Mathematical Modeling and Computational Physics
(MMCP2017)

Russia, Moscow region, Dubna, JINR, LIT
3-7 July 2017

Parallel evolutionary optimization algorithms for peptide-protein docking

S.V. Poluyan
svpoluyan@gmail.com

Dubna State University

Institute of System Analysis and Control
Postgraduate student

N.M. Ershov
ershovnm@gmail.com

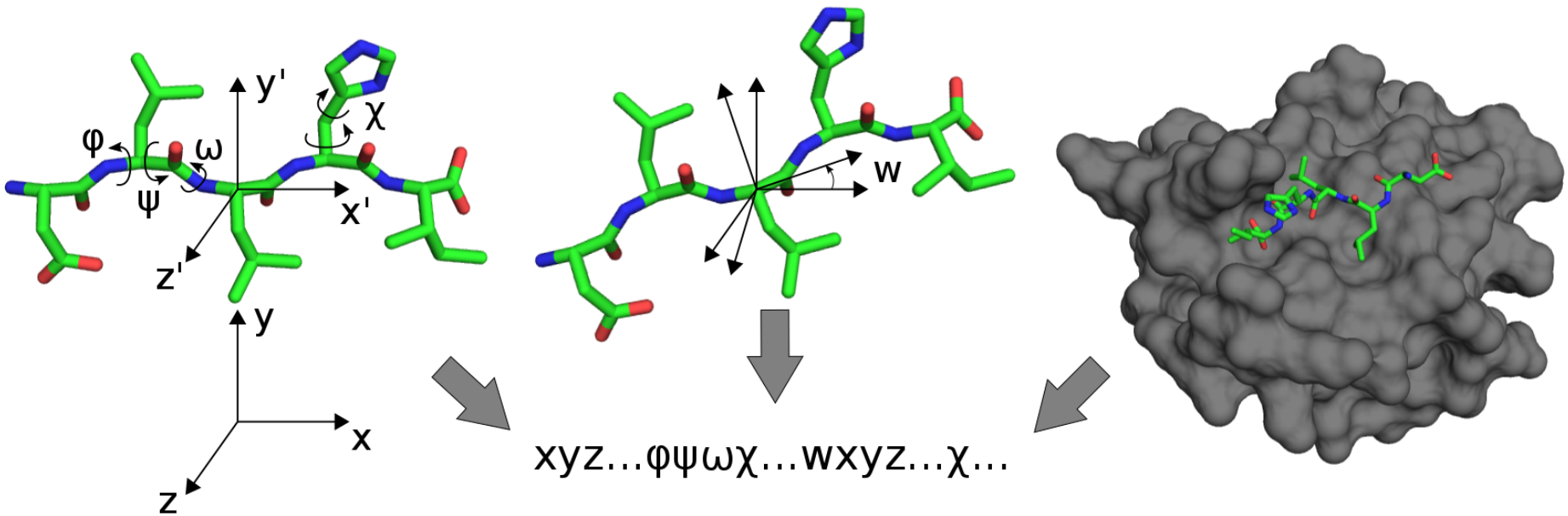
Lomonosov Moscow State University

Faculty of Computational Mathematics and Cybernetics

Protein-peptide docking

Protein and peptide structure:

- linear sequence of amino acids linked by peptide bonds;
- 20 standard amino acids.



Search space (real-coded variables):

- peptide backbone torsion angles: ϕ, $\psi \in [-\pi, \pi]$ and $\omega \in [\pi - \delta, \pi + \delta]$;
- protein and peptide side-chain torsion angles: χ_{1-4} $[-\pi, \pi]$;
- peptide translation and rotation.

Docking $\Rightarrow$ find the binding conformation with the lowest energy.

Rosetta all-atom force-field

Rosetta uses a mix of statistical and physical potentials:

- attractive and repulsive forces are modelled with the Lennard-Jones potential;
- Lazaridis–Karplus implicit solvation;
- Coulombic electrostatic potential with a distance-dependent dielectric.

Hydrogen bond terms

- backbone-backbone hbonds close in primary sequence;
- backbone-backbone hbonds distant in primary sequence;
- sidechain-backbone and sidechain-sidechain hydrogen bond energy.

Knowledge-based terms:

- Ramachandran preferences and probability of amino acid at ϕ/ψ;
- ω dihedral in the backbone (harmonic constraint on planarity);
- internal energy of sidechain derived from Dunbrack's statistics.

$$Energy = w_1 \cdot term_1 + w_2 \cdot term_2 + \dots$$

Weights on the score terms are calibrated. Energy score is not kcal/mol.

