Computational Analysis of Protein Coevolution and Interaction

Fábio Madeira, PhD student

CENTRIA-DI, Faculdade de Ciências e Tecnologia

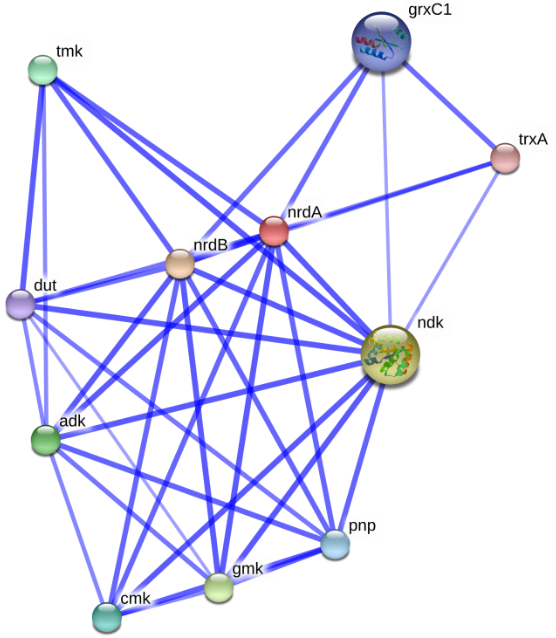
Universidade
Nova de Lisboa

Outline

1. Introduction
2. Motivation
3. Objectives
4. Results
5. Summary

Protein-protein interactions

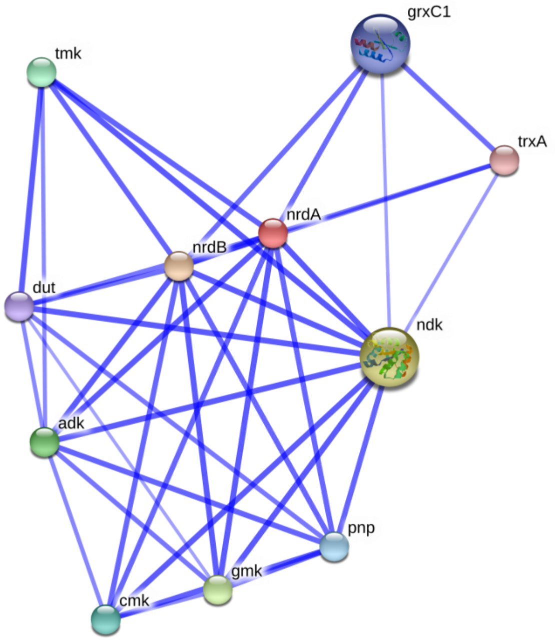
- Play a crucial role in biological systems



Protein interactions of RNR (STRING)

Protein-protein interactions

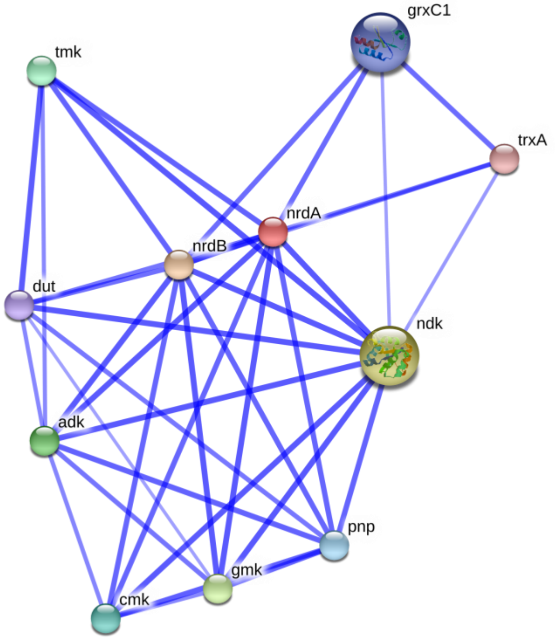
- Play a crucial role in biological systems
- Invaluable in expanding our understanding of diverse biological processes



Protein interactions of RNR (STRING)

Protein-protein interactions

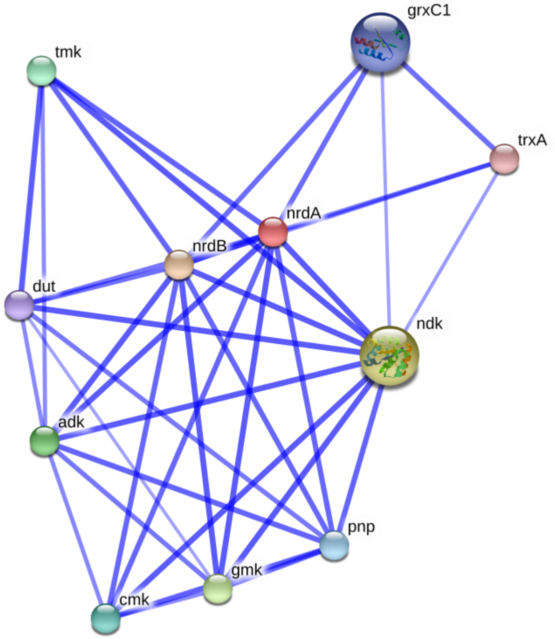
- Play a crucial role in biological systems
- Invaluable in expanding our understanding of diverse biological processes
- Hard to determine by experimental methods



Protein interactions of RNR (STRING)

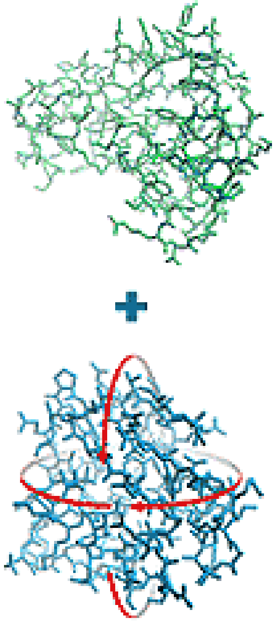
Protein-protein interactions

- Play a crucial role in biological systems
- Invaluable in expanding our understanding of diverse biological processes
- Hard to determine by experimental methods
- Hard to determine by computational methods



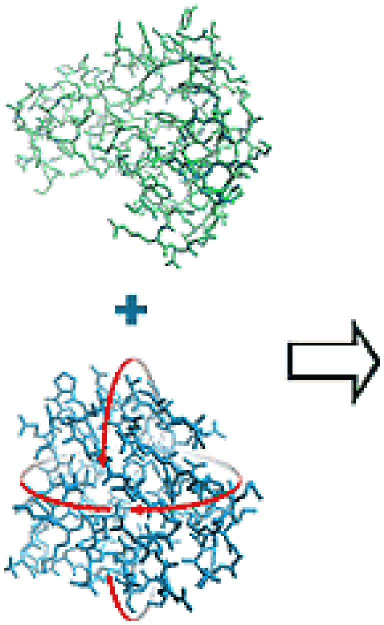
Protein interactions of RNR (STRING)

