Folding with Coevolution

Evan H. Baugh

ehb250@nyu.edu

Bonneau Lab

RosettaCon X

What is coevolution?

- when a “biological object” changes due to changes in another “biological object”

What is coevolution?

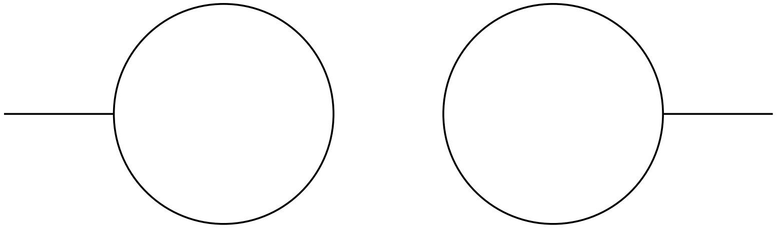
- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes

What is coevolution?

- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes
- indicates functional association or physical interaction

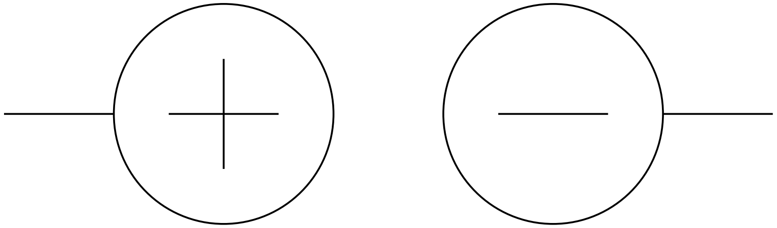
What is coevolution?

- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes
- indicates functional association or physical interaction



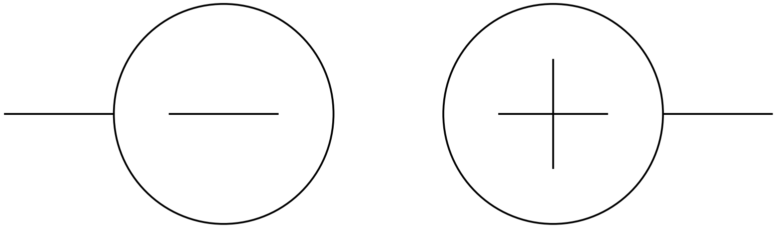
What is coevolution?

- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes
- indicates functional association or physical interaction



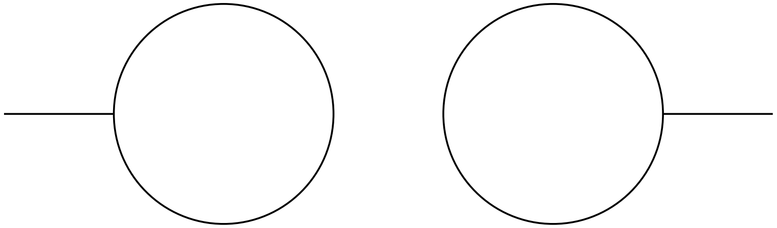
What is coevolution?

- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes
- indicates functional association or physical interaction



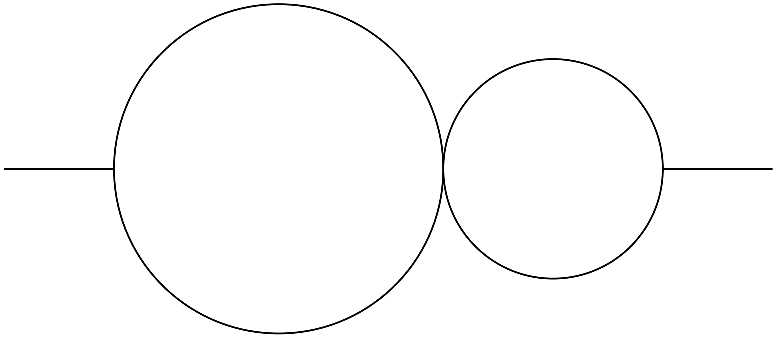
What is coevolution?

- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes
- indicates functional association or physical interaction



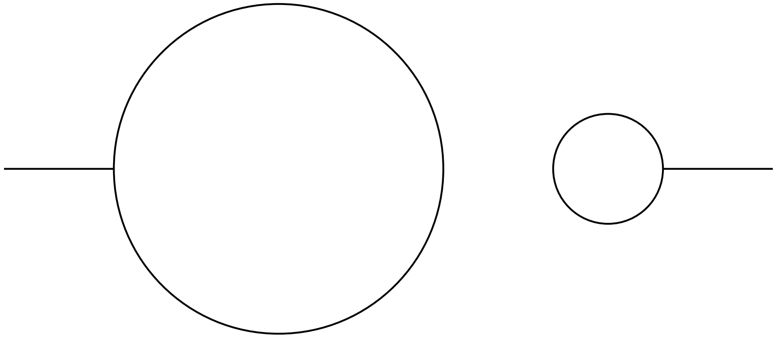
What is coevolution?

- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes
- indicates functional association or physical interaction



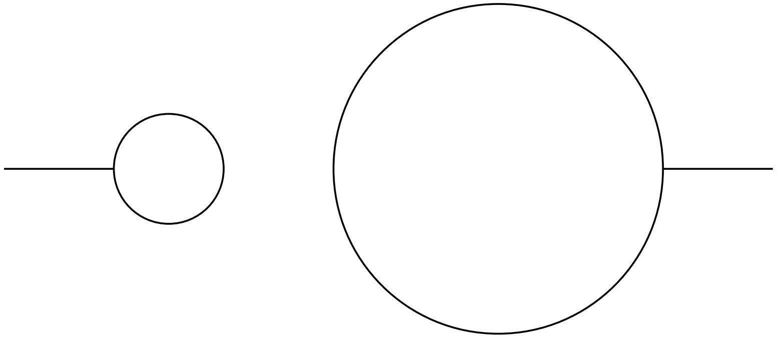
What is coevolution?

- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes
- indicates functional association or physical interaction



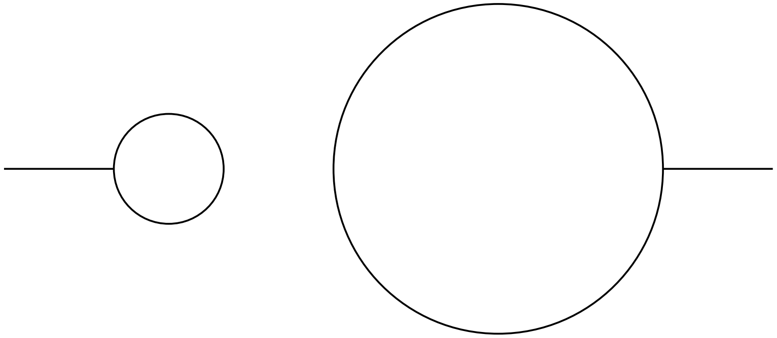
What is coevolution?

- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes
- indicates functional association or physical interaction



What is coevolution?

- when a “biological object” changes due to changes in another “biological object”
- correlated amino acid changes
- indicates functional association or physical interaction



- which correlations are meaningful?

Why do we care?

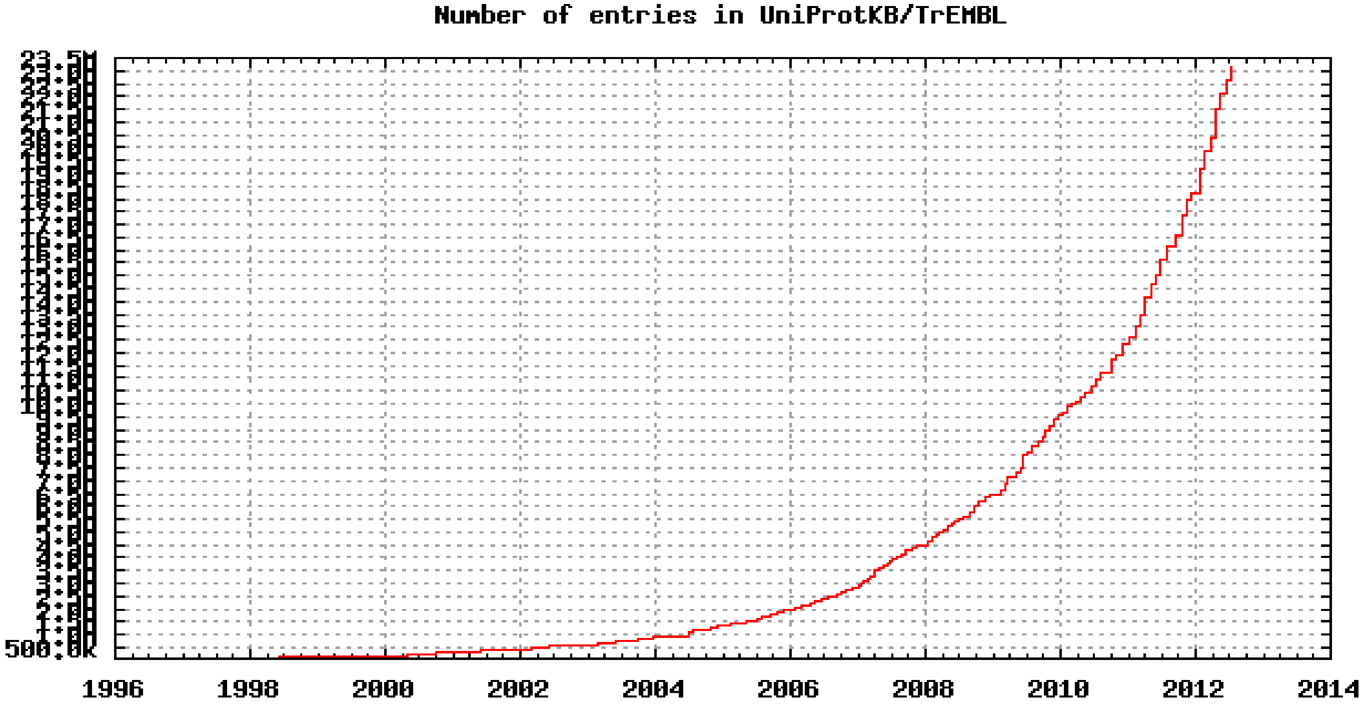
- coevolution is not currently used in Rosetta