Using knowledge-based information

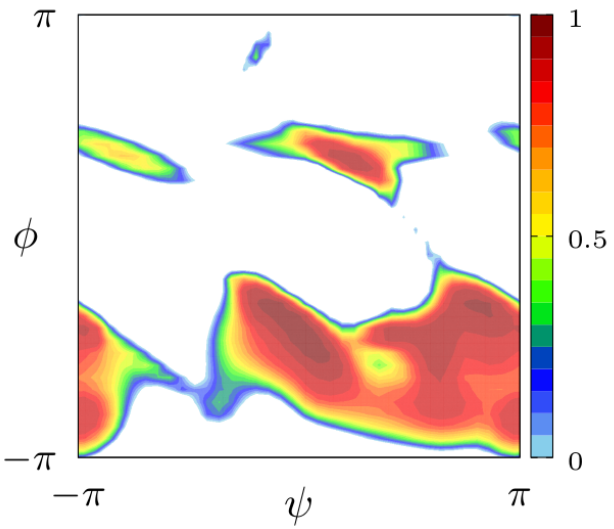
The main goal is to exclude impossible conformations.

Backbone torsion angles:

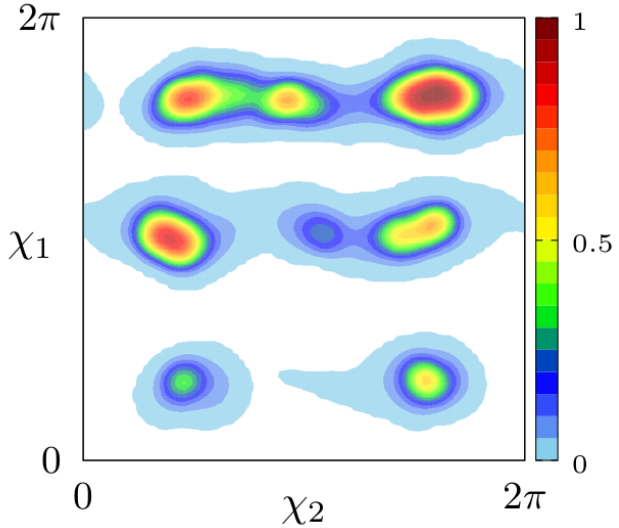
- ω angle in *trans*-state tends to be planar;
- Rosetta neighbor-dependent Ramachandran energy term.

Side-chain torsion angles:

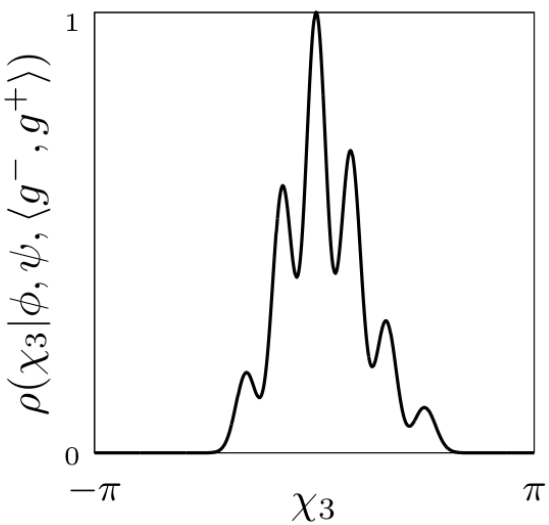
- centered on sp^3-sp^3 and sp^3-sp^2 hybridized bonds;
- peaks at approximately $60°$, $180°$, $300°$;



The neighbor-dependent Ramachandran probability distribution for asparagine.



The backbone-independent density of histidine side-chain torsion angles. (Top8000 data)



The backbone-dependent density for glutamine χ_3. (Dunbrack 2010 Library)

Global and local docking

Peptide structure:

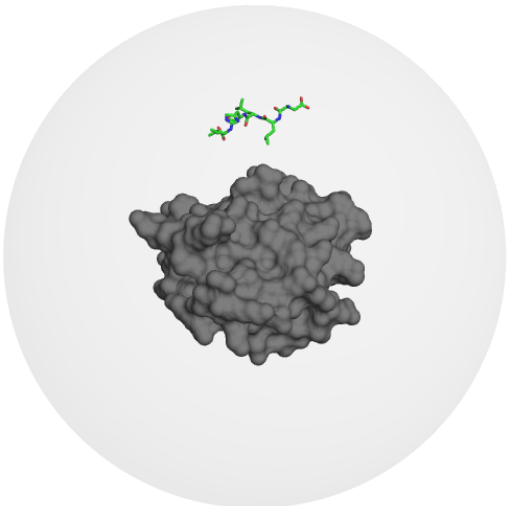
- **linear**;
- coil;
- α-helix;
- β-strand.