Protein docking

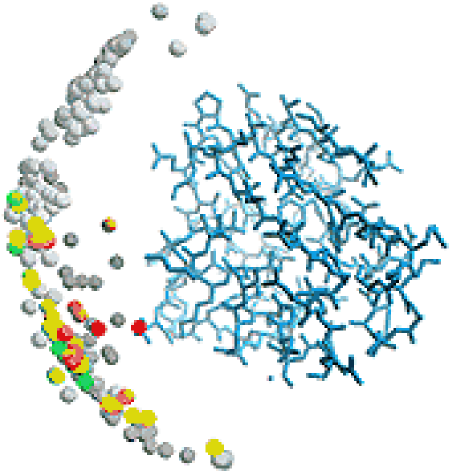


1. Search

Protein docking

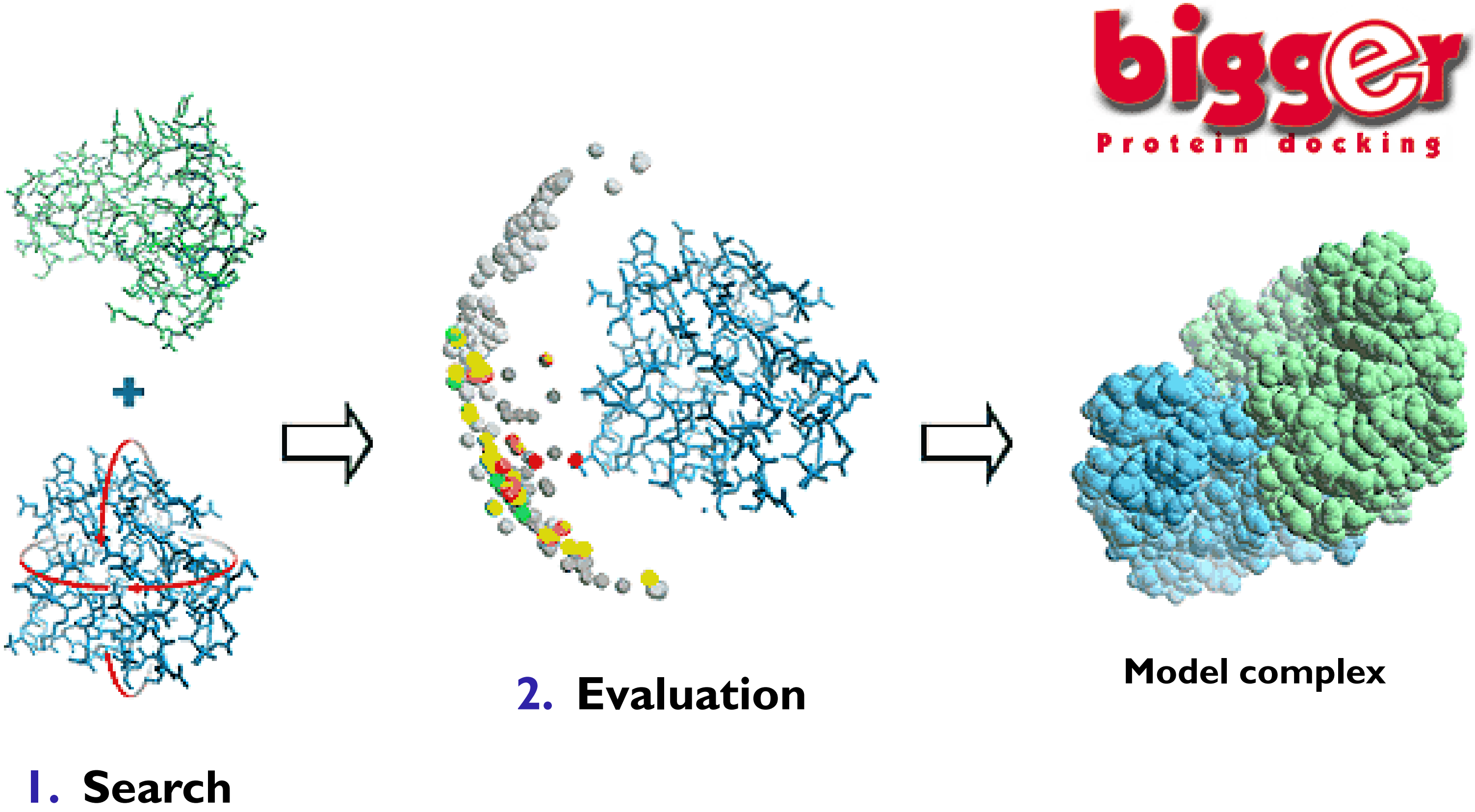


1. Search



2. Evaluation

Protein docking



1. Search

2. Evaluation

Model complex

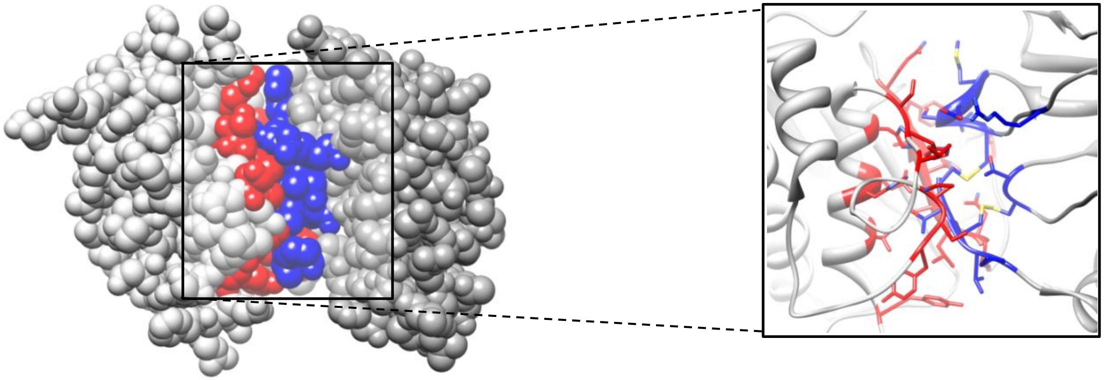
Protein coevolution (or correlated mutations)

- Accumulation of sequence changes in one protein triggered by changes in other regions of the same protein or in another protein

Protein coevolution (or correlated mutations)

- Accumulation of sequence changes in one protein triggered by changes in other regions of the same protein or in another protein

- Applied successfully for:
  - **Prediction of protein interactions**
  - **To guide protein docking**



The "classic" coevolution workflow

The “classic” coevolution workflow

1. **Homologous sequences**

DPSLDRPFISEGTTLKDLIYDMTT

VEGMIKLALSTASGLAHLHMEI

The “classic” coevolution workflow

1. Homologous sequences

```
DPSLDRPFISEGTTLKDLIYDMTT
-----EPRPDSGRDWSVELQEL--
-----KPGPDLGRDWSVELQEL--
-----EPESDSGRDWSAELPEL--
-----EPEPGSGGDCSEELPEL--
-----DPEPGSGGDCSEELPEL--
PEPEQEPEPDSGGDCSAELPEL--
RKQGLHSMNMMEAACSEPSLDL--
```

2. MSA computation

```
VEGMIKLALSTASGLAHLHMEI
WGSSLRMALSLAQGLAFLHEER
WGSSLRMALSLAQGLAFLHEER
WGSSLRMALSLAQGLAFLHEER
WGSSLSMALSLAEGLAFLHGRR
WGSSLSMALSLAEGLAFLHERR
--SSMSMALSLAQGLAFLHER-
--SSCRLAHSITRGLAYLHTRR
```

The “classic” coevolution workflow

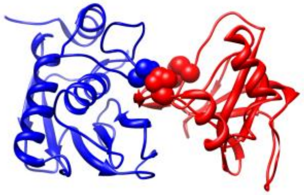
1. **Homologous sequences**

```
DPSLDRPFISEGTTLKDLIYDMTT
-----EPRPDSGRDWSVELQEL--  i
-----KPGPDLGRDWSVELQEL--
-----EPESDSGRDWSAELPEL--
-----EPEPGSGGDCSEELPEL--
-----DPEPGSGGDCSEELPEL--
PEPEQEPEPDSGGDCSAELPEL--
RKQGLHSMNMMEAACSEPSLDL--
              x
```

2. **MSA computation**

```
VEGMIKLALSTASGLAHLHMEI
WGSSLRMALSLAQGLAFLHEER
WGSSLRMALSLAQGLAFLHEER j
WGSSLRMALSLAQGLAFLHEER
WGSSLSMALSLAEGLAFLHGRR
WGSSLSMALSLAEGLAFLHERR
--SSMSMALSLAQGLAFLHER-
--SSCRLAHSITRGLAYLHTRR
     y
```

3. **Coevolution analysis**



Methods for finding protein coevolution

- Matrix-based
  - Physicochemical propensities and contact preferences (CPVN, CLM, etc.)
- Correlation-based
  - Correlation coefficient (Pearson's, Spearman's, Quartets, OMES, etc.)
- Statistical-based
  - Perturbation of MSA (MI, SCA, ELSC, etc.)
- Phylogenetic-based
  - Similarity of phylogenetic trees (e.g. Mirrortree, etc.)

Multiple Sequence Alignments

MSA =

```
QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
--------------------------------------ECPGHFGHIELAK
--------------------------------------ECPGHFGFIELAK
QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE
```

Multiple Sequence Alignments

MSA =

```
QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
--------------------------------------ECPGHFGHIELAK
--------------------------------------ECPGHFGFIELAK
QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE
```

- Insufficient evolutionary divergence
- Sample size effects
- Small taxa coverage
- False correlations as result of misalignments
- **Underlying assumption of independent mutations**

Outline

1. Introduction
2. Motivation
3. Objectives
4. Results
5. Summary

Motivation

- Coevolution data can help predict protein interaction and to guide docking

Motivation

- Coevolution data can help predict protein interaction and to guide docking

- BiGGER can use contact information to constraint the search space and improve the quality of the models

Motivation

- Coevolution data can help predict protein interaction and to guide docking
- BiGGER can use contact information to constraint the search space and improve the quality of the models
- Limitations of MSA, such as the assumption of independent mutations

Outline

1. Introduction
2. Motivation
3. Objectives
4. Results
5. Summary

Objectives

- The computational analysis of protein coevolution and interaction will focus on:

