GECCO 2015 Tutorial

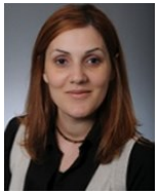
Evolutionary Algorithms for Protein Structure Modeling

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-2015/

GECCO'15 Companion, July 11–15, 2015, Madrid, Spain.
ACM 978-1-4503-3488-4/15/07.
http://dx.doi.org/10.1145/2739482.2756569

Brief Biosketch of Tutorial's Organizers



- **Amarda Shehu** is an Associate 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.

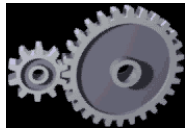
- **Kenneth De Jong** is a University Professor at George Mason University. He is a senior and well-known researcher in the EC community with a rich and diverse research profile. De Jong's research interests include genetic algorithms, evolutionary computation, machine learning, and adaptive systems. 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.

Computational Structural Biology: Three Fundamental Questions:

What does the part look like?

Mechanistic view == Shape/Form Governs Function

How do the parts move?



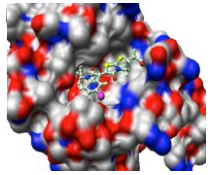
How do the parts come together?

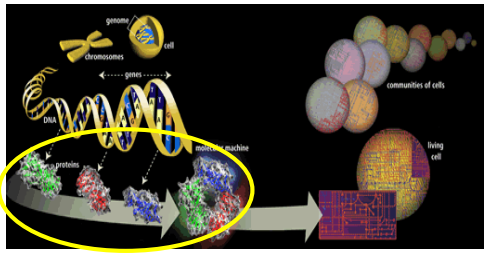
Mechanistic View: Form Governs Function

"Mechanistic View"

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

Protein molecule bound to small molecule





Source: Wikipedia

Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Two Problems on Protein Modeling
  - *De novo* Structure Prediction for Proteins with a Single Functional State
  - Mapping Energy Landscapes of Proteins with Multiple Functional States
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
- Conclusions, Discussion, & Questions

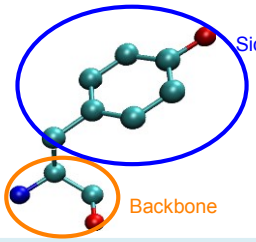
DNA

Amino Acid Sequence (AKA Protein)

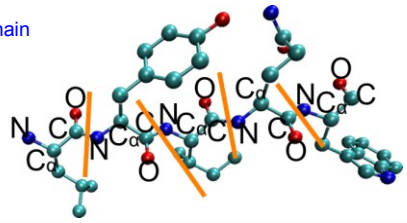
MKELVEMAVPENLVGAILGKGGKTLVEYQELTGARIQ
ISKKGEFLPGTRNRRVTITGSPAATQAAQYLISQRVT

Alphabet of 20 amino acids

Example of an Amino Acid



Amino acids link together to form a serial protein chain



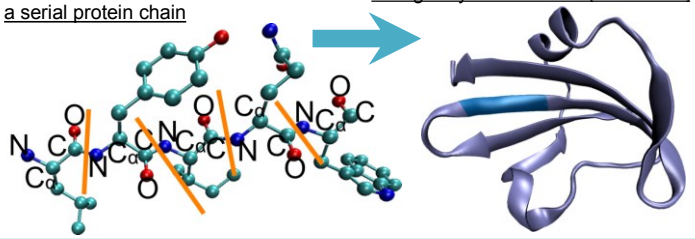
The Anfinsen Experiments: Protein Sequence Largely Determines Structure

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]

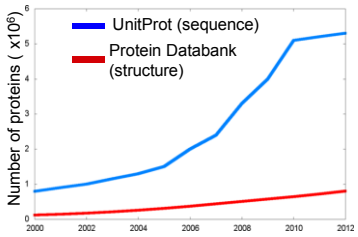


Agenda

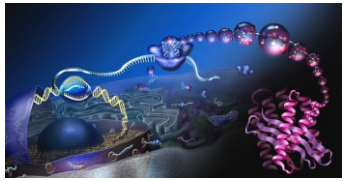
- Introduction: Proteins and their Role in Our Biology
- Overview of Two Interesting Problems on Protein Modeling
  - *De novo* Structure Prediction for Proteins with a Single Functional State
  - Mapping Energy Landscapes of Proteins with Multiple Functional States
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
- Conclusions, Discussion, & Questions

De Novo Protein Structure Prediction (PSP)

Need to model structure!