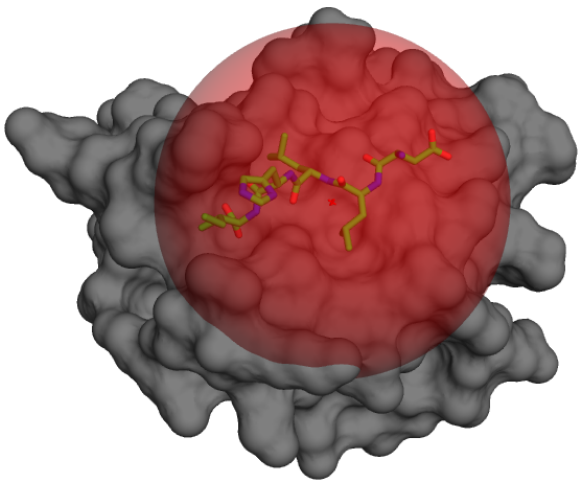
Protein interface:

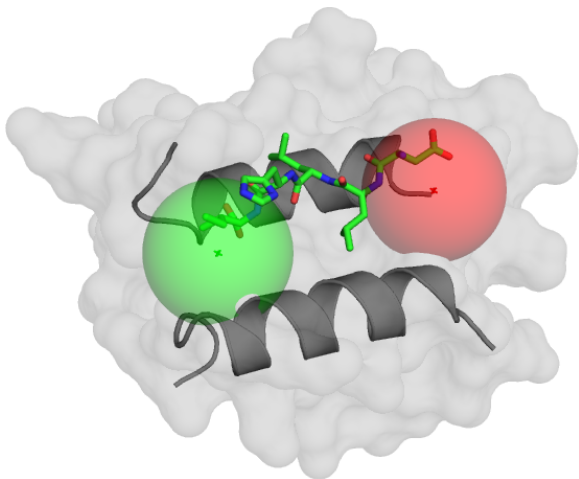
- β-sheet;
- 2-loop channel;
- **2-helix channel**;
- etc.

Approaches (Peptide translation vector – C_α atom from first residue):

- Global docking. Fully blind docking without prior knowledge of the binding site. All protein side-chain rotamers involved.
- Local docking at the binding site:
  - protein side-chain rotamers within local area;
  - peptide translation – one or two spheres.



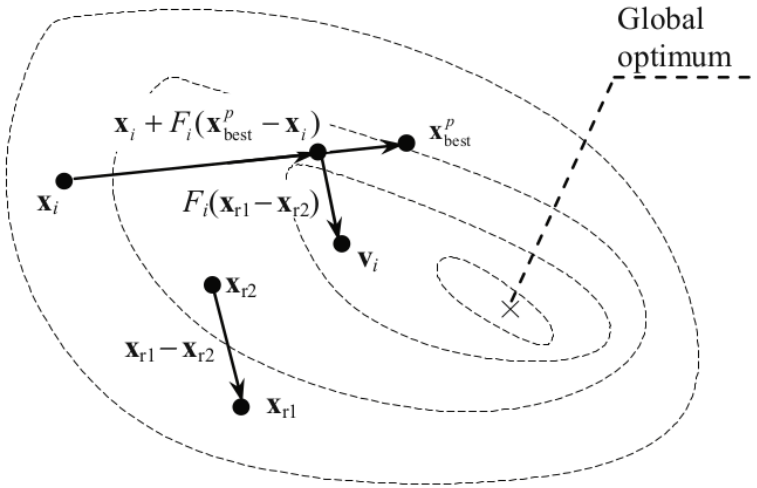




Selected algorithms

Evolutionary computation:

- stochastic optimization;
- heuristic algorithms;
- evolution of the population;
- easy to parallelize.



Evolutionary algorithms & strategies:

- JADE: Adaptive Differential Evolution With Optional External Archive;
- CSO: Competitive Swarm Optimizer.

