Modeling Metal Protein Complexes from Experimental Extended X-ray Absorption Fine Structure using Evolutionary Algorithms

Collin Price   Sheridan Houghten   Sergey Vassiliev   Doug Bruce
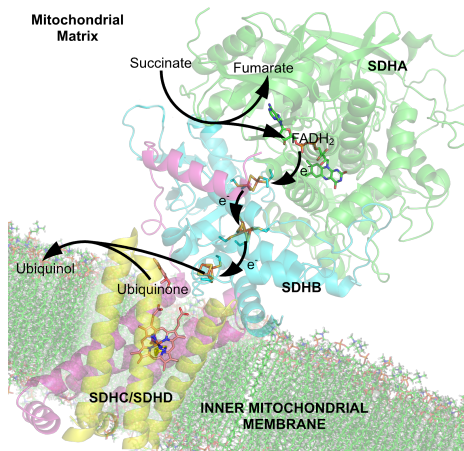
Brock University

October 3, 2014

Outline

Cofactor

- "Helper" molecules
- Assist in biochemical transformations
- Non-protein chemical compound
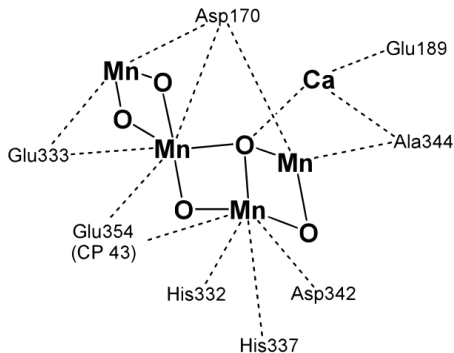- Exist within a protein

Figure : Cofactor Examples

Oxygen-Evolving Complex (OEC)

- The water-oxidizing enzyme of photosystem II.
- Water-splitting complex.
- Exists in 5 states: S_0 to S_4.
- Splits H_2O into hydrogen and oxygen.
- S_4 reacts with water to produce oxygen.
- Process is not completely understood.
- S_1 is examined in this work.

Figure : Metalloenzyme core of the OEC

Experimental Extended X-ray Absorption Fine Structure (EXAFS)

- A method used to measure the absorption coefficient of a material as a function of energy.
- X-ray is tuned to have the same wavelength as the target atom.
  - In the case of the OEC we targeted the four Manganese atoms.
- X-ray reacts with target atom causing the atom to lose an electron and interact with neighbouring atoms.
- Fluorescence energies are emitted and measured.

Experimental EXAFS Spectra

EXAFS Spectra in k space
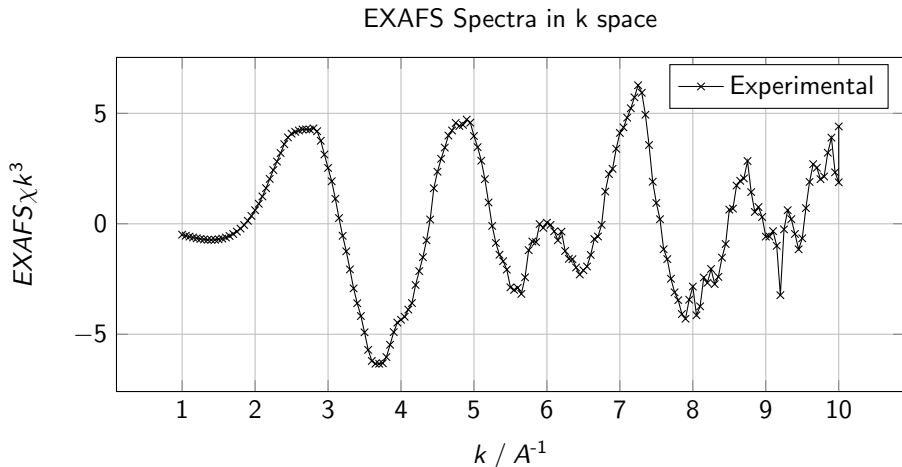
Figure : EXAFS Spectra of OEC in S_1

Structure Refinement Problem

- EXAFS spectra provides limited information about neighbouring atoms.
- We would like to know the 3-dimensional positions of each atom relative to each other.
- EXAFS spectra can be generated in simulations.
- If we match the calculated EXAFS spectra with the experimental EXAFS spectra it is likely we found the correct configuration.
- Goal is to minize the differences between spectra.