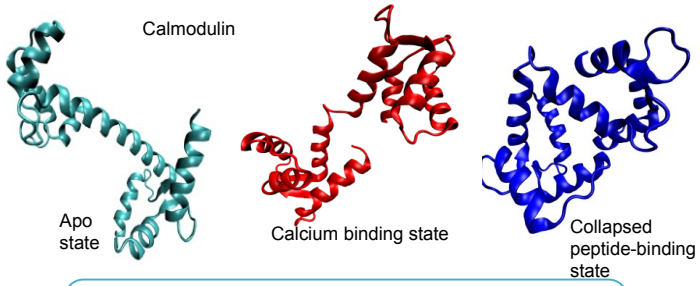
Protein chain folds as it comes out the ribosome translational machinery



Source: Science Art

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

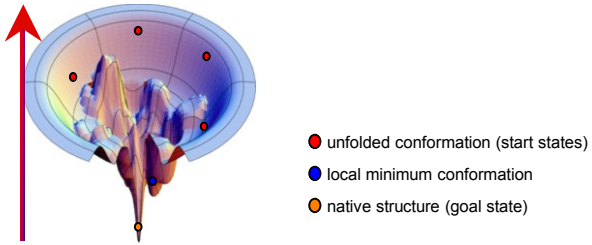
Structure Modeling for Proteins with Multiple Functional States



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

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



Implications of Energy Landscape View

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

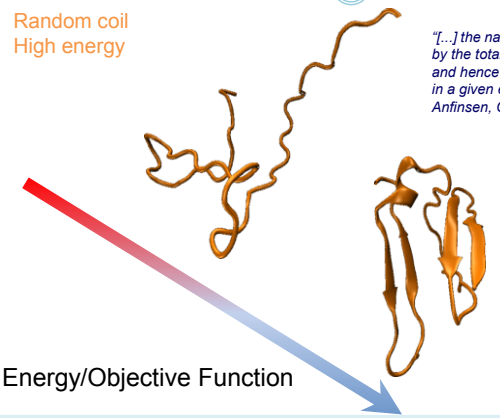
Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Two Interesting Problems on Protein Modeling
  - *De novo* Structure Prediction for Proteins with a Single Functional State
  - Mapping Energy Landscapes of Proteins with Multiple Functional States
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
- Conclusions, Discussion, & Questions

PSP 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

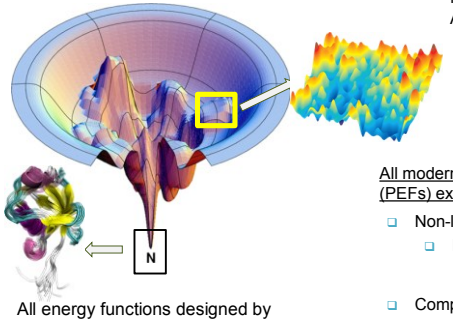
Energy/Objective Function

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

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:

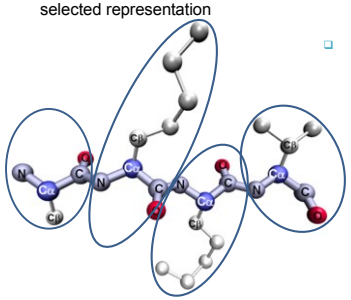
$$E = \alpha_1 \cdot E_{\text{Lennard-Jones}} + \alpha_2 \cdot E_{\text{H-Bond}} + \alpha_3 \cdot E_{\text{burial}} + \alpha_4 \cdot E_{\text{water}} + \alpha_5 \cdot E_{\text{Rg}}$$

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

No Ideal Representations, but Some are Better than Others

- Conformation = instantiation over selected representation



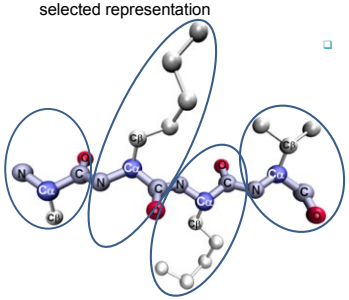
- Amino acid building blocks in protein chain are highly coupled!
- They impose spatial constrains on one another

Choice #1 for representation

- Cartesian coordinate-based representation
  - N atoms ➔ 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*