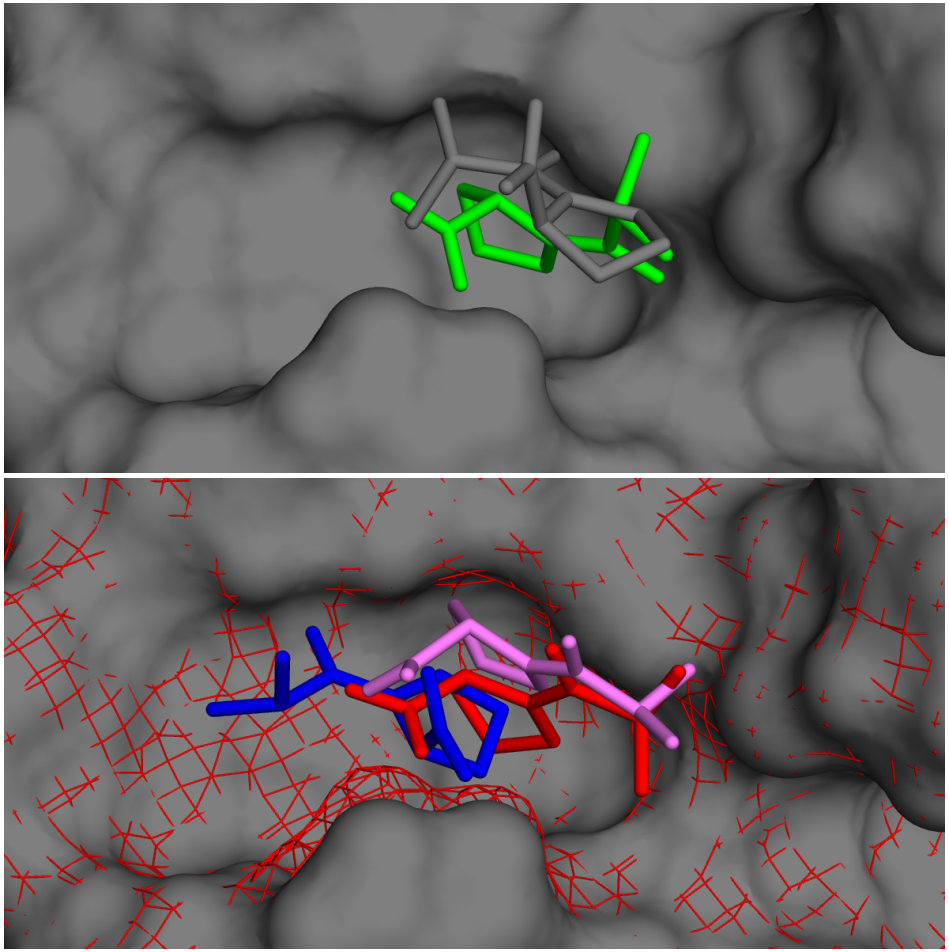
JADE operation:

- mutation: $v_i = x_i + F_i(x^p_{best} - x_i) + F_i(x_{r1} - \tilde{x}_{r2})$;
- crossover:

$$u_{i,j}(t) = \begin{cases} v_{j,i} & \text{if } \text{UniformRand}_j(0,1) \leqslant CR_i \text{ or } j = j_{rand} \\ x_{j,i} & \text{otherwise} \end{cases}$$

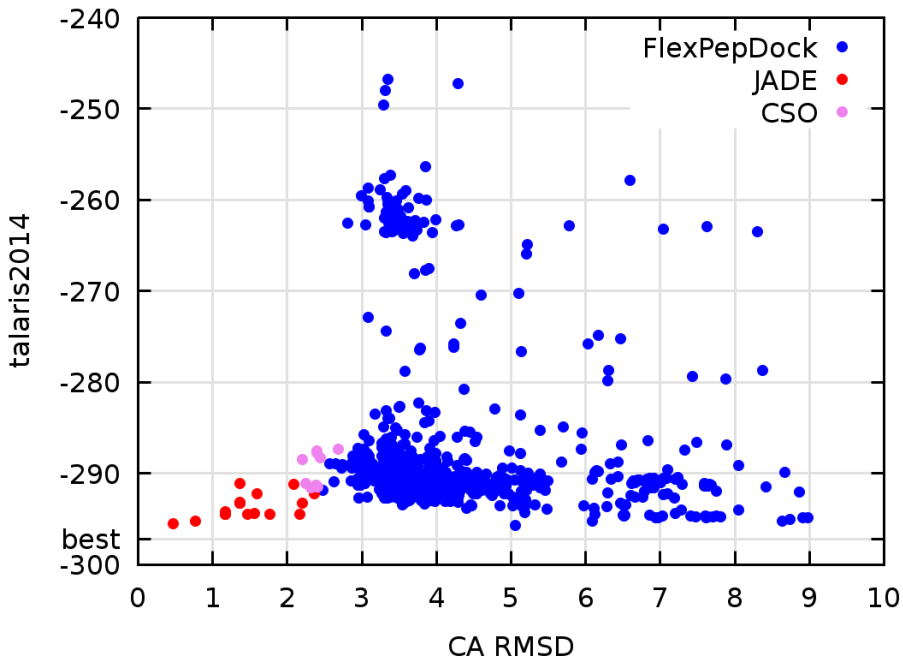
- selection;
- adaptation of F_i and CR_i.