Why do we care?

- coevolution is not currently used in Rosetta
- sequence based

Why do we care?

- coevolution is not currently used in Rosetta
- sequence based
- only requires sufficiently diverse sequences



http://www.ebi.ac.uk/uniprot/TrEMBLstats/relstat.html

Why do we care?

- coevolution is not currently used in Rosetta
- sequence based
- only requires sufficiently diverse sequences
- many possible applications in Rosetta

Why do we care?

- coevolution is not currently used in Rosetta
- sequence based
- only requires sufficiently diverse sequences
- many possible applications in Rosetta
  - decoy discrimination

Why do we care?

- coevolution is not currently used in Rosetta
- sequence based
- only requires sufficiently diverse sequences
- many possible applications in Rosetta
  - decoy discrimination
  - Abinitio

Why do we care?

- coevolution is not currently used in Rosetta
- sequence based
- only requires sufficiently diverse sequences
- many possible applications in Rosetta
  - decoy discrimination
  - Abinitio
  - docking

Why do we care?

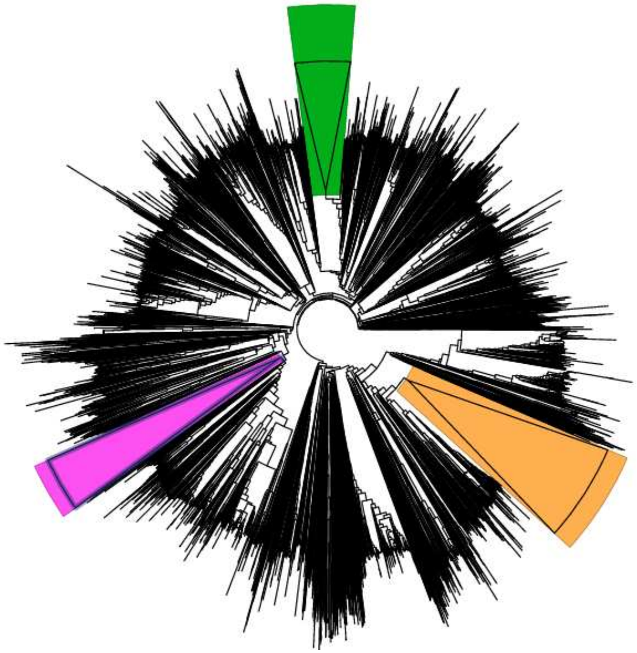
- coevolution is not currently used in Rosetta
- sequence based
- only requires sufficiently diverse sequences
- many possible applications in Rosetta
  - decoy discrimination
  - Abinitio
  - docking
  - ligand docking

Why now?

- coevolution is difficult to determine

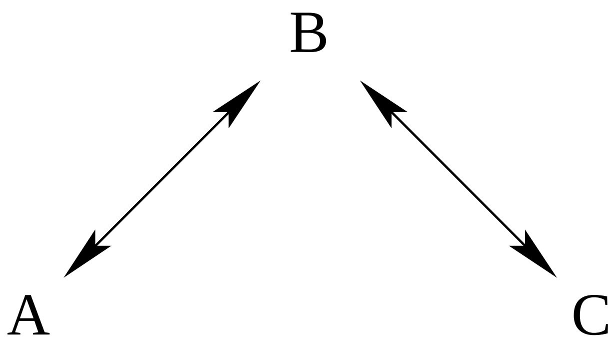
Why now?