No Ideal Representations, but Some are Better than Others

- Conformation = instantiation over selected representation



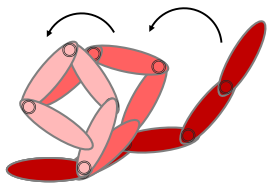
- Amino acid building blocks in protein chain are highly coupled!
- They impose spatial constrains on one another

Choice #2 for representation

- Collective-variable representation
  - N atoms ➔ K << 3N collective variables
  - Statistical analysis/dimensionality reduction of known (experimental) structures
    - Example: Principal Component Analysis
    - PCs = collective variables
  - Satisfies implicit constraints to extent satisfied in input structures
    - *Atoms "move together"*
    - *Opens opportunities for interesting EAs in new **reduced** search space*

No Ideal Representations, but Some are Better than Others

- Conformation = instantiation over selected representation

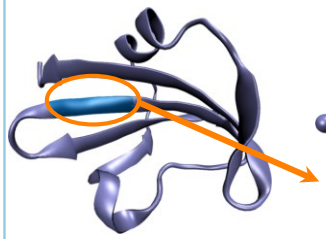


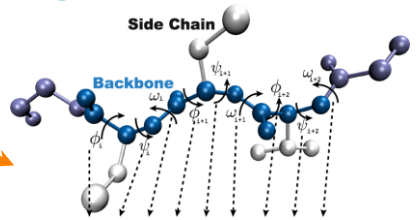
- Angular-based representations save in dimensions and are amenable to interesting perturbations
- Moreover, only some angles need to be modeled as variables (others, such as valence angles, can be ignored for structure prediction)

Choice #3 for representation

- Angular-based representation
  - N atom ➔ ~3N/7 parameters
  - Savings in dimensionality
    - Still hundreds of dimensions
    - Can be further reduced (molecular fragment replacement)
  - 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)

Angular-based Representation: State of the Art in *PSP*





- 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

PSP via Stochastic Optimization

- Conformational space: high-dimensional and continuous
- Fitness/energy surface/landscape: non-linear and multimodal
- 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

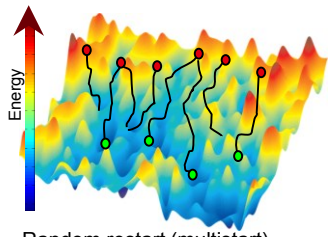
EAs for PSP in EC Community

Memetic/Hybrid Evolutionary Algorithms (HEAs)	Multi-Objective Evolutionary Algorithms (MOEAs)
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

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

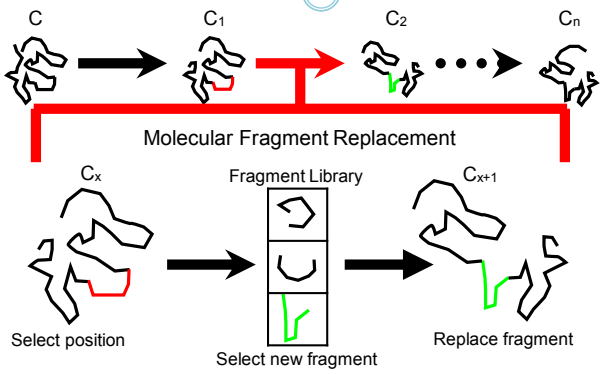


Random restart (multistart)

- 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 Two Interesting Problems on Protein Modeling
  - *De novo* Structure Prediction for Proteins with a Single Functional State
  - Mapping Energy Landscapes of Proteins with Multiple Functional States
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
- Conclusions, Discussion, & Questions

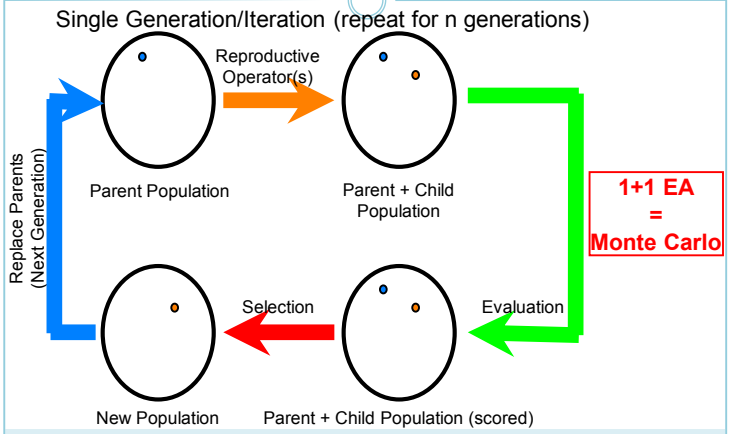
[25/52]