Local docking

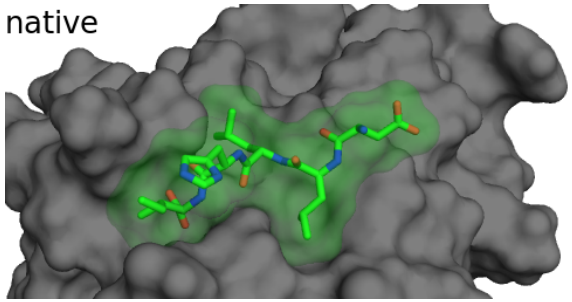


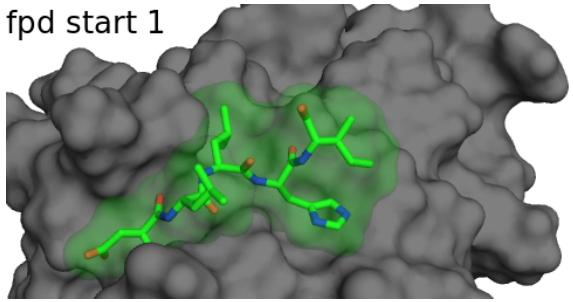
2CYH (PDB id):

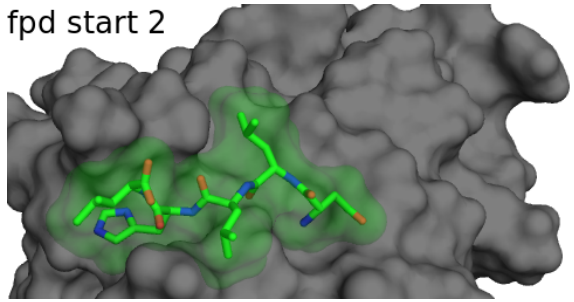
- peptide sequence: AP;
- sphere radius: 10 Å;
- dimension: 25.