- coevolution is difficult to determine
- sequence data are not independent



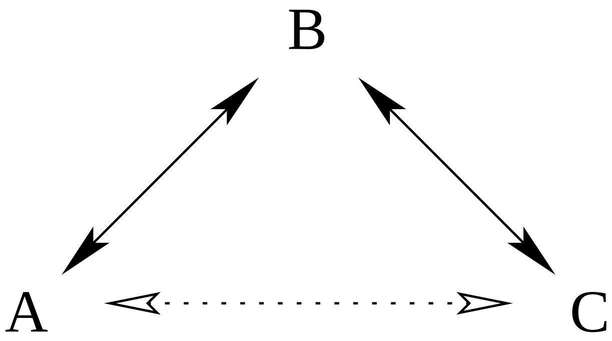
Why now?

- coevolution is difficult to determine
- transitivity of correlations



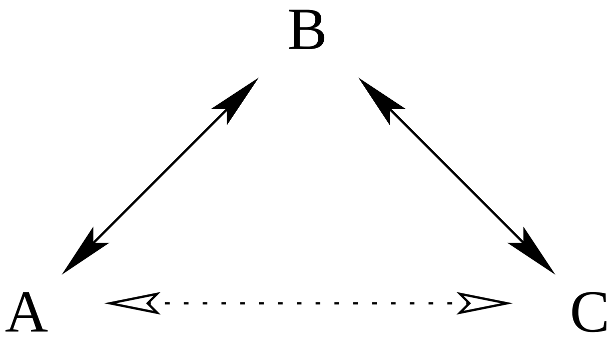
Why now?

- coevolution is difficult to determine
- transitivity of correlations



Why now?

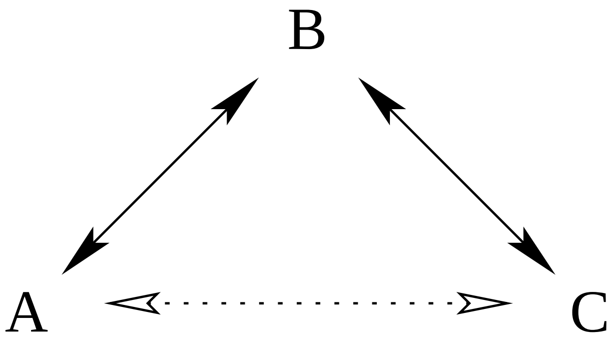
- coevolution is difficult to determine
- transitivity of correlations



- non-coding coevolution

Why now?

- coevolution is difficult to determine
- transitivity of correlations



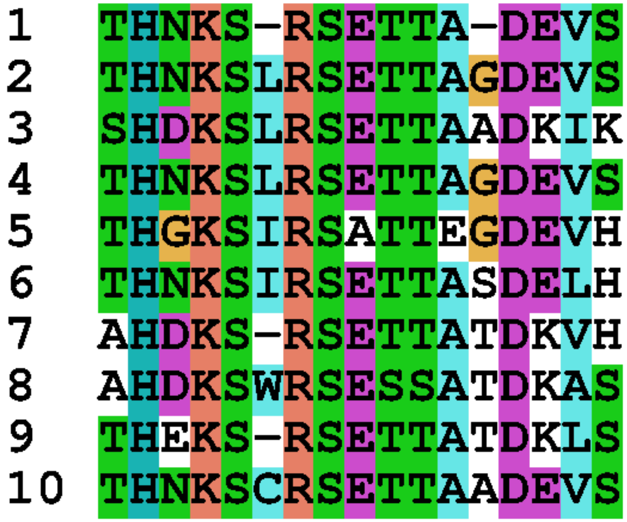
- non-coding coevolution
- new method: “Direct Information”

How can we detect coevolution?

- requires a diverse Multiple Sequence Alignment (MSA)

How can we detect coevolution?