Objectives

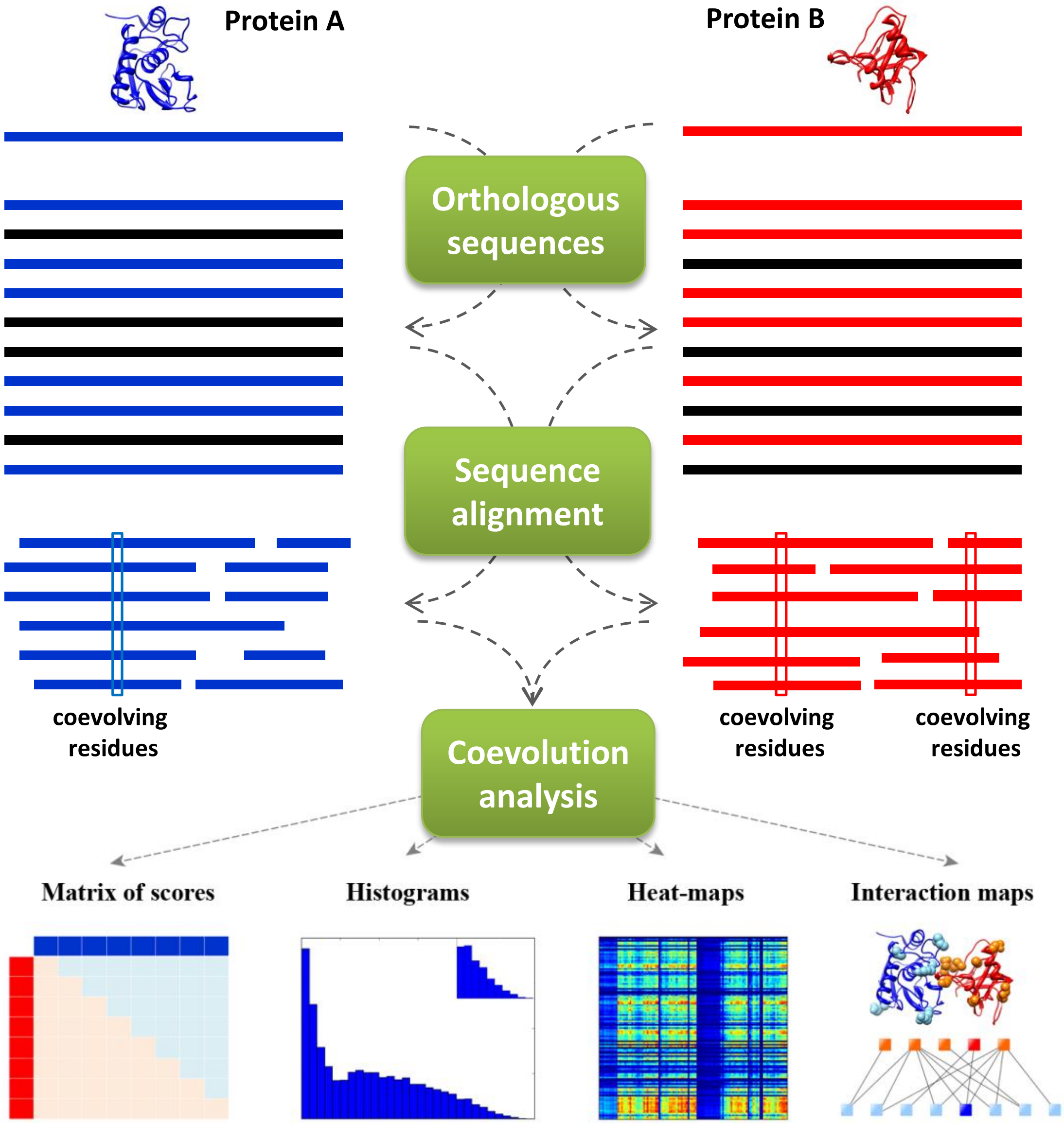
➢ The computational analysis of protein coevolution and interaction will focus on:

- Refinement of Multiple Sequence Alignments
- Integration of protein coevolution data in the protein docking workflow
- Development and assessment of different coevolution measures
- Development of new scoring methods to rank docking solutions

Outline

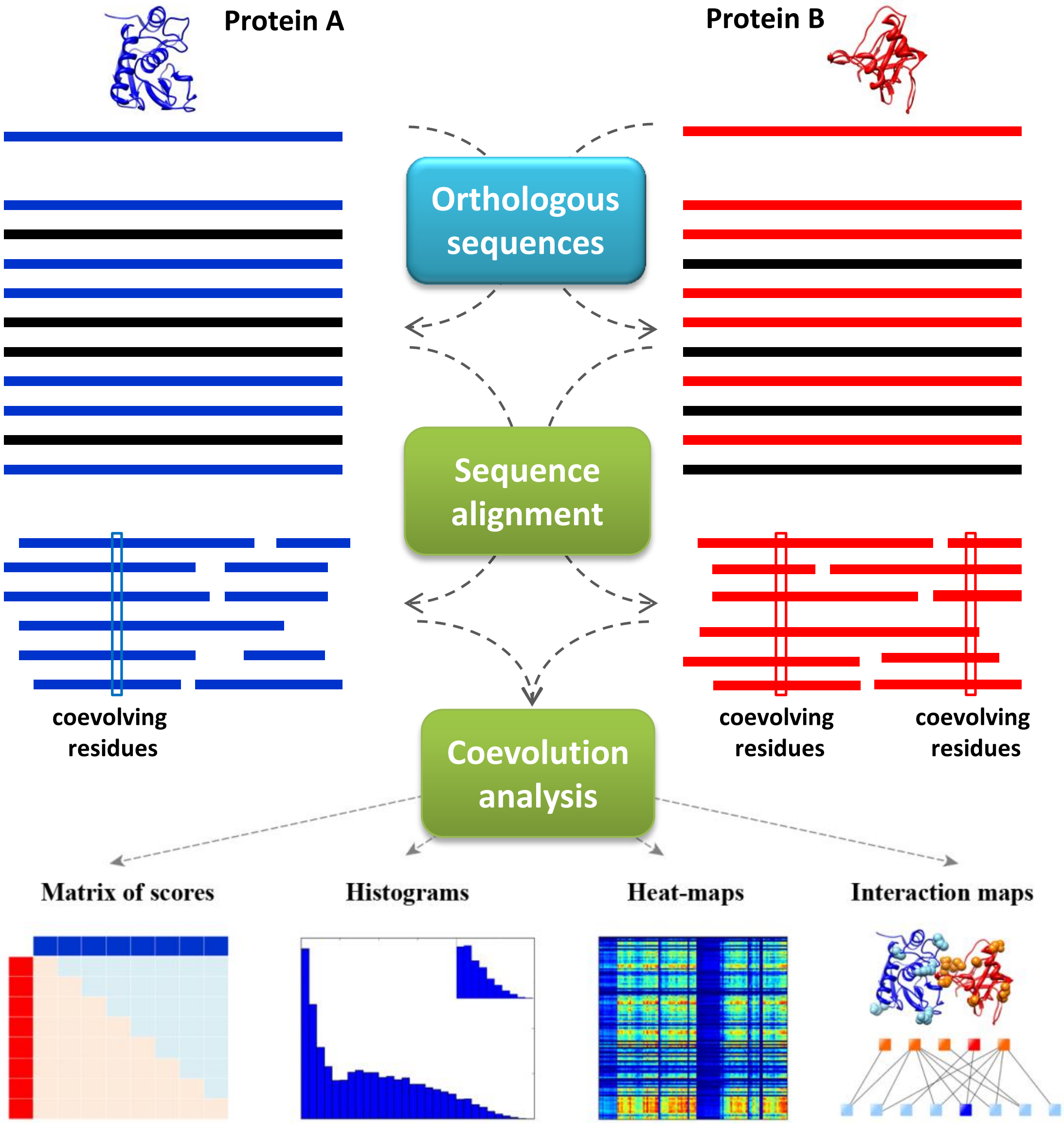
1. Introduction
2. Motivation
3. Objectives
4. Results
5. Summary

Pycoevol



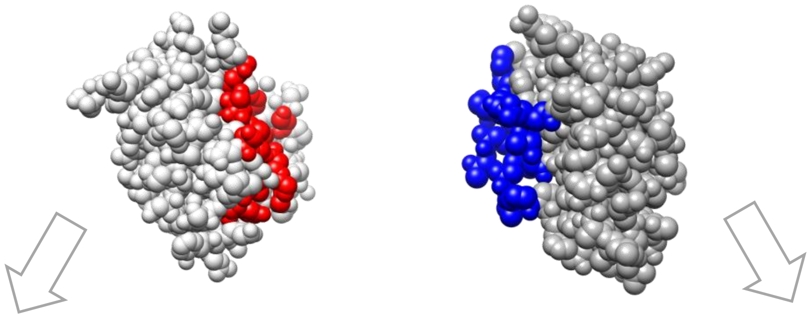
Madeira F. and Krippahl L. 2012. **PYCOEVOL: A Python workflow to study protein-protein coevolution**. BIOINFORMATICS 2012, pp.143-9.

Pycoevol



Madeira F. and Krippahl L. 2012. **PYCOEVOL: A Python workflow to study protein-protein coevolution**. BIOINFORMATICS 2012, pp.143-9.