Example Spectra Comparison
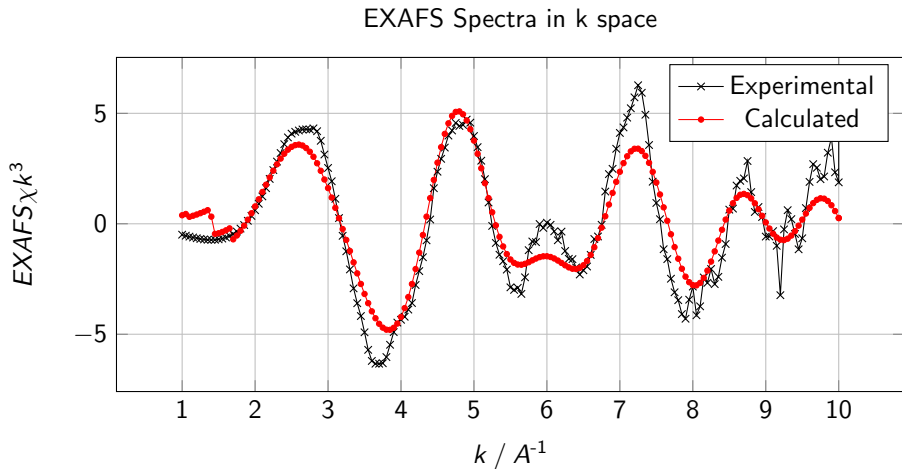
Figure : OEC EXAFS Spectra Comparison

Previous Work

- A model of the oxygen-evolving center of photosystem II predicted by structural refinement based on EXAFS simulations
    - *Sproviero, Eduardo M and Gascón, José A and McEvoy, James P and Brudvig, Gary W and Batista, Victor S*
- S_1-state model of the O_2-evolving complex of photosystem II
    - *Luber, Sandra and Rivalta, Ivan and Umena, Yasufumi and Kawakami, Keisuke and Shen, Jian-Ren and Kamiya, Nobuo and Brudvig, Gary W and Batista, Victor S*
- Density functional theory quantum mechanics and molecular mechanics hybrid (DFT-QM/MM)
- Monte Carlo method (R-QM/MM)
- Previous techniques would converge on local minima.

Problem Encoding

- Set of 3-dimensional coordinates in space, where each position is assigned to a specific atom in the atomic structure.
- Entire OEC atomic structure contains 1269 atoms.
- EXAFS calculations only require 79 specific atoms.
- The units of measurement for each atom position are measured in Angstroms (Å).
- Two representations used.

Problem Encoding

X	Y	Z
14.451	-13.346	1.133
15.336	-13.488	2.014
13.005	-13.364	1.452
0.019	0.011	0.045
...	...	...

Figure : Representation 1

X_0	14.451
Y_0	-13.346
Z_0	1.133
X_1	15.336
Y_1	-13.488
Z_1	2.014
X_2	13.005
...	...

Figure : Representation 2

Population Generation

Method 1

- Initial OEC atomic structure came from the crystallographic photosystem II structure in the Protein Data Bank.
- Randomly moving each atom produced too many unusable configurations.
- Needed to generate feasible structures for starting GA population.
- Used a molecular dynamics simulation to obtain these structures (NAMD).
- Produced 10 000 atomic structures.
- Evaluated each atomic structure and kept the top 300.

Population Generation

Method 2

- Initialize population based on seed atomic structure.
- Each atomic position within the atomic structure is randomly shifted.

14.451
-13.346
1.133
15.336
-13.488
2.014
13.005
...

$\Longrightarrow$

14.494
-13.370
1.132
15.295
-13.518
2.009
13.018
...