State-of-the-art EAs for PSP

- *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

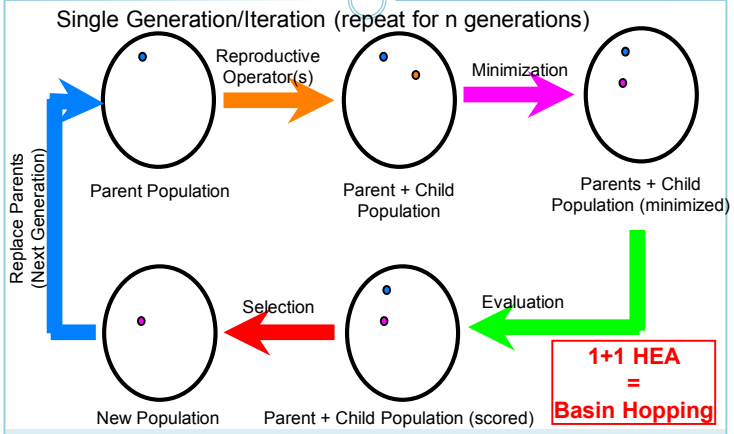
[26/52]

A Baseline EA for PSP



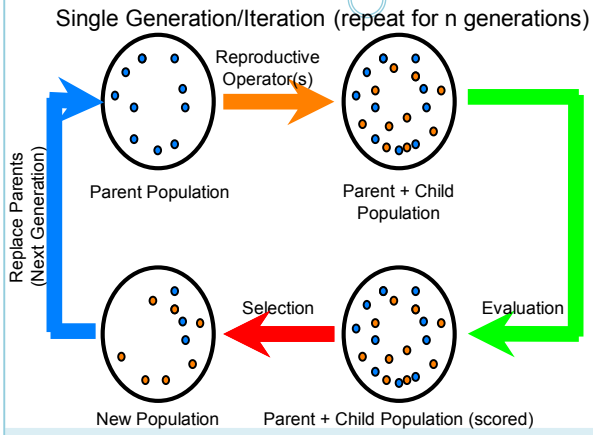
[27/52]

A Baseline HEA for PSP



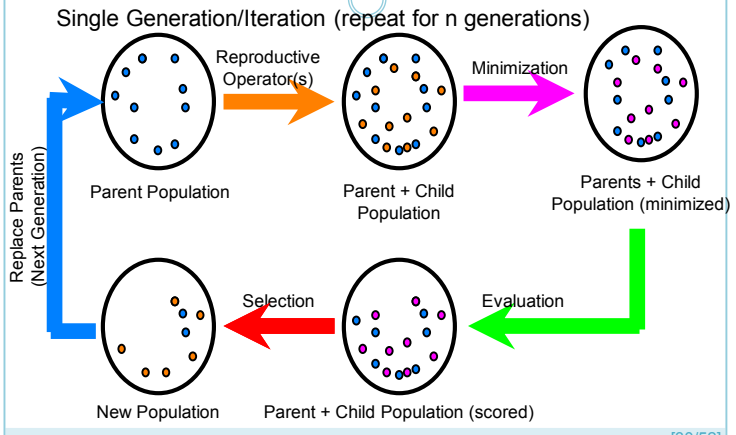
[28/52]

Population-based EAs for PSP



[29/52]

Population-based HEAs for PSP



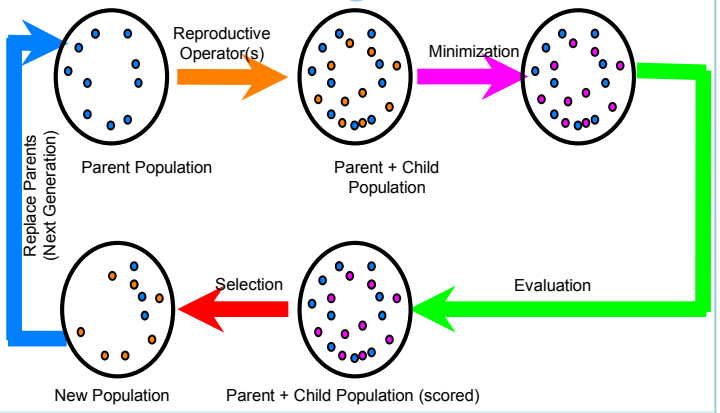
[30/52]