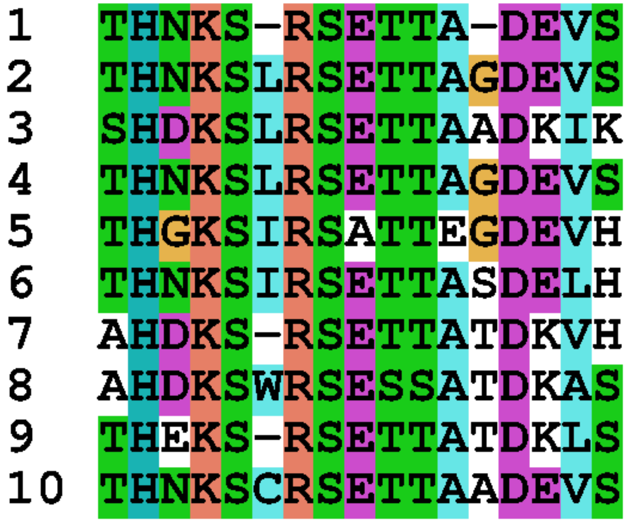
- requires a diverse Multiple Sequence Alignment (MSA)
- identify correlated pairs using “Direct Information” (DI)



```
1    THNKS-RSETTA-DEVS
2    THNKSLRSETTAGDEVS
3    SHDKSLRSETTAADKIK
4    THNKSLRSETTAGDEVS
5    THGKSIRSATTEGDEVH
6    THNKSIRSETTASDELH
7    AHDKS-RSETTATDKVH
8    AHDKSWRSESSATDKAS
9    THEKS-RSETTATDKLS
10   THNKSCRSETTAADEVS
```

How can we detect coevolution?

- requires a diverse Multiple Sequence Alignment (MSA)
- identify correlated pairs using “Direct Information” (DI)



```
1   THNKS-RSETTA-DEVS
2   THNKSLRSETTAGDEVS
3   SHDKSLRSETTAADKIK
4   THNKSLRSETTAGDEVS
5   THGKSIRSATTEGDEVH
6   THNKSIRSETTASDELH
7   AHDKS-RSETTATDKVH
8   AHDKSWRSESSATDKAS
9   THEKS-RSETTATDKLS
10  THNKSCRSETTAADEVS
```

- DI is the mutual information of a MSA specific distribution

Direct Information

MSA

Direct Information

1. extract site and pair frequencies from the MSA

$$\text{MSA} \rightarrow \text{frequencies}$$

Direct Information

1. extract site and pair frequencies from the MSA
2. downweight based on similarity* and incorporate pseudocounts

$$\text{MSA} \rightarrow \text{frequencies}$$

Direct Information

1. extract site and pair frequencies from the MSA
2. downweight based on similarity* and incorporate pseudocounts
3. determine the parameters of the constrained maximum entropy distribution ($P^{(dir)}$)
   use systematic small-coupling expansion to estimate parameters
   involves inversion of a connected correlation matrix

$$\text{MSA} \rightarrow \text{frequencies} \rightarrow P^{(dir)}$$

Direct Information

1. extract site and pair frequencies from the MSA
2. downweight based on similarity* and incorporate pseudocounts
3. determine the parameters of the constrained maximum entropy distribution ($P^{(dir)}$)
   use systematic small-coupling expansion to estimate parameters
   involves inversion of a connected correlation matrix
4. calculate the MI of $P^{(dir)}$ (the DI score)