Figure : Randomly shifting atomic positions by 0.05Å

Operators

- Crossover operator
    - One-point crossover
    - Least destructive to the individuals
- Mutation operator
    - Single atomic coordinate is randomly moved my 0.05Åusing Euclidean distance.
    - Tested moving atomic coordinates by 0.001Å, 0.005Å, 0.01Å, 0.025Å, 0.05Å, 0.1Å, 0.5Å, 1Å, and 5Å.
    - 0.05Å was chosen because it produced at least a 1% change in fitness for all chemical elements.
- Selection operator
    - 3-tournament selection

Fitness

- The goal of the experiment is to find an atomic structure that generates the same EXAFS spectra as the experimental EXAFS spectra.
- FEFF6 is used to simulate an XAFS experiment.
- IFEFFIT does post processing of the simulated EXAFS spectra.
- IFEFFIT was developed at CARS, the Consortium for Advanced Radiation Sources at The University of Chicago.
- The root-mean-square deviations (RMSD) is used to calculate the difference between the experimental and calculated EXAFS spectra.

$$RMSD = \sqrt{\frac{\sum_{t=1}^{n} \left(x_{1,t} - x_{2,t}\right)^2}{n}} \tag{1}$$

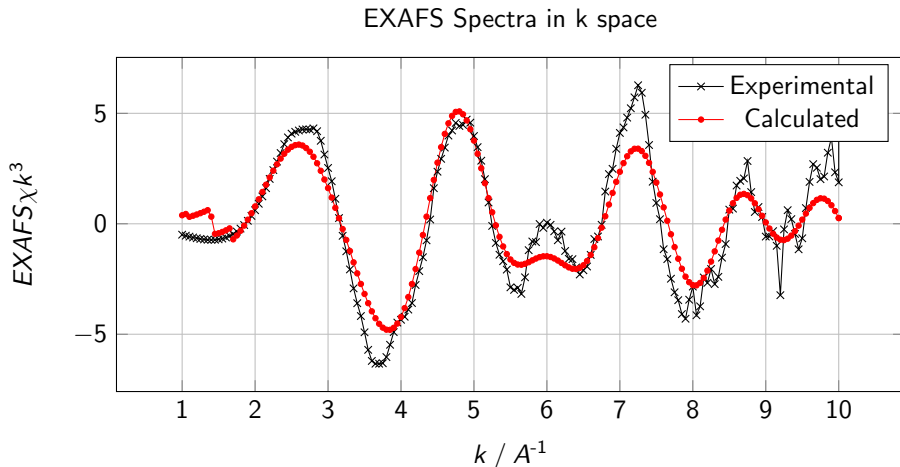# Example Spectra Comparison

Figure : OEC EXAFS Spectra Comparison

Genetic Algorithm (GA)

- Shown success in past studies in finding low-energy protein conformations.
- Higher potential of avoiding local minima.
- Algorithm is able to produce a population of candidate solutions for future examination.

Restarting Genetic Algorithm (RGA)

- Variation of the Recentering-Restarting Genetic Algorithm.
- Used to maintain diversity within the population.
- Genetic Algorithm is run until the population has reached a set convergence rate.
- Duplicate individuals are removed from the population and new individuals are injected into the population.

Algorithm 1 Restarting the population

1: **if** population has converged to minimum diversity **then**
2:    remove all duplicate individuals;
3:    **while** population not full **do**
4:       insert random draw from generated individuals into population;
5:    **end while**
6: **end if**

Experiment Setup

Table : System Parameters used in GA and RGA

Settings	Values
Runs	30
Generations	Max. 30, Until Converged
Population Size	50
Crossover Rate	80, 70
Mutation rate	10, 20, 30
Elitism	True, False
Number of restart attempts (RGA)	3, 5
Max convergence percentage before restarting (RGA)	5%, 10%

Experiment Setup

- Initial population: molecular dynamics simulation.
- Representation 1: list of 3-dimensional coordinates.

