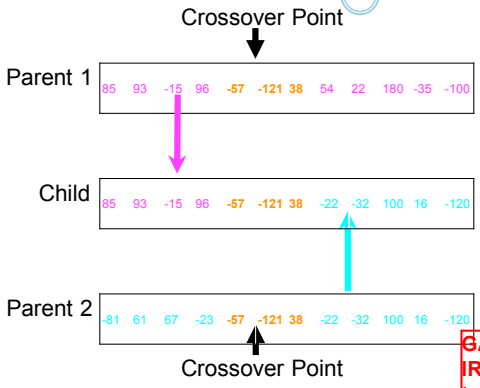
Rich Algorithmic Settings to Explore Role of Perturbation Operators, MOO, and more



Hybrid Genetic Algorithm (GA): HEA + Crossover



Homologous Crossover for Protein Structures



Mean reduction in conformational sampling due to crossover:

1 point	11%
2 point	21%
Homologous	7%

GA outperforms in lowest IRMSD Monte Carlo techniques like FixIt
(Olson et al. GECCO 2013)

Multi-objective EAs for Handling Imperfections in Energy Function

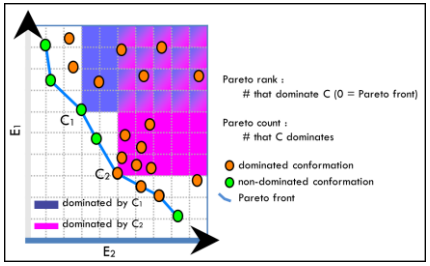
Evaluation on multiple objectives rather then on aggregate energy score

$E = \alpha_1 \cdot E_{Lennard\text{-}Jones} + \alpha_2 \cdot E_{H\text{-}Bond} + \alpha_3 \cdot E_{burial} + \alpha_4 \cdot E_{water} + \alpha_5 \cdot E_{Rg}$

E_1 = $\alpha_1 \cdot E_{Lennard\text{-}Jones} + \alpha_2 \cdot E_{H\text{-}Bond}$

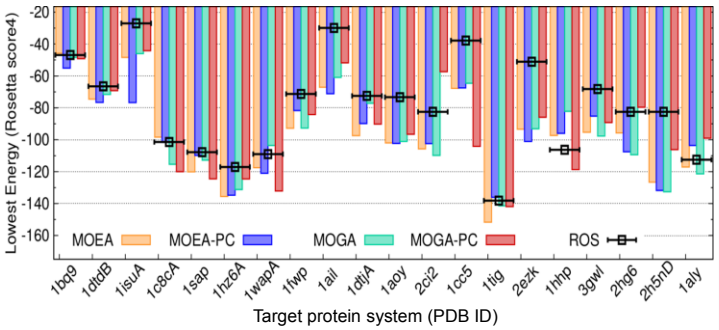
E_2 = $\alpha_3 \cdot E_{burial} + \alpha_4 \cdot E_{water} + \alpha_5 \cdot E_{Rg}$

Multi-objective



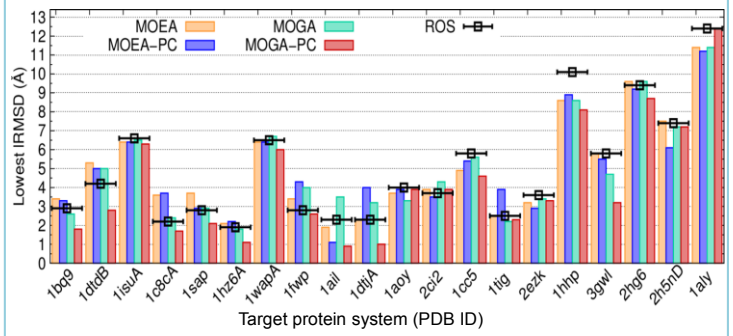
Sorting by Pareto rank, Pareto count, and total energy in this order in Olson et al. BiCoB 2014 and ACM BCB 2013
Truncation-based selection

Enhanced Exploration Capability over Monte Carlo-based Algorithms



Olson et al. BiCoB 2014

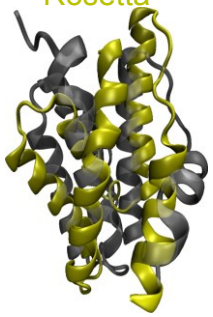
Higher Accuracy over Monte Carlo-based Algorithms