EAs with Domain-specific Insight for PSP

- 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
- Nr. of generations: capped by total nr. of energy evaluations
- Comparison setting:
  - Versus Rosetta (Monte Carlo-based)
  - Explored issues: magnitude of perturbation operator, cost of improvement operator

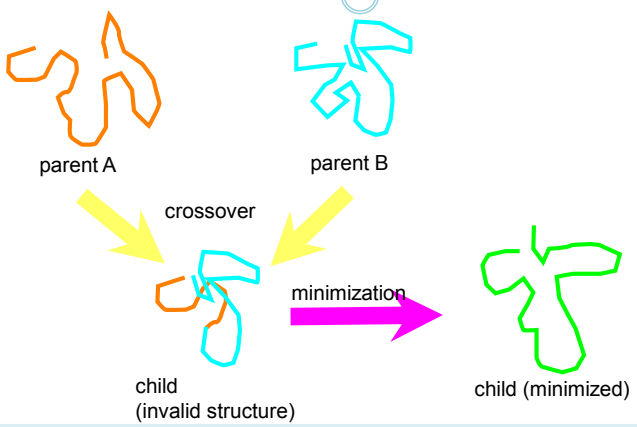
[31/52]

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



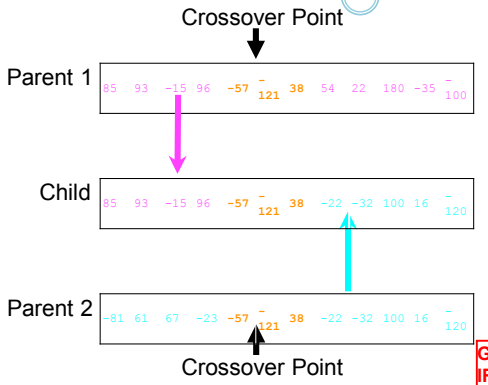
[32/52]

Hybrid Genetic Algorithm (GA) for PSP: HEA + Crossover



[33/52]

Homologous Crossover for Protein Structures



Mean reduction in conformational sampling due to crossover:

1 point	11%
2 point	21%
Homologous	7%

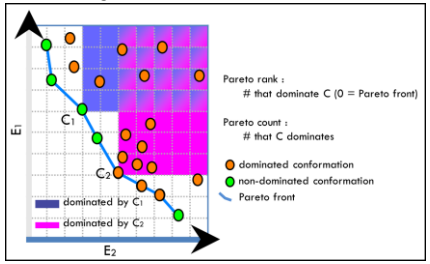
GA outperforms in lowest IRMSD Monte Carlo techniques like FixIt
(Olson et al. GECCO 2013)

[34/52]

Multi-objective EAs for Handling Imperfections in Energy Function

Evaluation on multiple objectives rather then on aggregate energy score

Multi-objective



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

E_1 E_2

Sorting by Pareto rank, Pareto count, and total energy in this order

Truncation-based selection

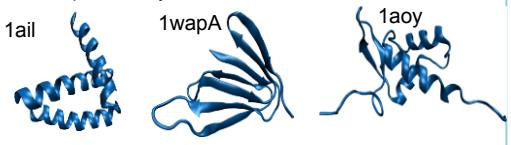
[Olson et al. BiCoB 2014 and ACM BCB 2013]

[35/52]

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

Experimentally-determined native structures



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

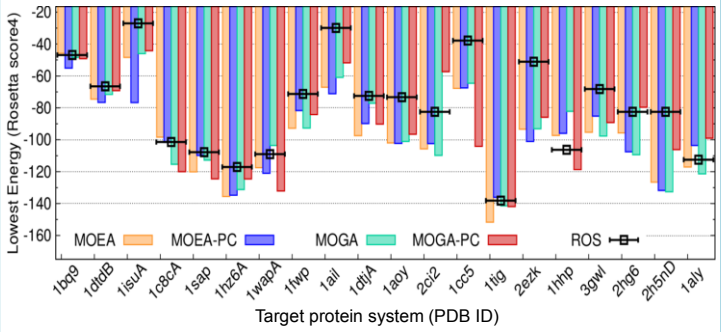
10 million energy function evaluations 2.4 GHz Core i7
12-24 hours of CPU user time (depending on protein length)
Repeated five times (total 50 million energy evaluations)