Table : Results from GA Experiments

Gen.	Crossover	Mutation	Elitism	Best	Avg. Best
30	80%	20%	False	1.2471	1.3518
30	80%	10%	False	1.1880	1.3610
30	70%	30%	False	1.1173	1.2942
30	80%	20%	True	1.2287	1.3294
30	80%	10%	True	1.2349	1.3658
30	70%	30%	True	1.0533	1.3044

Table : Results from RGA Experiments

Gen.	Xover	Mut.	Elitism	Conv. Rate	Restarting	Best	Avg. Best
	80%	20%	True	10%	3		
	80%	20%	True	5%	3		
	80%	20%	True	10%	5		
	80%	20%	True	5%	5		
	70%	30%	True	10%	3		
	70%	30%	True	5%	3		
	70%	30%	True	10%	5		
	70%	30%	True	5%	5		

Analysis

Genetic Algorithm

- Mann-Whitney U tests performed.
- Experiments using a crossover rate of 70% and a mutation rate of 30% performed statistically better than others.
- May indicate problem prefers exploration over exploitation.

Restarting Genetic Algorithm

- Mann-Whitney U tests performed.
- RGA experiments performed statistically better than all GA experiments.
- Experiment with the highest number of restarts and lowest convergence criteria performed statistically better than others.
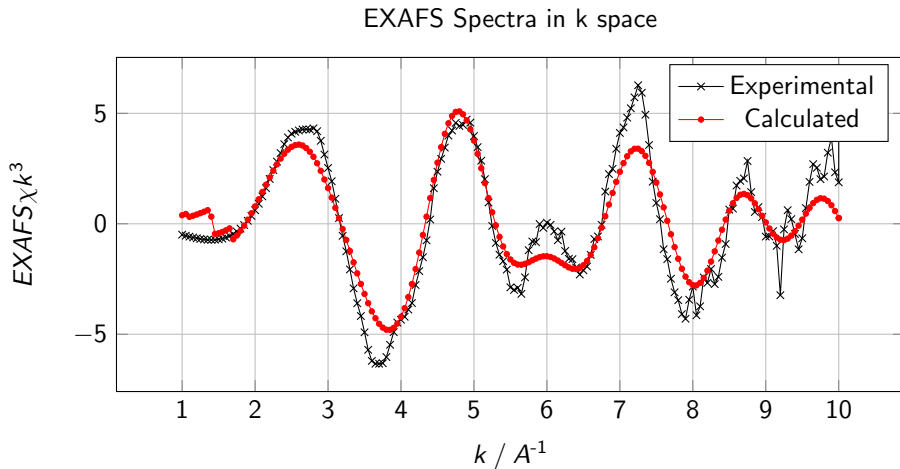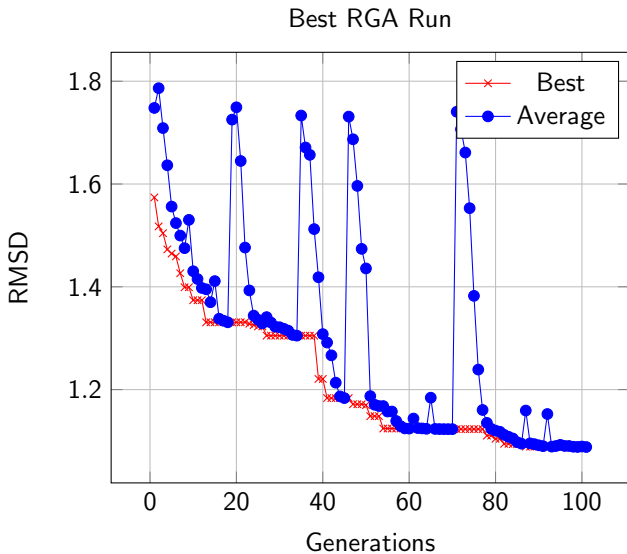- Results show GA experiments converging early on local optima.

Best RGA Spectra

Figure : OEC EXAFS Spectra Comparison

Results

Best RGA Run

Post-Optimization

- Results from GA and RGA were successful but we could do better.
- Convinced by guest speaker to try Particle Swarm Optimization.
- Convinced by fellow student to try Differential Evolution.
- Used the computational intelligence library, CILIB.

Particle Swarm Optimization (PSO) & Differential Evolution (DE)

- Excel in continuous space problems.
- Offers more fluid movement through the search space.
- Still gain a population of candidate solutions.