Olson et al. BiCoB 2014

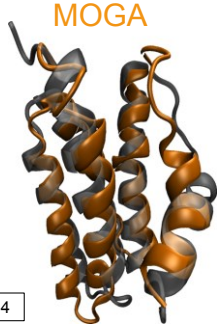
Better Models Over State of the Art

Rosetta



PDB ID
3gwl

MOGA



Olson et al. BiCoB 2014

5.8 Å RMSD to native

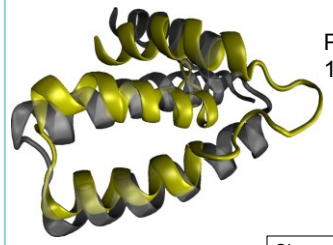
3.2 Å RMSD to native

Compared to experimentally-determined native structure (in gray)

[44/65]

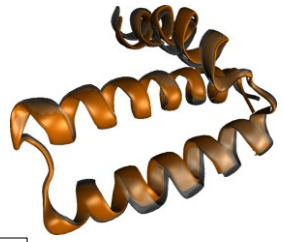
Better Models Over State of the Art

Rosetta



PDB ID
1dtjA

MOGA



Olson et al. BiCoB 2014

4.5 Å RMSD to native

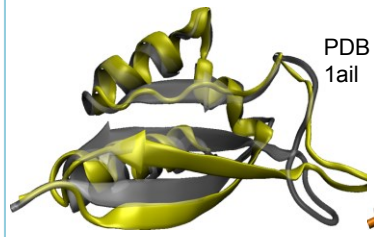
0.9 Å RMSD to native

Compared to experimentally-determined native structure (in gray)

[45/65]

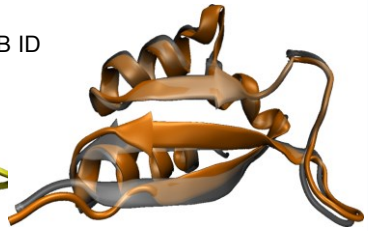
Better Models Over State of the Art

Rosetta



PDB ID
1ail

MOGA



Olson et al. BiCoB 2014

2.3 Å RMSD to native

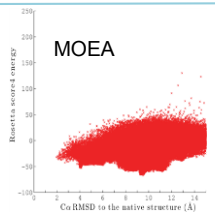
1.0 Å RMSD to native

Compared to experimentally-determined native structure (in gray)

[46/65]

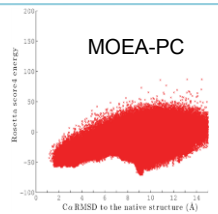
Improved Funneling and better Exploration

MOEA

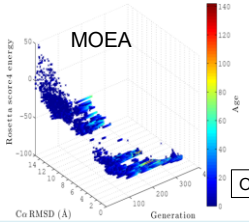


PDB ID : 1ail

MOEA-PC

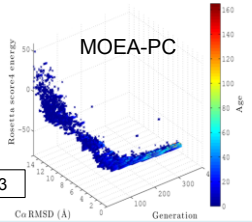


MOEA



PDB ID : 1dtjA

MOEA-PC



Olson et al. ACM BCB 2013

[47/65]

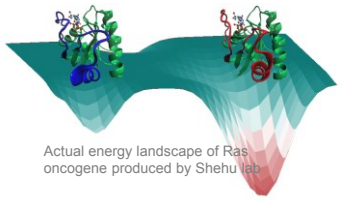
Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Three Interesting Problems on Protein Modeling
  - *De novo* Protein Structure Prediction
  - Protein-Protein Assembly
  - Protein Energy Landscape Mapping
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
  - Protein-Protein Docking/Assembly
- Conclusions, Discussion, & Questions

[48/65]

Single-basin View Does not Apply to All Proteins: Structure Tuning for Function Tuning

- Existence of multiple basins
  - Exploited by multi-functional proteins to modulate function
- Some proteins have degenerate landscapes
  - *Are prone to non-deleterious mutations*
  - *Example: disease-causing mutations in RAS oncogene, ALS SOD1, and more*



Actual energy landscape of Ras oncogene produced by Shehu lab

- HEAs produce a discrete representation of the energy landscape through the population of sampled minima
- More information obtained about energy landscape than utilized in strict context of *de novo* structure prediction

[49/65]

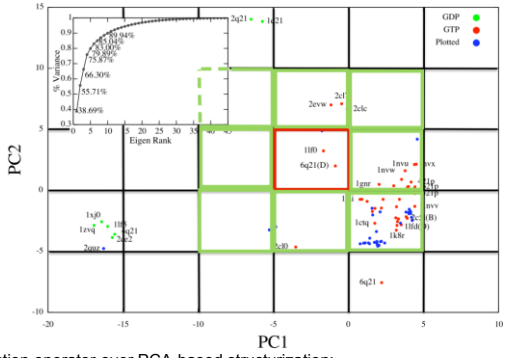
Enhancing Exploration via Structurization