Comparison with Rosetta on several metrics (including lowest lRMSD)

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

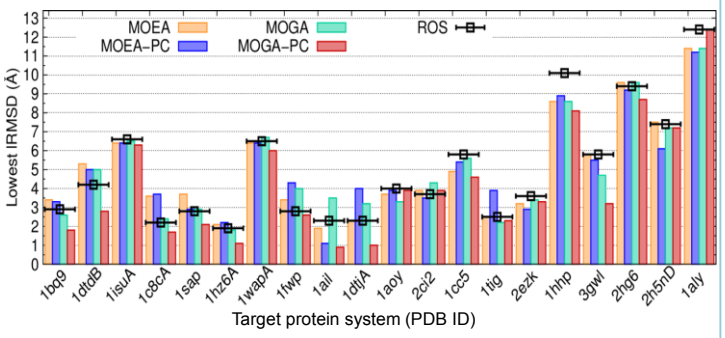
[36/52]

Enhanced Exploration Capability over Monte Carlo-based Algorithms



Olson et al. BiCoB 2014

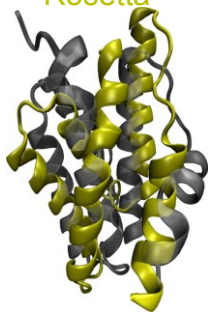
Higher Accuracy over Monte Carlo-based Algorithms



Olson et al. BiCoB 2014

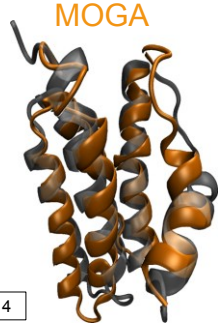
Better Models Over State of the Art

Rosetta



PDB ID
3gwl

MOGA



Olson et al. BiCoB 2014

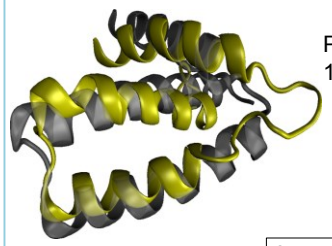
5.8 Å RMSD to native

3.2 Å RMSD to native

Compared to experimentally-determined native structure (in gray)

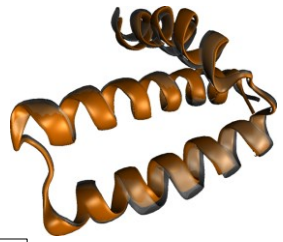
Better Models Over State of the Art

Rosetta



PDB ID
1dtjA

MOGA



Olson et al. BiCoB 2014

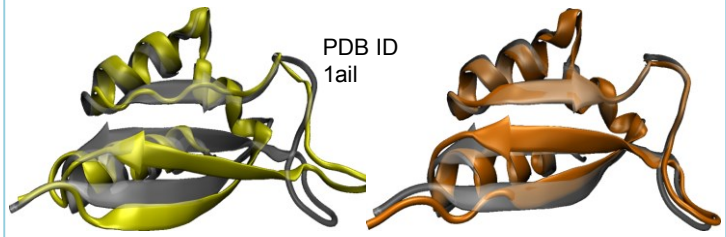
4.5 Å RMSD to native

0.9 Å RMSD to native

Compared to experimentally-determined native structure (in gray)

Better Models Over State of the Art

Rosetta MOGA



PDB ID 1ail

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)

[41/52]

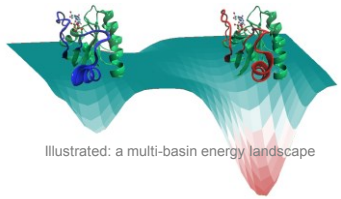
Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Two Interesting Problems on Protein Modeling
  - *De novo* Structure Prediction for Proteins with a Single Functional State
  - Mapping Energy Landscapes of Proteins with Multiple Functional States
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
- Conclusions, Discussion, & Questions

[42/52]