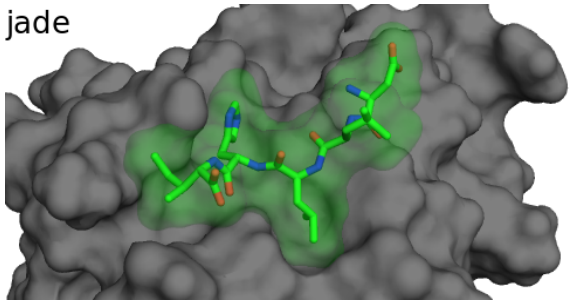


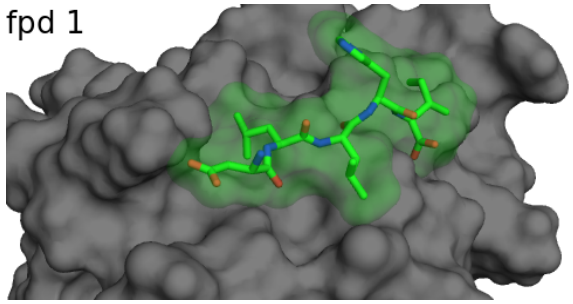
Local docking

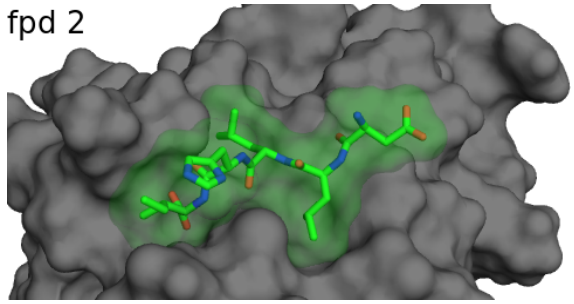






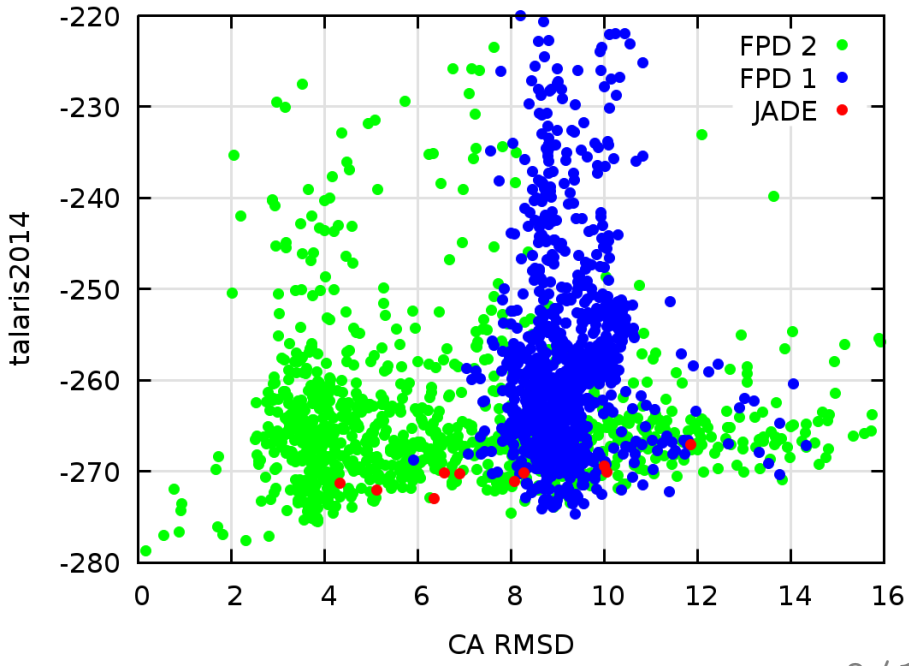




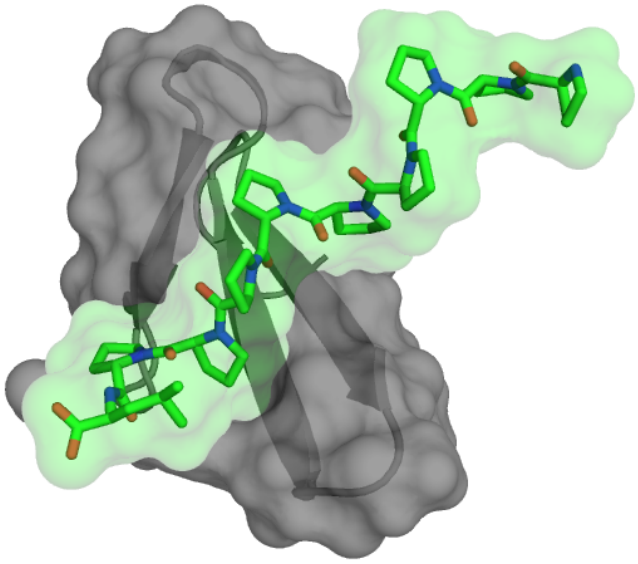


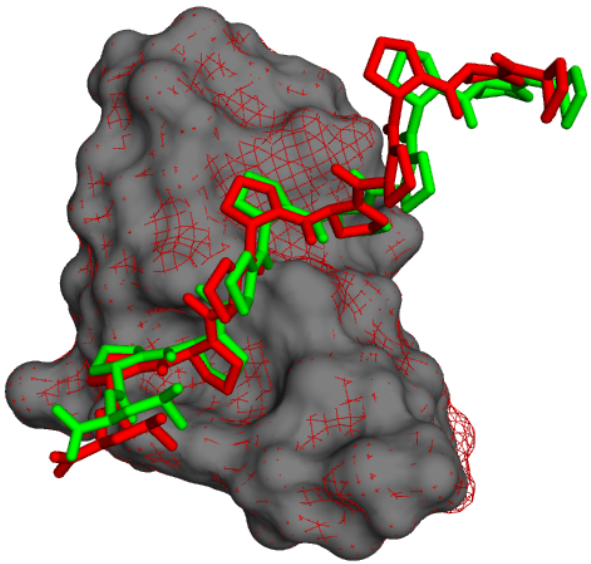
1JWG (PDB id):

- peptide sequence: DLLHI;
- spheres radius: 5 Å;
- dimension: 54.