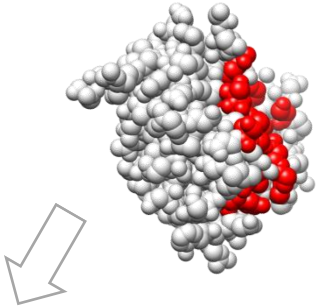
Matching of taxa

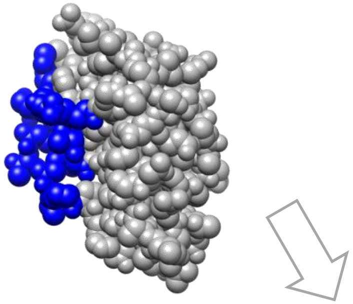


```
Organism1  DPSLDRPFISEGTTLKDLIYDMTT
Organism2  -----EPRPDSGRDWSVELQEL--
Organism3  -----KPGPDLGRDWSVELQEL--
Organism4  -----EPESDSGRDWSAELPEL--
Organism5  -----EPEPGSGGDCSEELPEL--
Organism6  -----DPEPGSGGDCSEELPEL--
Organism7  PEPEQEPEPDSGGDCSAELPEL--
Organism8  RKQGLHSMNMMEAACSEPSLDL--
```

```
Organism3   VEGMIKLALSTASGLAHLHMEI
Organism1   WGSSLRMALSLAQGLAFLHEER
Organism9   WGSSLRMALSLAQGLAFLHEER
Organism8   WGSSLRMALSLAQGLAFLHEER
Organism5   WGSSLSMALSLAEGLAFLHGRR
Organism10  WGSSLSMALSLAEGLAFLHERR
Organism2   --SSMSMALSLAQGLAFLHER-
Organism11  --SSCRLAHSITRGLAYLHTRR
```

Matching of taxa

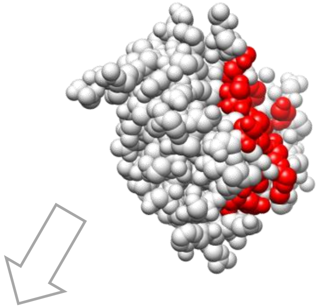


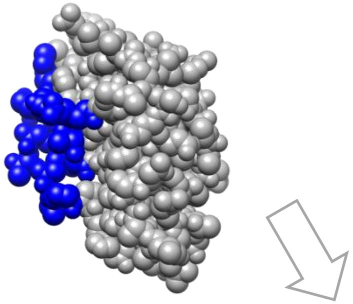


Organism1	DPSLDRPFISEGTTLKDLIYDMTT
Organism2	-----EPRPDSGRDWSVELQEL--
Organism3	-----KPGPDLGRDWSVELQEL--
Organism4	-----EPESDSGRDWSAELPEL--
Organism5	-----EPEPGSGGDCSEELPEL--
Organism6	-----DPEPGSGGDCSEELPEL--
Organism7	PEPEQEPEPDSGGDCSAELPEL--
Organism8	RKQGLHSMNMMEAACSEPSLDL--

Organism3	VEGMIKLALSTASGLAHLHMEI
Organism1	WGSSLRMALSLAQGLAFLHEER
Organism9	WGSSLRMALSLAQGLAFLHEER
Organism8	WGSSLRMALSLAQGLAFLHEER
Organism5	WGSSLSMALSLAEGLAFLHGRR
Organism10	WGSSLSMALSLAEGLAFLHERR
Organism2	--SSMSMALSLAQGLAFLHER-
Organism11	--SSCRLAHSITRGLAYLHTRR

Matching of taxa

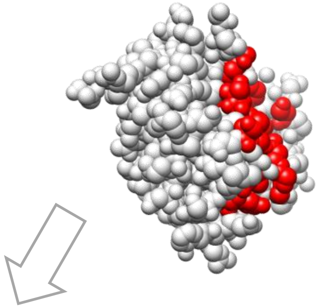




Organism1	DPSLDRPFISEGTTLKDLIYDMTT
Organism2	-----EPRPDSGRDWSVELQEL--
Organism3	-----KPGPDLGRDWSVELQEL--
~~Organism4~~	~~-----EPESDSGRDWSAELPEL--~~
Organism5	-----EPEPGSGGDCSEELPEL--
~~Organism6~~	~~-----DPEPGSGGDCSEELPEL--~~
~~Organism7~~	~~PEPEQEPEPDSGGDCSAELPEL--~~
Organism8	RKQGLHSMNMMEAACSEPSLDL--

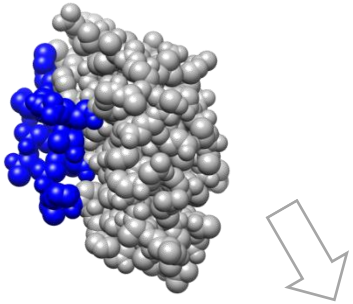
Organism3	VEGMIKLALSTASGLAHLHMEI
Organism1	WGSSLRMALSLAQGLAFLHEER
~~Organism9~~	~~WGSSLRMALSLAQGLAFLHEER~~
Organism8	WGSSLRMALSLAQGLAFLHEER
Organism5	WGSSLSMALSLAEGLAFLHGRR
~~Organism10~~	~~WGSSLSMALSLAEGLAFLHERR~~
Organism2	--SSMSMALSLAQGLAFLHER-
~~Organism11~~	~~--SSCRLAHSITRGLAYLHTRR~~

Matching of taxa



Organism1	DPSLDRPFISEGTTLKDLIYDMTT
Organism2	-----EPRPDSGRDWSVELQEL--
Organism3	-----KPGPDLGRDWSVELQEL--
~~Organism4~~	~~-----EPESDSGRDWSAELPEL--~~
Organism5	-----EPEPGSGGDCSEELPEL--
~~Organism6~~	~~-----DPEPGSGGDCSEELPEL--~~
~~Organism7~~	~~PEPEQEPEPDSGGDCSAELPEL--~~
Organism8	RKQGLHSMNMMEAACSEPSLDL--

Organism1	DPSLDRPFISEGTTLKDLIYDMTT
Organism3	-----KPGPDLGRDWSVELQEL--
Organism2	-----EPRPDSGRDWSVELQEL--
Organism5	-----EPEPGSGGDCSEELPEL--
Organism8	RKQGLHSMNMMEAACSEPSLDL--



Organism3	VEGMIKLALSTASGLAHLHMEI
Organism1	WGSSLRMALSLAQGLAFLHEER
~~Organism9~~	~~WGSSLRMALSLAQGLAFLHEER~~
Organism8	WGSSLRMALSLAQGLAFLHEER
Organism5	WGSSLSMALSLAEGLAFLHGRR
~~Organism10~~	~~WGSSLSMALSLAEGLAFLHERR~~
Organism2	--SSMSMALSLAQGLAFLHER-
~~Organism11~~	~~--SSCRLAHSITRGLAYLHTRR~~

Organism1	WGSSLRMALSLAQGLAFLHEER
Organism3	VEGMIKLALSTASGLAHLHMEI
Organism2	--SSMSMALSLAQGLAFLHER-
Organism5	WGSSLSMALSLAEGLAFLHGRR
Organism8	WGSSLRMALSLAQGLAFLHEER

Pairwise distances and orthologs selection

- p-distance *[Jukes and Cantor, 1969]*

$$d = p = \frac{N_{dif}}{N_{total}}$$

- Jukes-Cantor *[Jukes and Cantor, 1969]*

$$d = -\frac{19}{20}log(1 - p * \frac{20}{19})$$

- Kimura distance *[Kimura, 1983]*

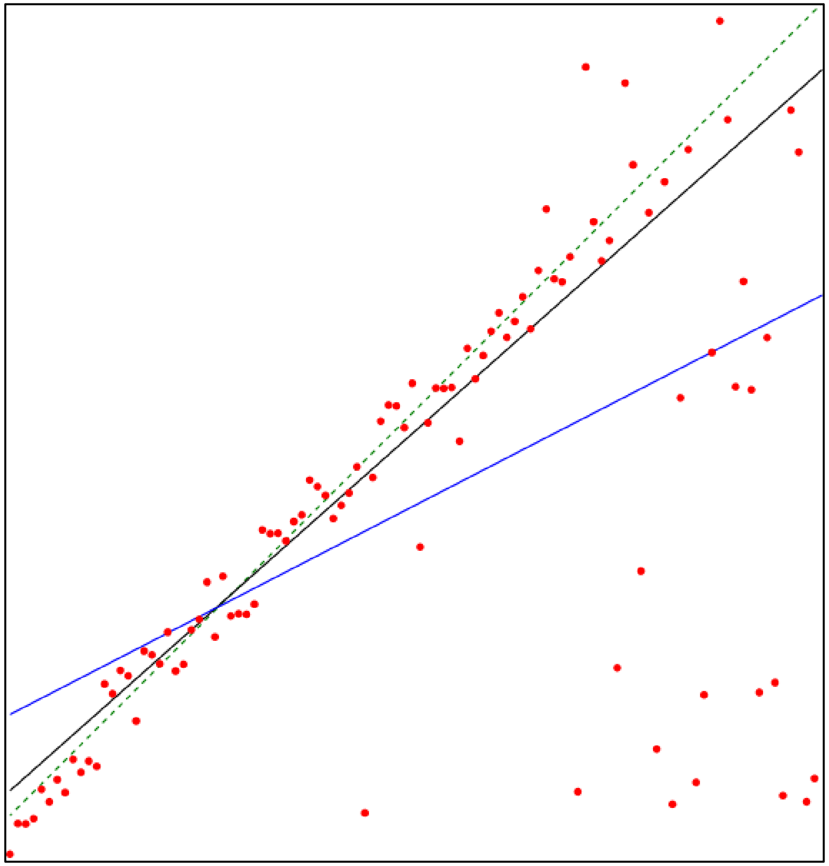
$$d = -ln(1 - p - 0.2^2)$$

- Pairwise score using Dayhoff or PAM matrices [*Gonnet, 2000*]