- 1. By exploiting experimentally-known disease-causing structures to define search space of interest
  - How: Through Principal Component Analysis (PCA) of known structures
  - Cartesian coordinate-based representations can be used and groups of atoms can be moved together in directions encoded by PCs
- 2. PCA-obtained embedding defines meaningful structurization of search space
  - *Can be used to enhance diversity via a local selection operator*
- *Demonstration: on RAS oncogene to explain the structural impact of cancer-causing mutations*

Clausen et al. J Chem Theory and Comp 2014

[50/65]

PCA-based Structurization and Local Selection

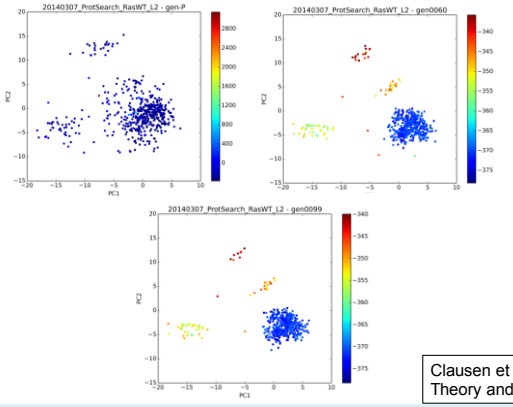


Local selection operator over PCA-based structurization:
Red cell indicates where child falls. Cells outlined in green (l=1) are included for comparison. All black cells are included if l=2.

Clausen et al. J Chem Theory and Comp 2014

[51/65]

Energy Landscape of RAS

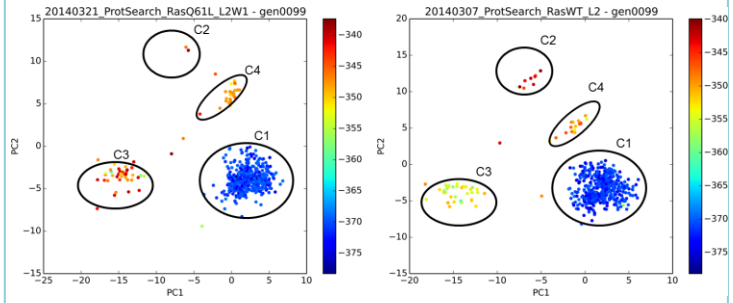


Clausen et al. J Chem Theory and Comp 2014

Ability to Explain Changes to Landscape Upon Sequence Mutations

Clausen et al. J Chem Theory and Comp 2014

Q61L mutation vs. wildtype (WT)

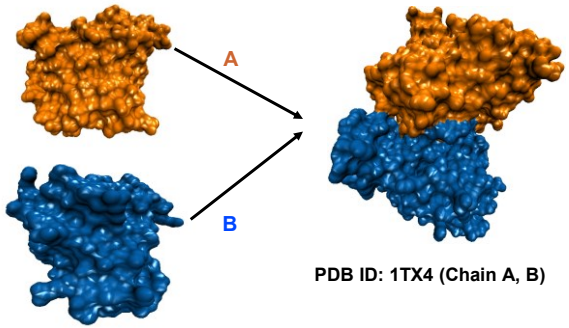


Q61L vs. WT: Fewer structures left in C2, higher-energy ones in C3, most structures in C1. Results indicate Q61L mutant increases rigidity of native state, possibly affecting transition of RAS between its on and off states.

Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Three Interesting Problems on Protein Modeling
  - *De novo* Protein Structure Prediction
  - Protein-Protein Assembly
  - Protein Energy Landscape Mapping
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
  - Protein-Protein Docking/Assembly
- Conclusions, Discussion, & Questions

Goal of Protein-Protein Docking



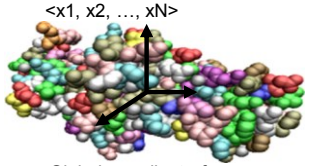
PDB ID: 1TX4 (Chain A, B)

- **Given the 3D structures of unbound protein chains.**
- **Predict the native-like low-energy bound configuration formed by the interactions of these chains.**

Challenges in Protein-Protein Docking

Unit A

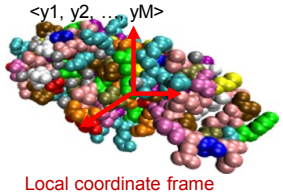
Cartesian-based representation:
<x1, x2, …, xN>



Global coordinate frame
<0, 0, 0, **i**, **j**, **k**>

Unit B

<y1, y2, …, yM>



Local coordinate frame
<x0, y0, z0, **u**, **v**, **w**>

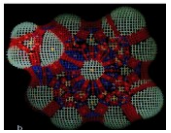
For each configuration of unit A, there are many possibilities for configurations for unit B.