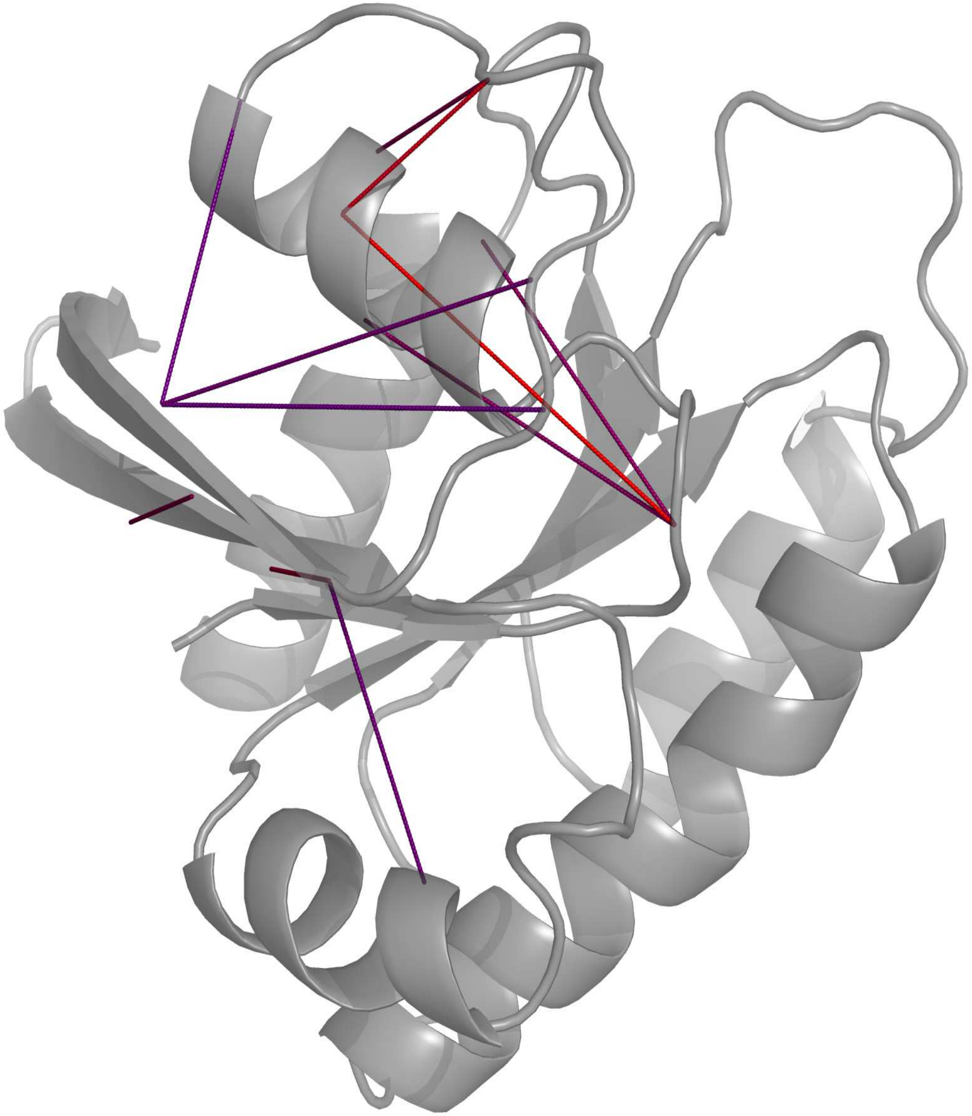
MSA $\rightarrow$ frequencies $\rightarrow$ $P^{(dir)}$ $\rightarrow$ DI

Direct Information

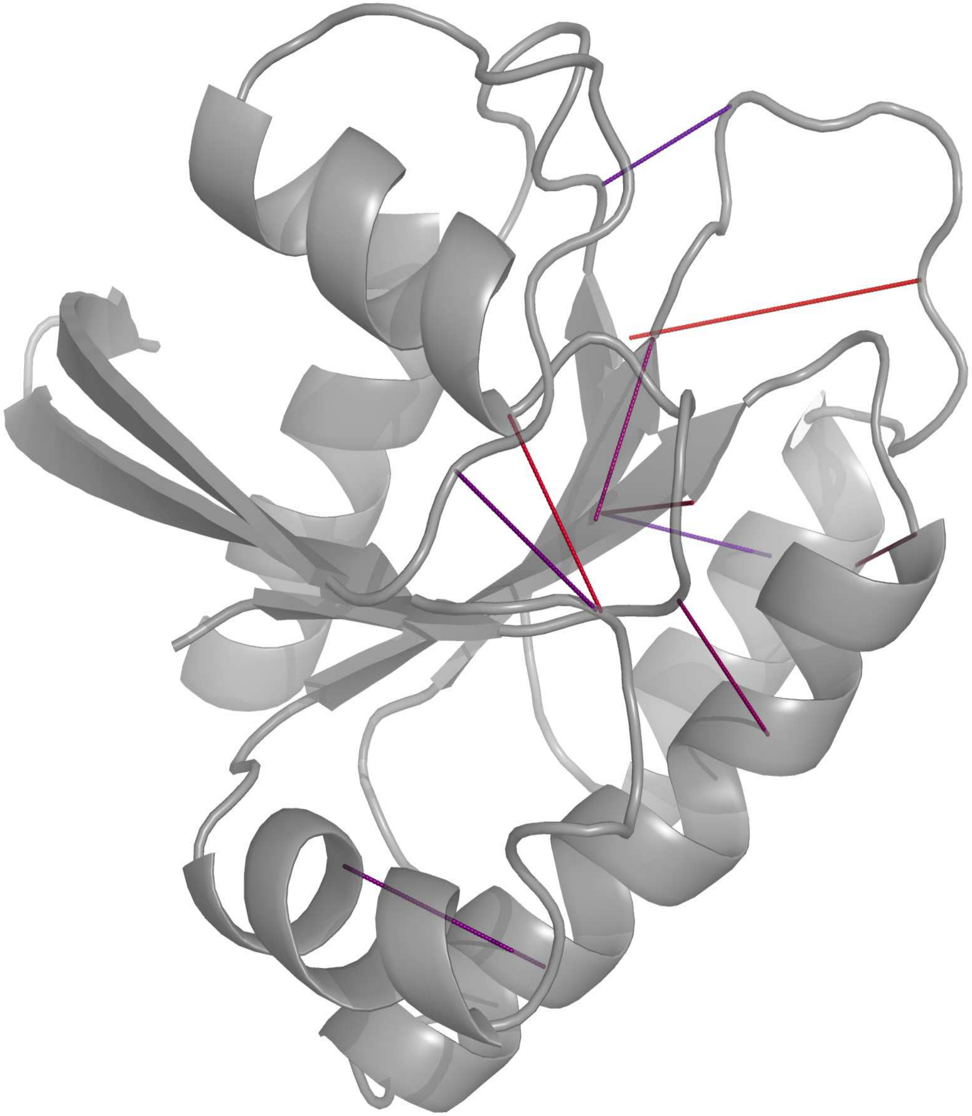
1. extract site and pair frequencies from the MSA
2. downweight based on similarity* and incorporate pseudocounts
3. determine the parameters of the constrained maximum entropy distribution ($P^{(dir)}$)
   use systematic small-coupling expansion to estimate parameters
   involves inversion of a connected correlation matrix
4. calculate the MI of $P^{(dir)}$ (the DI score)
5. identify correlated pairs