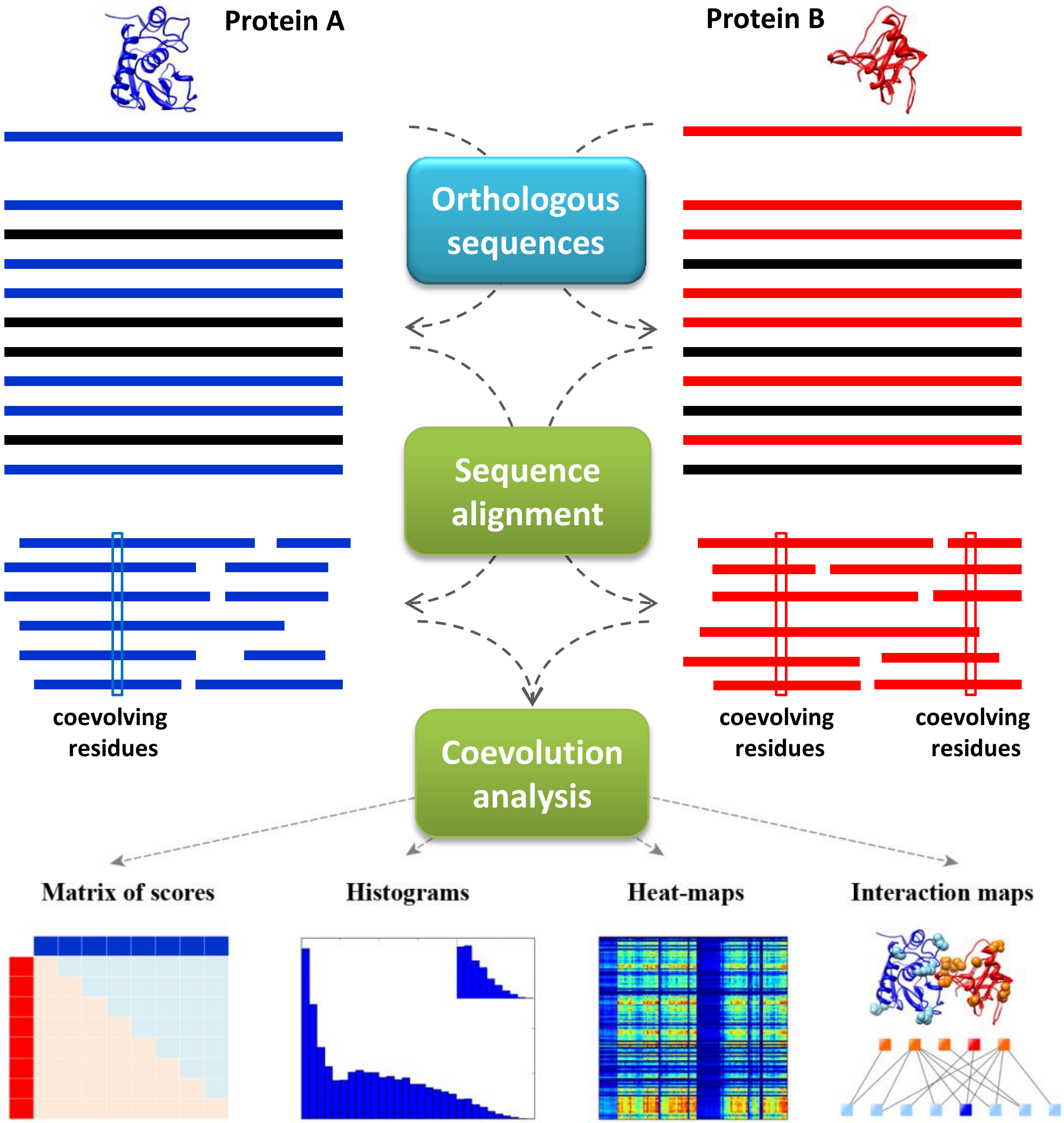
$$d = \sum \frac{1 - S_{ij}}{S_{ii}} * \frac{1 - S_{ij}}{S_{jj}}$$

Theil-Sen regression and sequence removal

- *Median m of the slopes $(y_j - y_i)/(x_j - x_i)$ determined by all pairs of sample points $\in (x,y)$, which have distinct x-coordinates [Theil, 1950] and [Sen, 1968]*

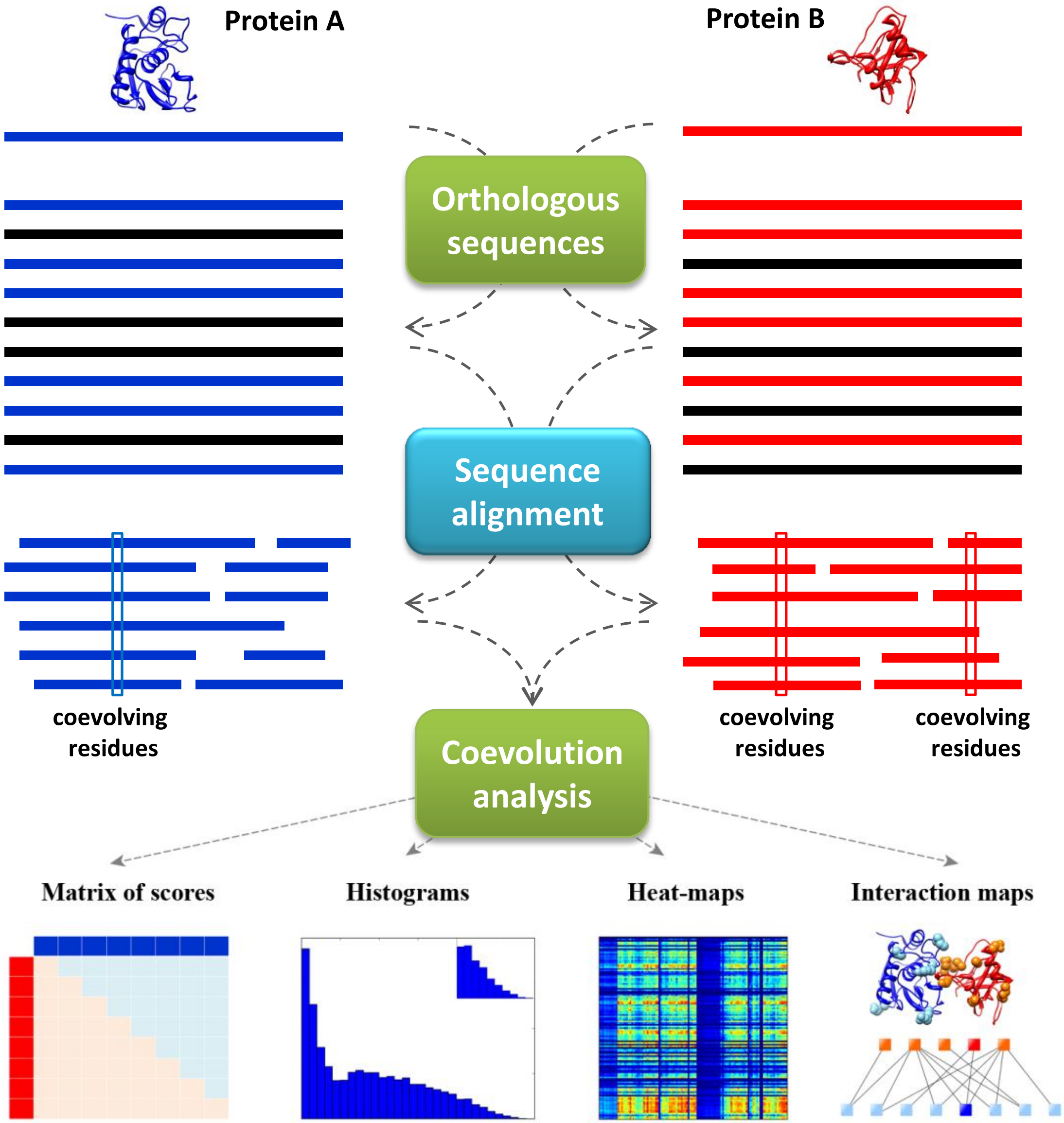


Pycoevol



Madeira F. and Krippahl L. 2012. **PYCOEVOL: A Python workflow to study protein-protein coevolution**. BIOINFORMATICS 2012, pp.143-9.

Pycoevol



Madeira F. and Krippahl L. 2012. **PYCOEVOL: A Python workflow to study protein-protein coevolution**. BIOINFORMATICS 2012, pp.143-9.