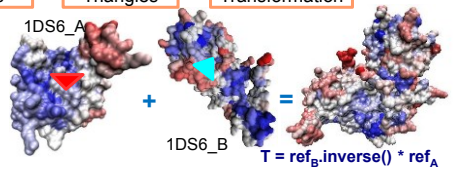
For rigid docking: space is SE(3)

Domain-specific Insight to Reduce Search Space for Rigid Protein-Protein Docking

- Limit to rigid-body transformations that align geometrically-complementary regions on molecular surfaces of units
- Origin of idea from geometric hashing in vision
- Applicability founded on mechanistic view of molecular interactions

Connolly Surface → Critical Points → Surface Triangles → Rigid Body Transformation





Discrete representation of molecular surface through notion of critical points

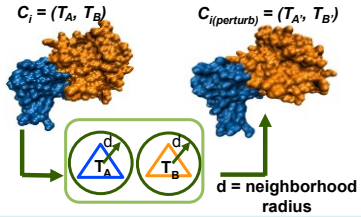
Triangles defined over surface to represent regions of interest for aligning

Simple way of defining transformations in SE(3) that align two complementary triangles

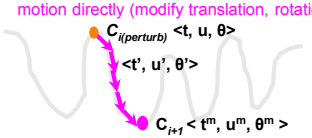
BH for Rigid Protein-Protein Docking

- Further domain-specific insight: limit search space to transformations that match only evolutionary-conserved regions (putative interaction surfaces) [Hashmi et al. CSBW 2011]
- Incorporated in 1+1 HEA (BH) [Hashmi et al. Proteome Science 2013]

Perturbation: search in neighborhood of active triangles for nearby rigid-body motion resulting in similar interface



Minimization: sample in vicinity of rigid body motion directly (modify translation, rotation)



BH for Rigid Protein-Protein Docking is Effective

Lowest lRMSD to native dimeric structure (Å)

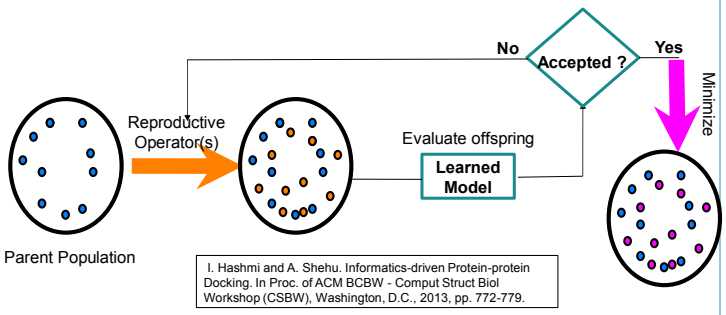
PDB ID (Chains)	Size (nr. atoms)	Inbar et al. [1]	Hashmi et al. [2]	Hashmi et al. [3]
1C1Y (A, B)	1376, 658	1.2	1.3	1.8
1DS6 (A, B)	1413, 1426	1.2	1.8	3.4
1TX4 (A, B)	1579, 1378	1.4	2.4	2.4
1WWW (W, Y)	862, 782	11.4	2.2	2.6
1FLT (V, Y)	770, 758	1.5	1.1	2.7
1IKN (A, C)	2262, 916	1.2	2.0	2.1
1IKN (C, D)	916, 1589	2.0	2.0	4.1
1VCB (A, B)	755, 692	0.7	2.1	3.4
1VCB (B, C)	692, 1154	1.7	1.3	2.7
1OHZ (A, B)*	1027, 416	1.8	1.7	2.7
1T6G (A, C)*	2628, 1394	1.6	2.5	3.6
1ZHI (A, B)*	1597, 1036	25.3	1.7	4.6
2HQS (A, C)*	3127, 856	29.1	2.2	2.6
1QAV (A, B)	663, 840	1.44	1.0	2.6
1G4Y (B, R)	682, 1156	0.8	2.3	4.1
1CSE (E, I)	1920, 522	0.7	1.5	2.4
1G4U (R, S)	1398, 2790	1.0	2.2	3.2

*CAPRI targets

[3] I. Hashmi and A. Shehu. A Probabilistic Search Algorithm for Decoy Sampling in Protein-protein Docking. Proteome Science 11(Suppl1):S6, 2013.
[2] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu. Guiding Protein Docking with Geometric and Evolutionary Information. J Bioinf and Comp Biol 10(3):1242002, 2012.
[1] Y. Inbar, H. Benyamini, R. Nussinov, and H. J. Wolfson. Prediction of multimolecular assemblies by multiple docking. J Mol Biol 349(2):435–447, 2005.