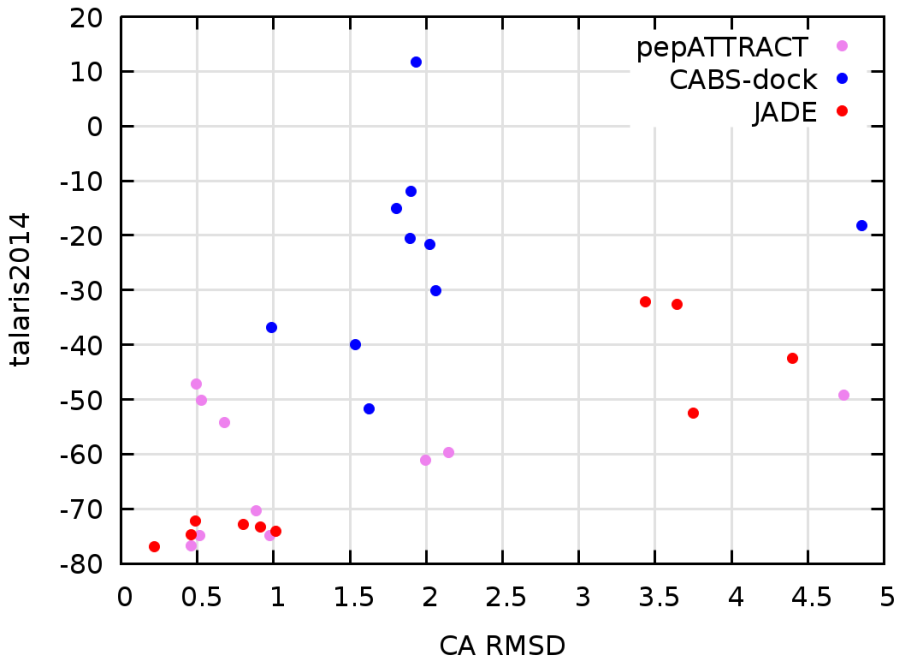
Global docking





2HO2 (PDB id):

- peptide sequence: PPPPPPPPPL;
- protein length – 33 residues;
- dimension: 93.



Parallel realization

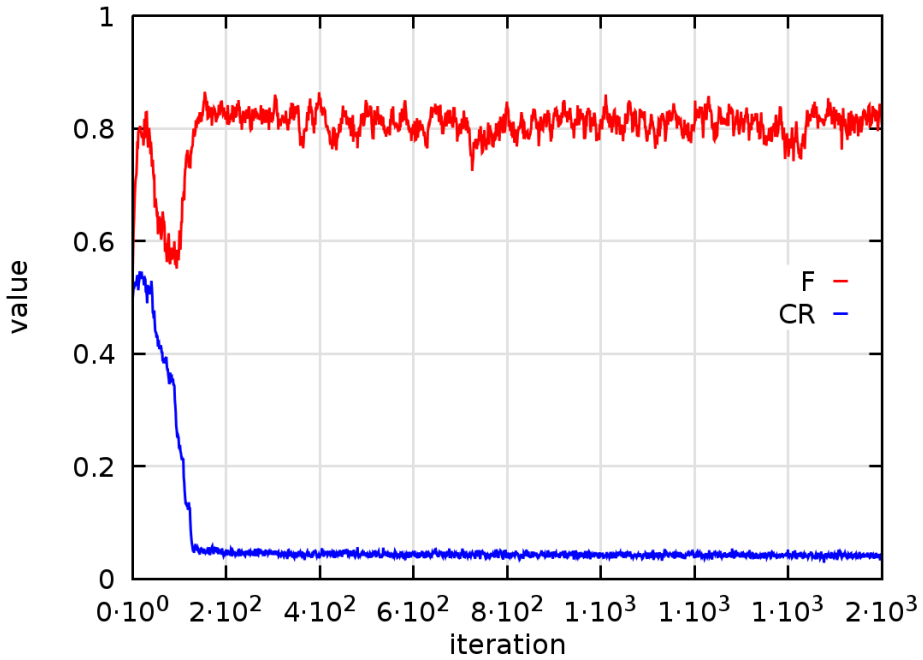
Implementation features:

- cumulative distribution functions (CDF) require a lot of memory;
- CDF worst-case complexity is $O(k \cdot \log_2 n)$,
  where n is step size for angle, k – dimension;
- both algorithms are implemented using C++ with MPI/OpenMP;
- benefits of using HybriLIT cluster (JINR) are presented in table.