Refinement of Multiple Sequence Alignments

```
QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
--------------------------------------ECPGHFGHIELAK
--------------------------------------ECPGHFGFIELAK
QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE
```

- Constraint Programming

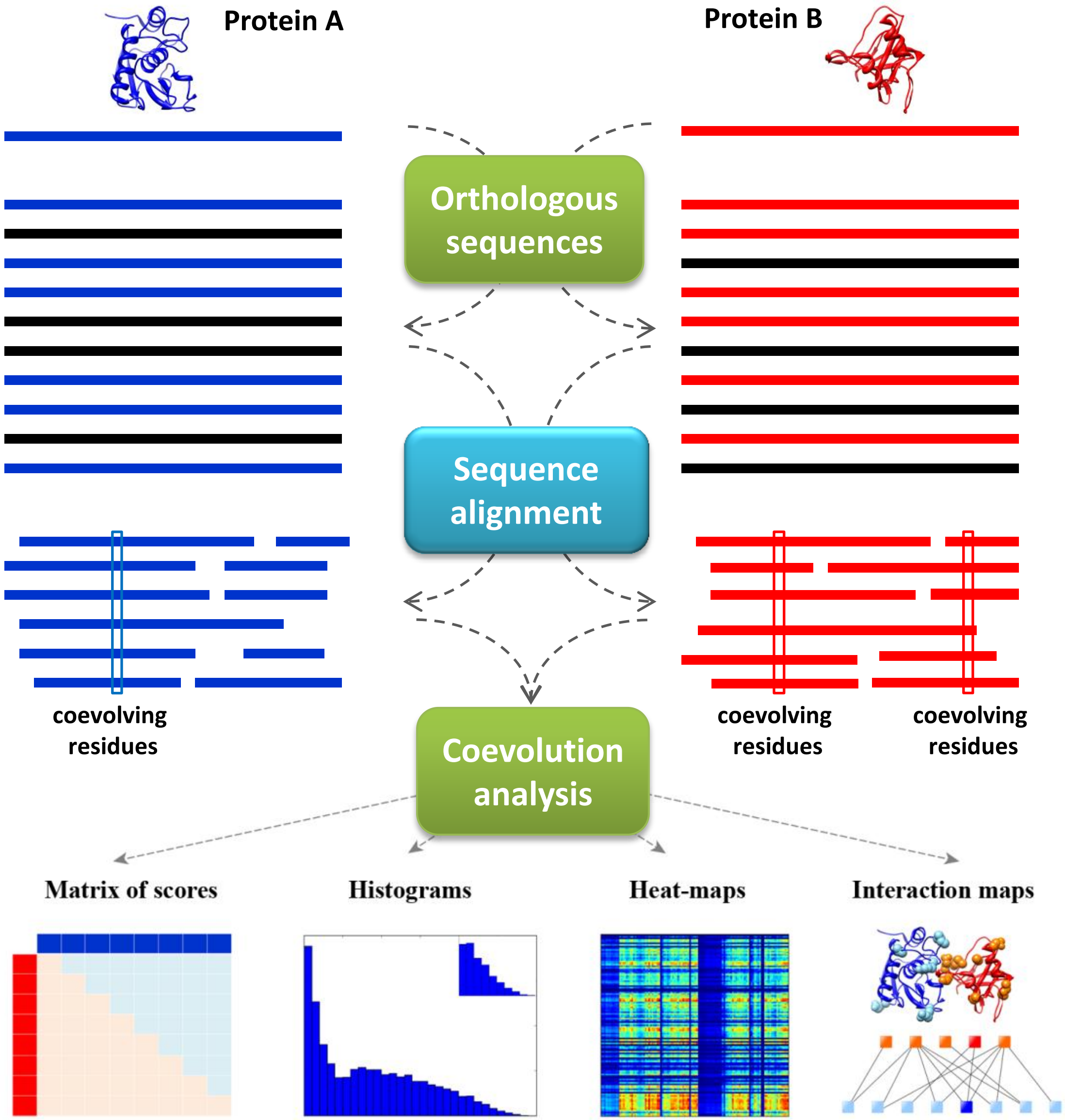
misalignments

Accounts for correlated mutations

```
QFGLFSPEEIRASSVAL--IRYPETLE--NGVPKESGLVCAGHFGHIELVK
QFGLFSPEEIKRMSVVH--VEYPETMDEQRQRPRTKGLECPGHFGHIELAT
QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
ELGVLDPEIIKKISVCE--IVNVDIYK--DGFPREGGLYCPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
QFGILSPDEIRQMSVIH--VEHSETTE--KGKPKVGGLECPGHFGYLELAK
--------------------------------------ECPGHFGHIELAK
--------------------------------------ECPGHFGFIELAK
QFEIFKERQIKSYAVCL--VEHAKSYA--NAADQSGEAECPGHFGYIELAE
QFEVFKEAQIKAYAKCI--IEHAKSY--EHGQPVRGGIECPGHFGYVELAE
```

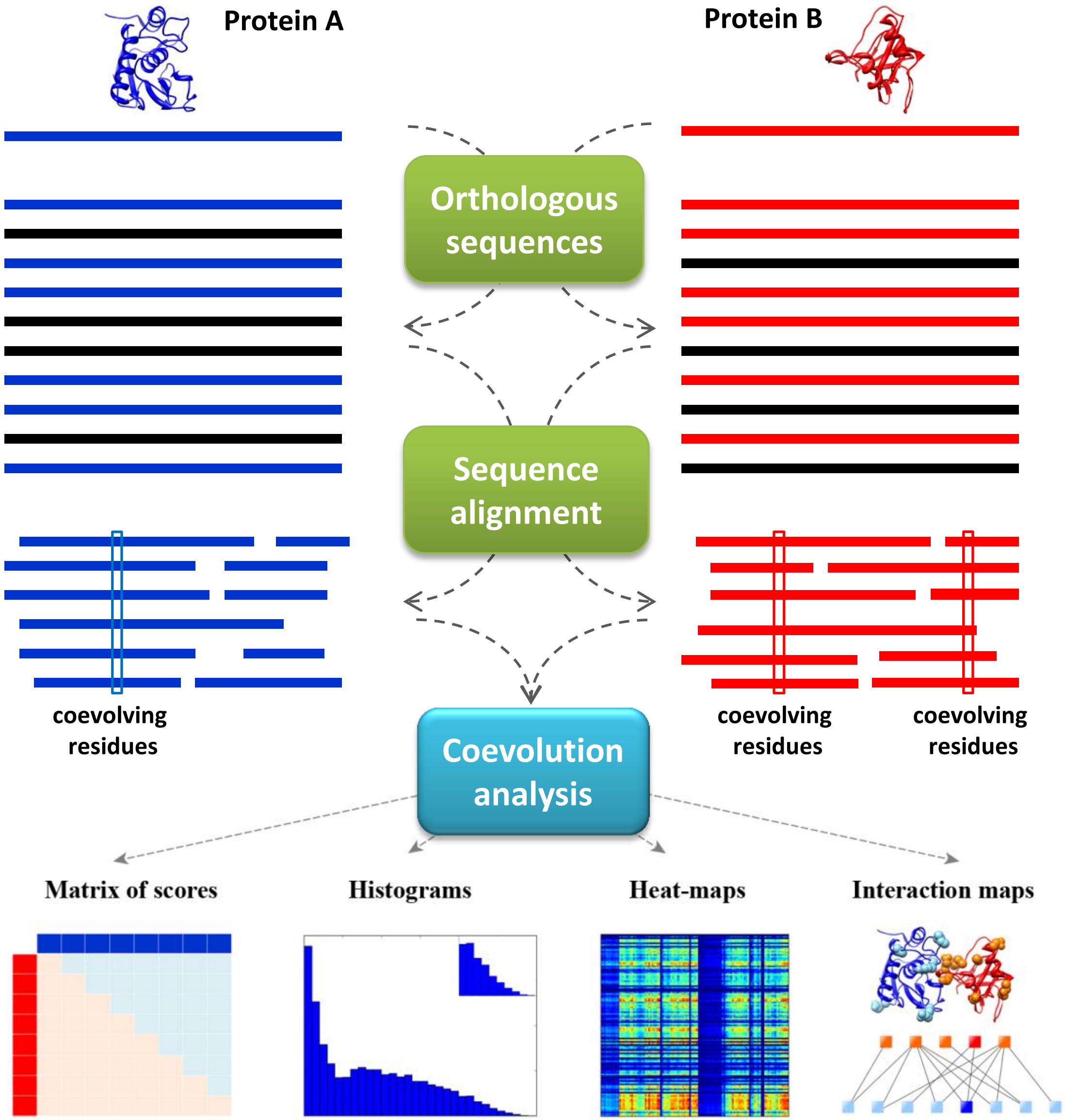
Correia, M. Madeira, F., Barahona, P. and Krippahl, L. 2011. **Improving Multiple Sequence Alignments with Constraint Programming and Local Search**. WCB11, pp.37-44.

Pycoevol



Madeira F. and Krippahl L. 2012. **PYCOEVOL: A Python workflow to study protein-protein coevolution**. BIOINFORMATICS 2012, pp.143-9.

Pycoevol



Madeira F. and Krippahl L. 2012. **PYCOEVOL: A Python workflow to study protein-protein coevolution**. BIOINFORMATICS 2012, pp.143-9.