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*



Illustrated: a multi-basin energy landscape

- 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

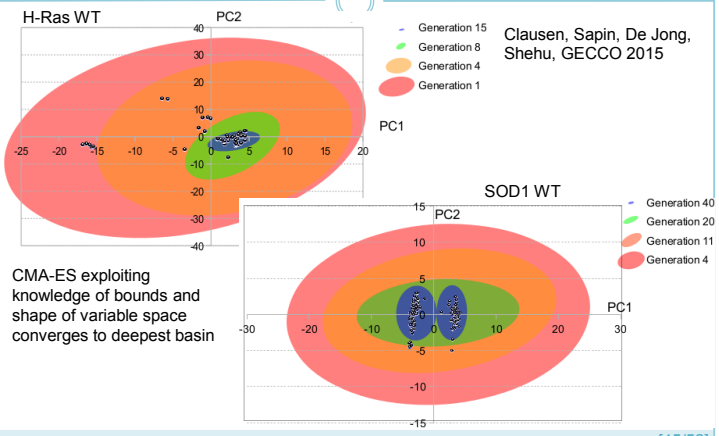
[43/52]

EAs for Modeling Multi-Basin Proteins: Energy Landscape Mapping

- *De novo* setting computationally impractical
  - Direction: exploit domain-specific knowledge (example: experimental structures)
- Idea #1: directly in initial population
  - Other algorithmic components as in EAs for PSP
- Idea #2: indirectly to define effective phenotypic representation
  - Discrete to continuous mapping (experimental structures to collective variables through PCA) [Clausen et a. 2014-2015]
- Other evolutionary strategies
  - CMA-ES vs. customized EAs

[44/52]

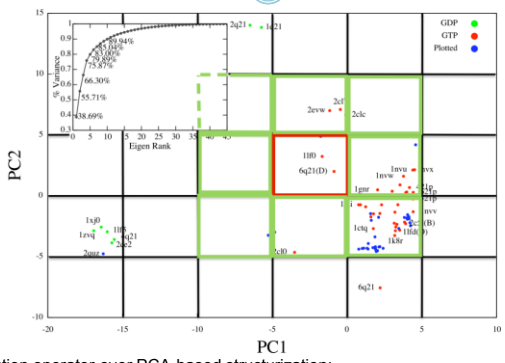
CMA-ES: Analysis of Convergence in Multimodal Landscapes



Clausen, Sapin, De Jong, Shehu, GECCO 2015

CMA-ES exploiting knowledge of bounds and shape of variable space converges to deepest basin

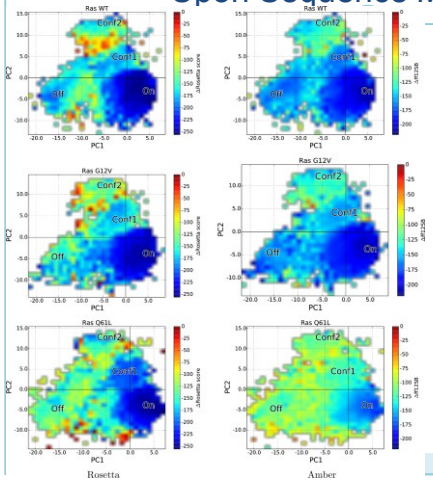
Customized EA (SIfTer): PCA-based Representation and Local Selection



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, Shehu
J Comput Biol 2015

SIfTer: Ability to Explain Changes to Landscape Upon Sequence Mutations



Detailed analysis of application on a conformational switching onco-protein, catalytic domain of H-Ras

Clausen et al.
PLoS Comput Biol 2015

Comparison of energy landscapes obtained for wildtype and variant sequences of H-Ras indicate mutations change landscape

thus affecting transition of catalytic domain of H-RAS between its on and off functional states.

Agenda

- Introduction: Proteins and their Role in Our Biology
- Overview of Two Interesting Problems on Protein Modeling
  - *De novo* Structure Prediction for Proteins with a Single Functional State
  - Mapping Energy Landscapes of Proteins with Multiple Functional States
- Corresponding Optimization Problems and Computational Challenges
- Overview of Corresponding State-of-the-art EAs for
  - *De novo* Protein Structure Prediction
  - Protein Energy Landscape Mapping
- Conclusions, Discussion, & Questions

Directions For EC Researchers

- Low-dimensional, constraint-satisfying representations
- Adaptations, enhancements for protein-protein docking
  - Handle structure modeling of multimeric protein 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

- Algorithms (PSP-EA, SIfTER) available as executables for linux at: http://www.cs.gmu.edu/~ashehu/?q=OurTools.
- Operate 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.

Acknowledgment of funding:
National Science Foundation Grant No. 1421001 and No. 1144106 and the Jeffress Trust Program in Interdisciplinary Research Award