Experiment Setup

- Initial population: random.
- Representation 2: list of real numbers.

Table : System Parameters used in GA and RGA

Settings	Values
Runs	30
Generations	30
Population Size	50
Initial Movement Radius	0.05, 0.25

Results

- Initial RMSD from RGA experiment was 1.0877.

Table : Local Search Results

Algorithm	Initial Movement Radius	Pop. Size	Gen.	Average Best
PSO	± 0.05Å	50	30	0.8976
DE	± 0.05Å	50	30	**1.1354**
PSO	± 0.25Å	50	30	1.2411
DE	± 0.25Å	50	30	1.7220

- DE was not able to improve upon previously found candidate solutions.
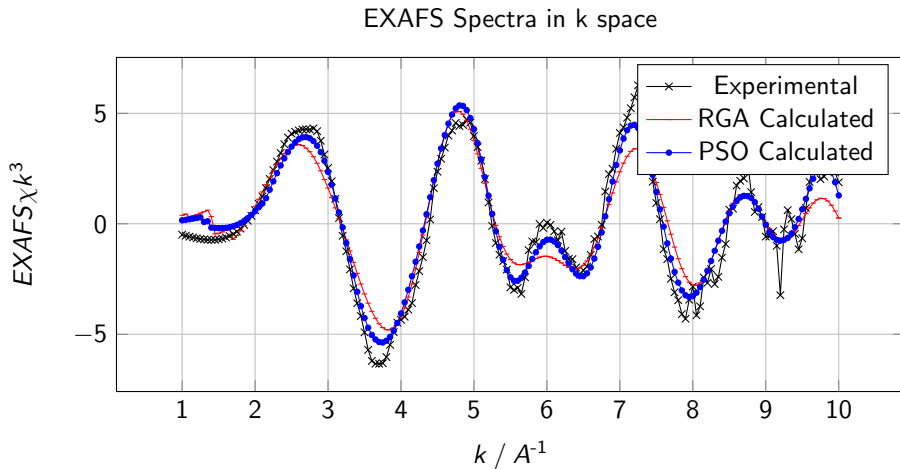
Analysis

Best Result

Figure : OEC EXAFS Spectra Comparison for PSO Post-Optimization

Experimental Design

Parameters

Results

Analysis

Comparison

Table : Comparison of Results

Algorithm	Best Results
DFT-QM/MM	1.2643
R-QM/MM	1.2403
GA	1.1452
RGA	1.0877
Post-Op PSO	0.7287
Post-Op DE	1.0877
PSO	0.7287
DE	1.0877

Subset Testing

- Individuals contain 79 atoms.
- Wanted reduce the search space.
- Tested keeping certain chemical elements rigid during experiments.

Table : Chemical Element Breakdown of OEC

Element	Count
Mn	4
Ca	1
O	26
C	14
N	6
H	28
Total	79

GA Parameters

Table : GA Subset Parameters

Runs	10
Population size	50
Crossover rate	0.7
Mutation rate	0.3
Elitism	True
Number of restart attempts	5
Max convergence percentage before restarting	5%

Results

Table : Results from subsets experiments

Flexible Atoms	Rigid Atoms	Best	Average
Mn, Ca, C, O, N, H	None	1.1998	1.2580
Mn, Ca, C, O, N	H	**1.1697**	1.2621
Mn, Ca, C	O, N, H	2.4413	N/A
Mn, Ca, O	C, N, H	1.2531	N/A
Mn, Ca, N	C, O, H	2.4881	N/A
Mn, Ca, C, O	N, H	**1.1648**	N/A
Mn, Ca, C, N	O, H	2.5143	2.5649
Mn, Ca, O, N	C, H	N/A	N/A
Mn, Ca	C, O, N, H	2.4916	2.5088

Note: N/A means there were too many invalid solutions.

Conclusion

- ...

Future Work

- Attempt direct comparison of PSO and DE with GA and RGA.
- Perform another molecular dynamics simulation on best result found from experimentation.
- Incorporate force fields into the algorithms to guarantee chemically feasible atomic structures.
- Multi-objective optimization using RMSD and potential energy.

Questions?