Methods for finding protein coevolution

- Contact Preferences, Volume Normalized (**CPVN**) *[Glaser et al, 2001]*
- Contact PDB-derived Likelihood Matrix (**CLM**) *[Singer et al, 2002]*
- Residue-residue Volume Normalized (**VOL**) *[Esque et al, 2010]*

$$CM_{x,y} = \sum_{i}^{n} \sum_{j}^{n} \frac{S_{ij}}{n}$$

Methods for finding protein coevolution

- Pearson's correlation (**Pearson**) *[Göbel et al, 1994]*

- Spearman's rank correlation (**Spearman**) *[Pazos et al, 1997]*

- McLachlan Based Substitution Correlation (**McBASC**) *[Fodor and Aldrich, 2004]*

- ***Quartets*** *[Galitsky, 2002]*

$$CM_{x,y} = \frac{1}{N^2} \sum_i \sum_j \frac{W_{ij}(S_{xij} - \langle S_x \rangle)(S_{yij} - \langle S_y \rangle)}{\sigma_x \sigma_y}$$

- Observed Minus Expected Squared (**OMES**) *[Kass and Horovitz, 2002]*

$$CM_{x,y} = \sum_l^L \frac{(N_{obs} - \frac{C_{xi}C_{yj}}{N_{valid}})^2}{N_{valid}}$$

Methods for finding protein coevolution

- Mutual Information (**MI**) *[Gloor et al, 2005]*

- MI by pair Entropy (**MI/E**) *[Martin et al, 2005]*

- Row and Column Weighed MI (**RCW MI**) *[Gouveia-Oliveira et al, 2007]*

$$MI_{x,y} = \sum_i \sum_j P(x_i y_j) log \frac{P(x_i y_j)}{P(x_i)P(y_j)}$$

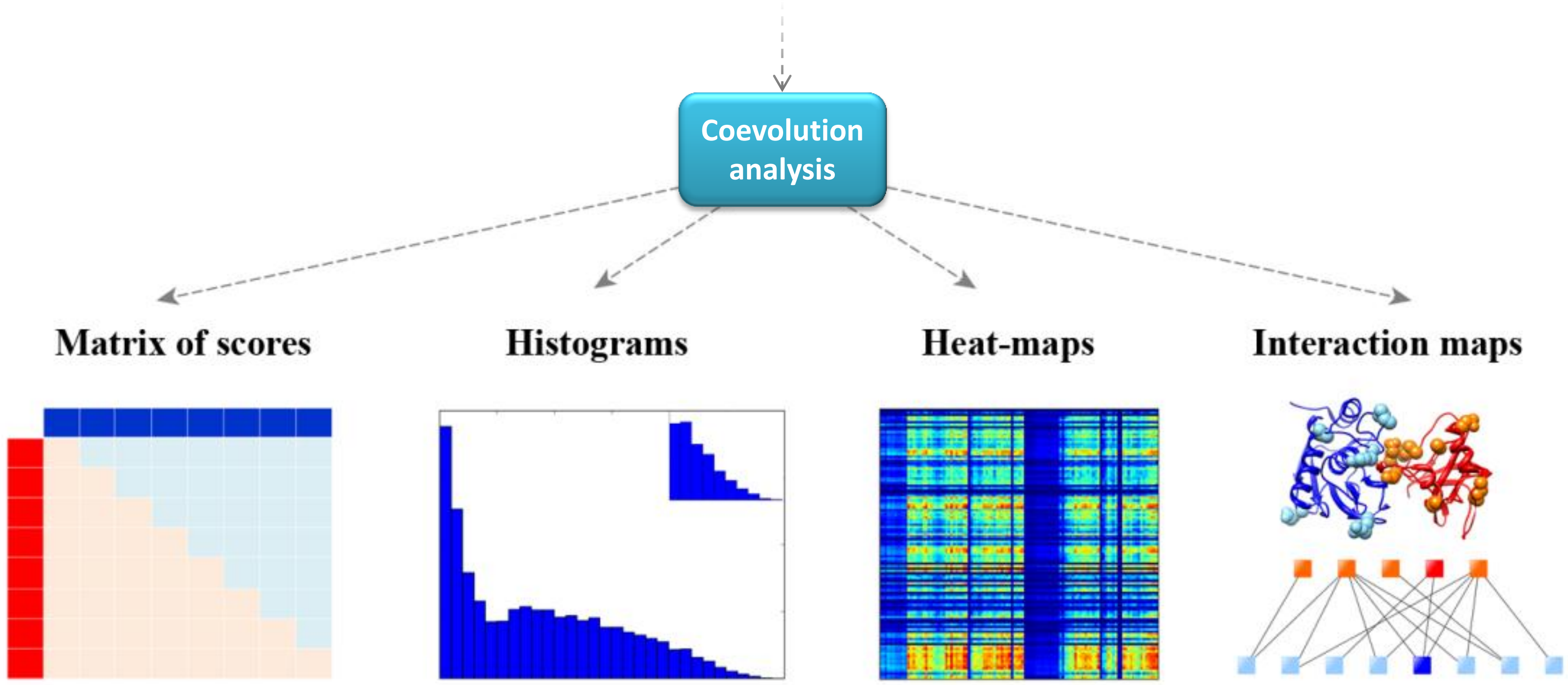
- Statistical Coupling Analysis (**SCA**) *[Lockless and Ranganathan, 1999]*

$$\Delta\Delta G_{x,y} = \sqrt{\sum_i (lnP^i_{x|\delta y} - P^i_x)^2}$$

- Explicit Likelihood of Subset Covariation (**ELSC**) *[Dekker et al, 2004]*

$$\Lambda^y_x = \prod_r^{20} \frac{\binom{N_{r,y}}{n_{r,y}}}{\binom{N_{r,y}}{m_{r,y}}}$$