Room for Incorporating Learned Models

- Interaction energy is poor predictor of native interface
- Learned models can be incorporated to evaluate child configurations



I. Hashmi and A. Shehu. Informatics-driven Protein-protein Docking. In Proc. of ACM BCBW - Comput Struct Biol Workshop (CSBW), Washington, D.C., 2013, pp. 772-779.

Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Three Interesting Problems on Protein Modeling
  - *De novo* Protein Structure Prediction
  - Protein-Protein Assembly
  - Protein Energy Landscape Mapping
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
  - Protein-Protein Docking/Assembly
- Conclusions, Discussion, & Questions

Exciting Directions For EC Researchers

- Big, bigger, biggest in de novo structure prediction
  - Key work on representations to handle large proteins
- Extending protein-protein docking algorithms
  - Handle multimeric setting to compute structures of multimeric assemblies
- Mapping of energy landscapes to move beyond single-basin view
  - Key to detailed understanding of sequence-structure-function relationship
- Algorithmic design – effective and efficient
  - Local improvement, reproductive operators, single- vs. multi-objective evaluation, global vs. local selection, data-driven structurizations, cellular vs. spatial, island models, co-evolutionary models, and more
- Room for injecting ideas from machine learning, computational (statistical) physics to obtain more powerful EA frameworks

Further Reading: Reviews and Web

- A. Shehu. Probabilistic Search and Optimization for Protein Energy Landscapes. In Handbook of Computational Molecular Biology (Chapman & Hall/CRC Computer & Information Science Series), second edition, (Editors: Srinivas Aluru and Mona Singh), 2013.
- B. Olson. Evolving Local Minima in the Protein Energy Surface. Ph.D. Thesis, George Mason University, July, 2013.
- B. Olson, I. Hashmi, K. Molloy, and A. Shehu. Basin Hopping as a General and Versatile Optimization Framework for the Characterization of Biological Macromolecules. Advances in Artificial Intelligence J, 674832, 2012.
- K. A De Jong. Evolutionary Computation: A Unified Approach. MIT Press, 2006.
- Internet: http://www.cs.gmu.edu/~ashehu
  http://www.cs.gmu.edu/~kdejong

Further Reading: References in This Tutorial

On Energy Landscape View

C. B. Anfinsen. Principles that govern the folding of protein chains. Science 181(4096): 223-230, 1973.
K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. Nat Struct Biol 4(1): 10-19, 1997.
J. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 48:545-600, 1997.

On Fragment Replacement and Fragment libraries in Rosetta

P. Bradley, K. M. S. Misura, and D. Baker, Toward high-resolution de novo structure prediction for small proteins. Science 309(5742):1868–1871, 2005.

On Basin Hopping (design of variation and improvement operators) for De Novo Protein Structure Prediction and protein-protein docking

B. Olson and A. Shehu. Rapid sampling of local Minima in protein energy surface and effective reduction through a multi-objective filter. Proteome Science 11(Suppl1):S12, 2013.
I. Hashmi and A. Shehu. HopDock: A probabilistic search algorithm for decoy sampling in protein-protein docking.
B. Olson and A. Shehu. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface. Proteome Science 10(Suppl1):S5, 2012.

On Baseline EA for De Novo Protein Structure Prediction

S. Saleh, B. Olson, and A. Shehu. A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction. BMC Structural Biology J 13(Suppl1):S4, 2013.

On Crossover Design in a GA for De Novo Protein Structure Prediction

B. Olson, K. A. De Jong, and A. Shehu. Off-lattice protein structure prediction with homologous crossover. In Proc. of GECCO 2013, pp. 287-294.

On MOEAs for De Novo Protein Structure Prediction

B. Olson and A. Shehu. Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction. In Proc. of BiCoB 2014.
B. Olson and A. Shehu. Multi-objective stochastic search for sampling local minima in the protein energy surface. In Proc. of ACM BCB 2014, pp. 430-439.

How to Get Started

- **HEA-PSP: A Hybrid Evolutionary Search Framework with Various Crossover Implementations for Ab-initio Protein Structure Prediction**
- An executable for linux is available at: http://www.cs.gmu.edu/~ashehu/?q=OurTools.
- Operates over Rosetta, which can be downloaded from the Baker lab under an academic license.
- All algorithms presented today can be run through the use of configuration files, as described in above link.