Threads / Nodes	4 / 1	8 / 1	12 / 1	16 / 1	24 / 1	24 / 2	48 / 2
Acceleration	3.31	6.28	9.03	11.84	17.1	15.85	31.1
Efficiency	0.82	0.78	0.75	0.74	0.71	0.66	0.64

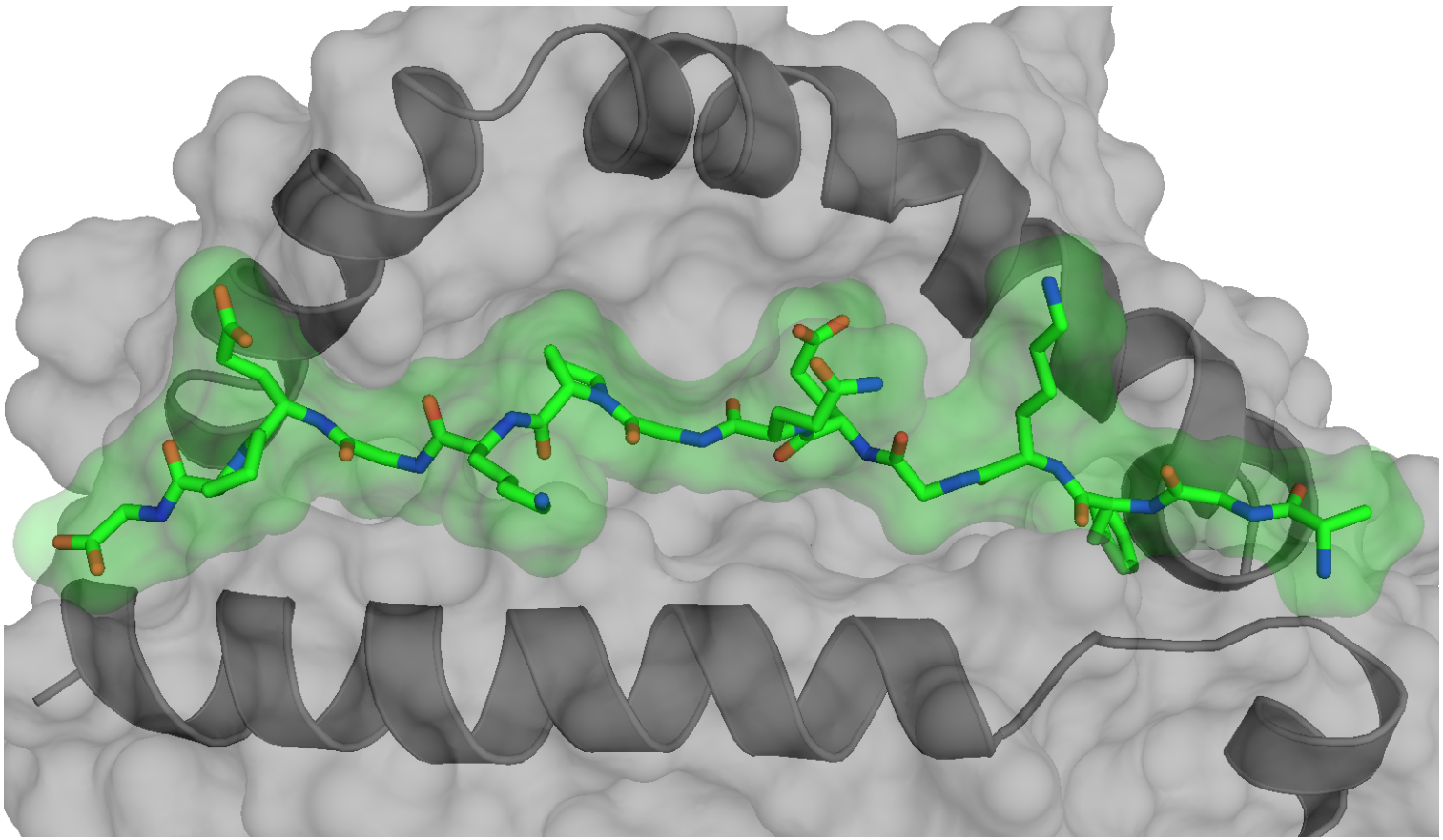
Conclusions

- Good results only for short peptides (2-3 residues) and dimension up to 30.
- Poor performance.
- Population degeneration.
- Low crossover probability – no global search;
- High mutation probability.
- Meta-optimization for F and CR or different adaptation scheme – same result.
- It is hard to modify algorithm to produce local search. For instance, with Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.



Future work

- There are lots of statistics.
- Different energy evaluation (multi-objective optimization).
- Estimation of distribution algorithms.
- Bayesian optimization.
- DR1-RA (PDB id 2FSE) complex with peptide AGFKGEQGPKGEPG.



Thank you!