Output results



Coevolution analysis

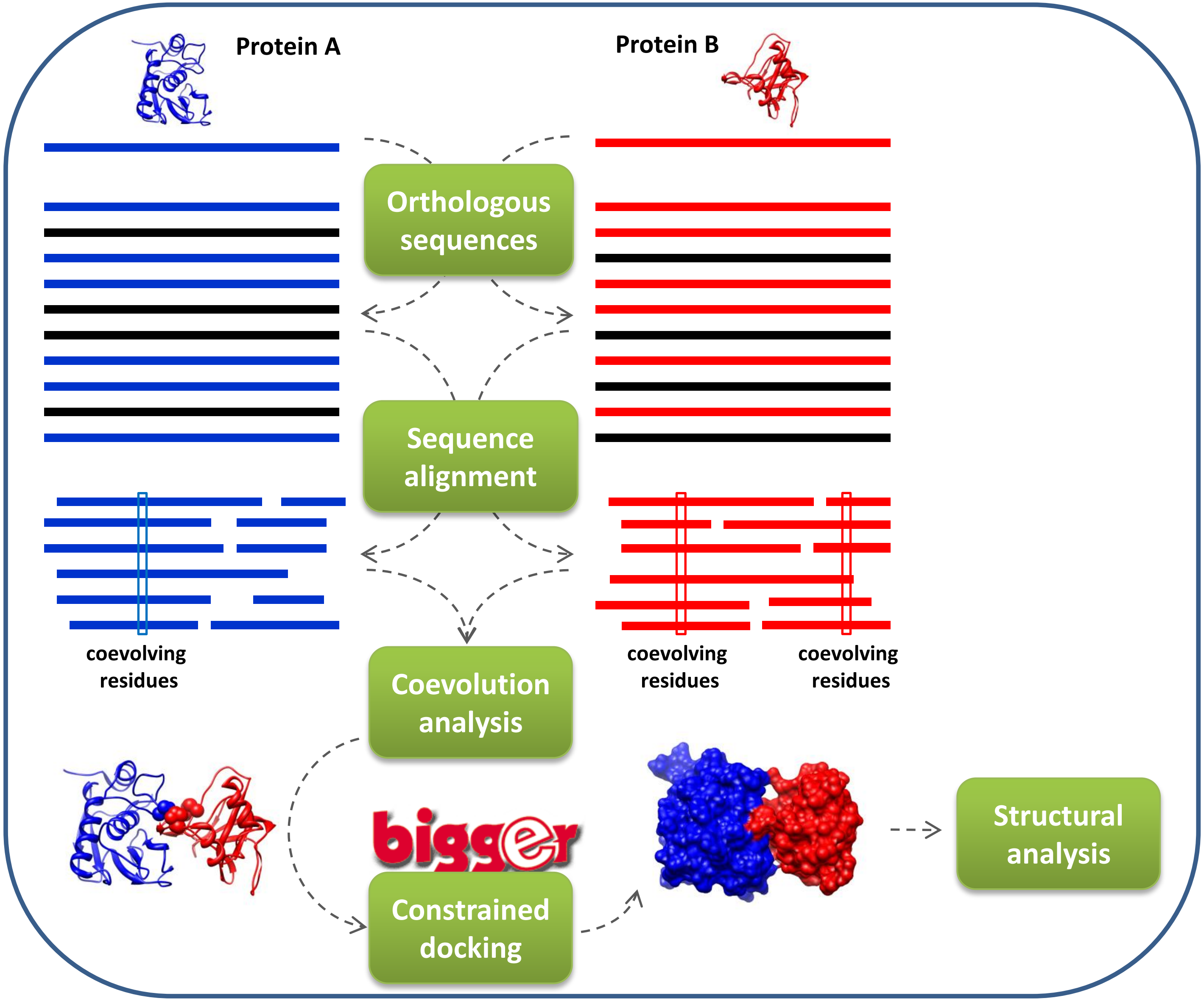
Matrix of scores

Histograms

Heat-maps

Interaction maps

Development of an integrated system



Protein A

Protein B

Orthologous sequences

Sequence alignment

coevolving residues

coevolving residues

coevolving residues

Coevolution analysis

bigger

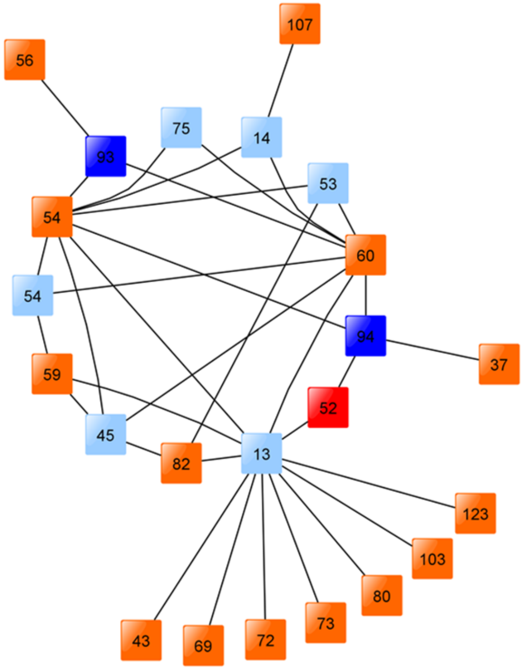
Constrained docking

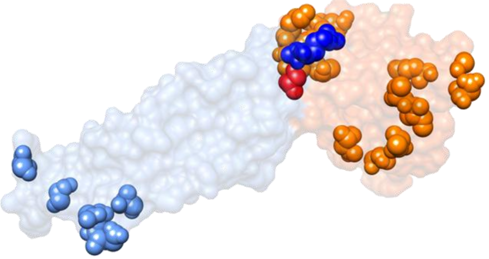
Structural analysis

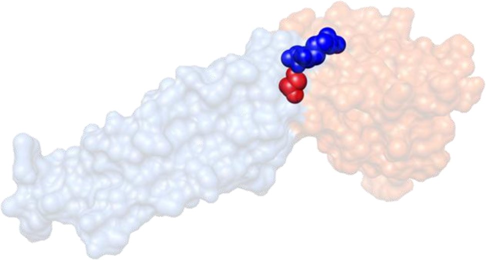
Example

Protein complex TGF-β3/TGF-β receptor type II (1ktz)

- Most residues were at the surface level
- 3 out of 23 residues were at the interface
- 8 residues from TGF-β3
- 15 residues from TGB-β receptor

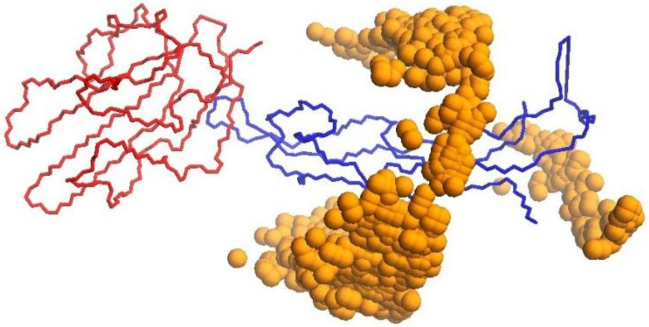






Improving protein-protein docking

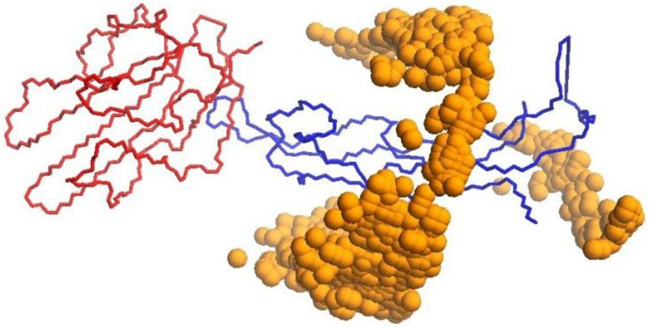
- Possible pairs of contacts = 11772
- Tested contacts = 120
- Reduced the search space to 0.01% of possible contact points



Difficult complex to model

Improving protein-protein docking

- Possible pairs of contacts = 11772
- Tested contacts = 120
- Reduced the search space to 0.01% of possible contact points



Difficult complex to model

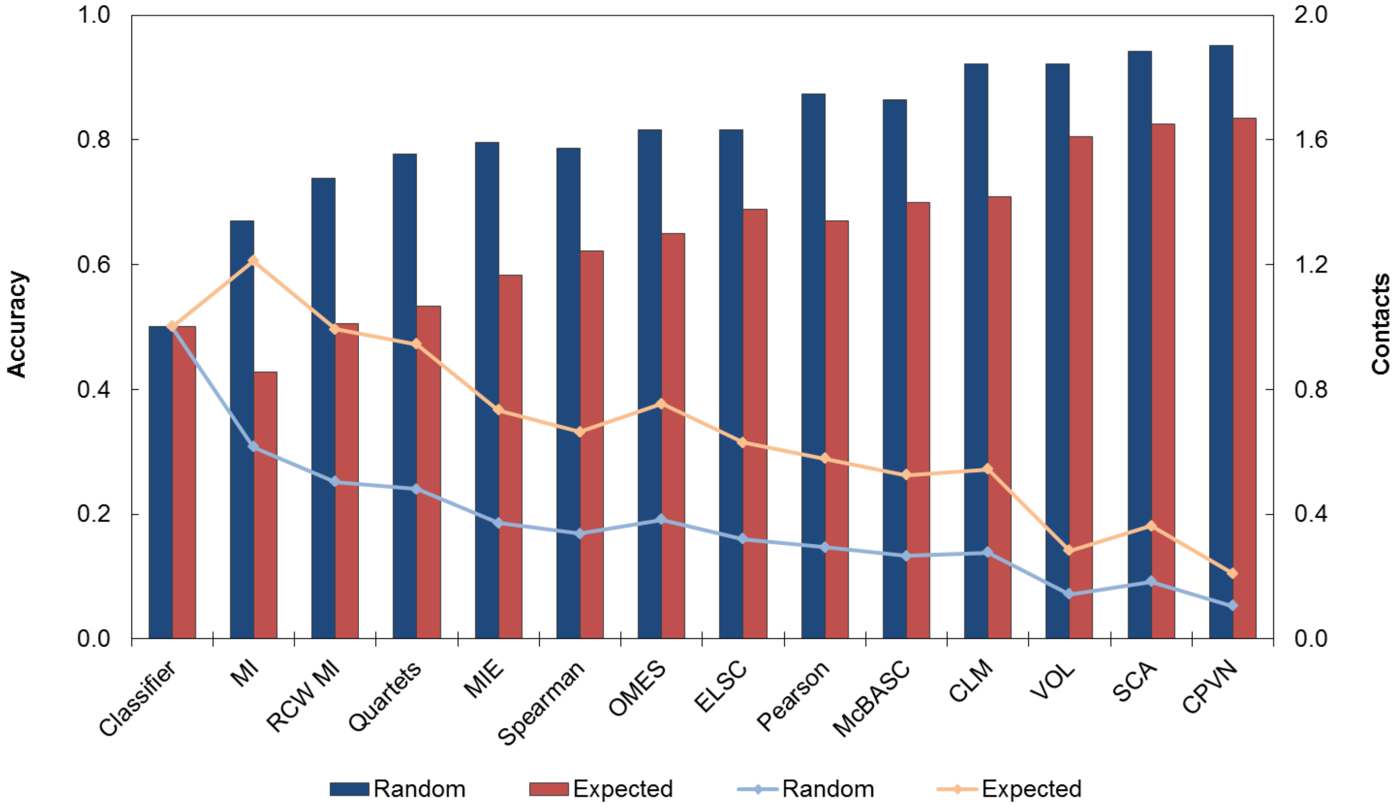
- Given the complexity of protein-protein docking, finding even only one positive interface contact can help constraint the search space and improve the accuracy of constrained docking algorithms such as BiGGER [*Palma et al, 2000*]

Evaluation of coevolution methods

- Survey which method is better suited for the identification of interface contact points using a large protein complex dataset (unpublished)

Evaluation of coevolution methods

- Survey which method is better suited for the identification of interface contact points using a large protein complex dataset (unpublished)



Outline

1. Introduction
2. Motivation
3. Objectives
4. Results
5. Summary

Pycoevol

- Set of tools for the study of inter-protein coevolution and interaction

Pycoevol

- Set of tools for the study of inter-protein coevolution and interaction
- Implements the classic workflow with extended capabilities

Pycoevol

- Set of tools for the study of inter-protein coevolution and interaction
- Implements the classic workflow with extended capabilities
- Automates access to remote databases and third-party applications

Pycoevol

- Set of tools for the study of inter-protein coevolution and interaction
- Implements the classic workflow with extended capabilities
- Automates access to remote databases and third-party applications
- Simplifies the coevolution analysis and the interpretation of results

Pycoevol

- Set of tools for the study of inter-protein coevolution and interaction
- Implements the classic workflow with extended capabilities
- Automates access to remote databases and third-party applications
- Simplifies the coevolution analysis and the interpretation of results
- Implemented in Python and platform independent

Pycoevol

- Set of tools for the study of inter-protein coevolution and interaction
- Implements the classic workflow with extended capabilities
- Automates access to remote databases and third-party applications
- Simplifies the coevolution analysis and the interpretation of results
- Implemented in Python and platform independent
- Open source

Pycoevol

- Set of tools for the study of inter-protein coevolution and interaction
- Implements the classic workflow with extended capabilities
- Automates access to remote databases and third-party applications
- Simplifies the coevolution analysis and the interpretation of results
- Implemented in Python and platform independent
- Open source
- Source code at https://github.com/fmadeira/pycoevol

Thanks for your attention!

Computational Analysis of Protein Coevolution and Interaction

Fábio Madeira - fmadeira@campus.fct.unl.pt

CENTRIA-DI, Faculdade de Ciências e Tecnologia

Universidade
Nova de Lisboa