$$\text{MSA} \rightarrow \text{frequencies} \rightarrow P^{(dir)} \rightarrow \text{DI} \rightarrow \text{pairs}$$

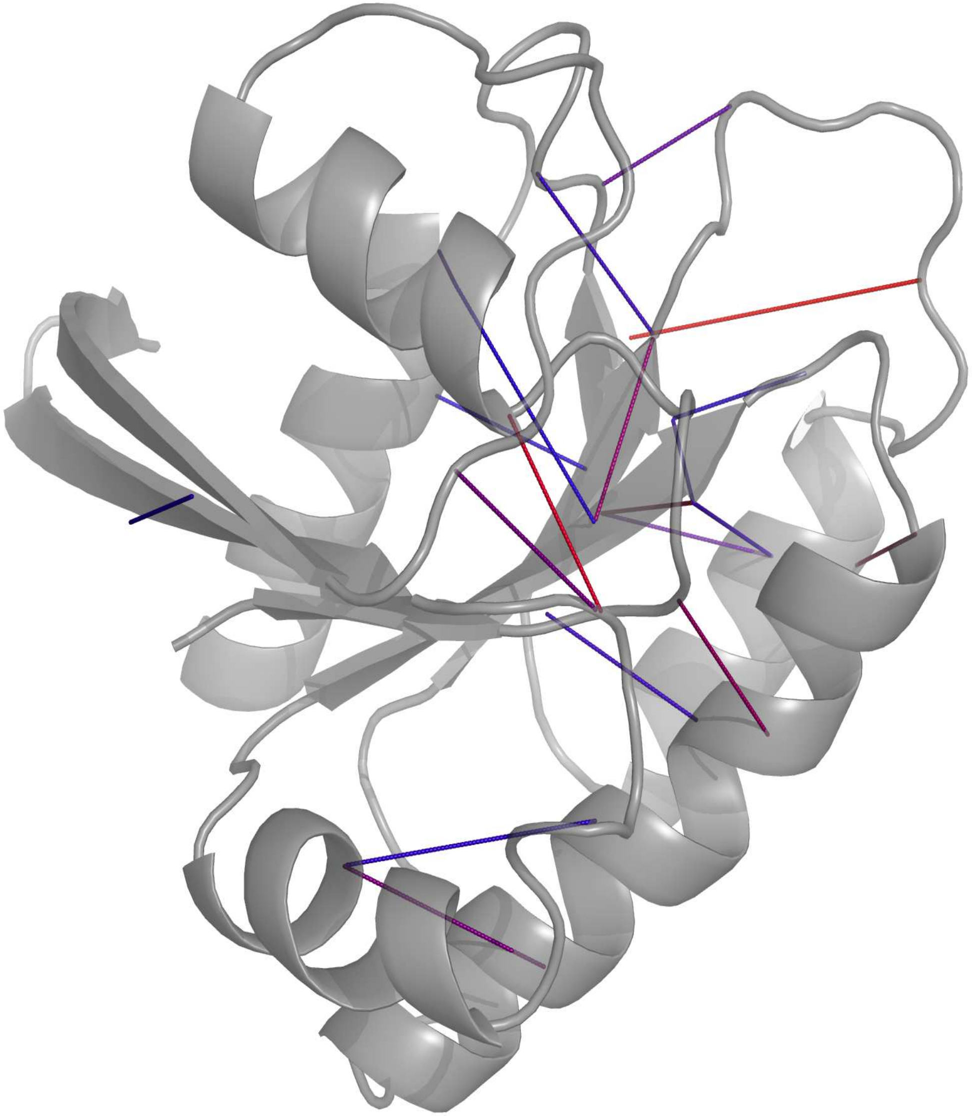
MI



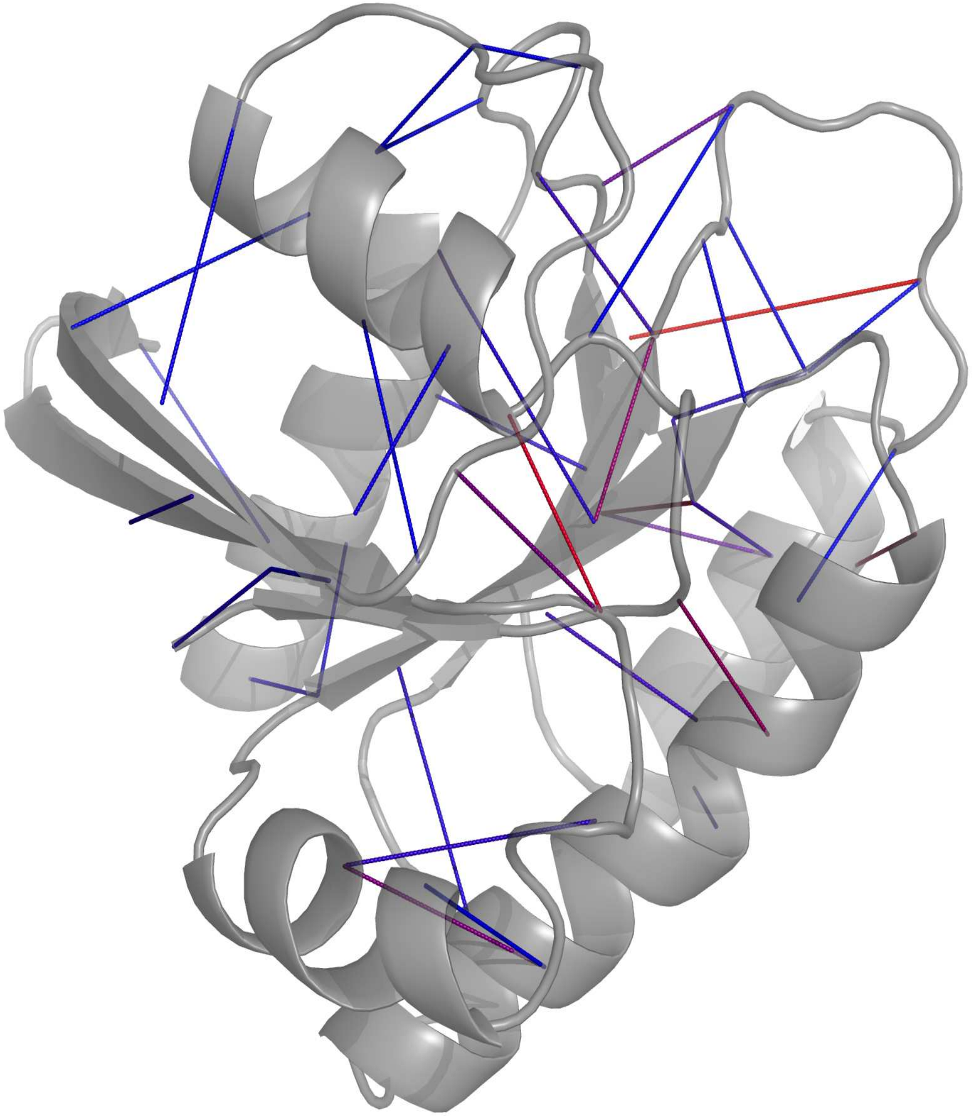
DI



DI



DI

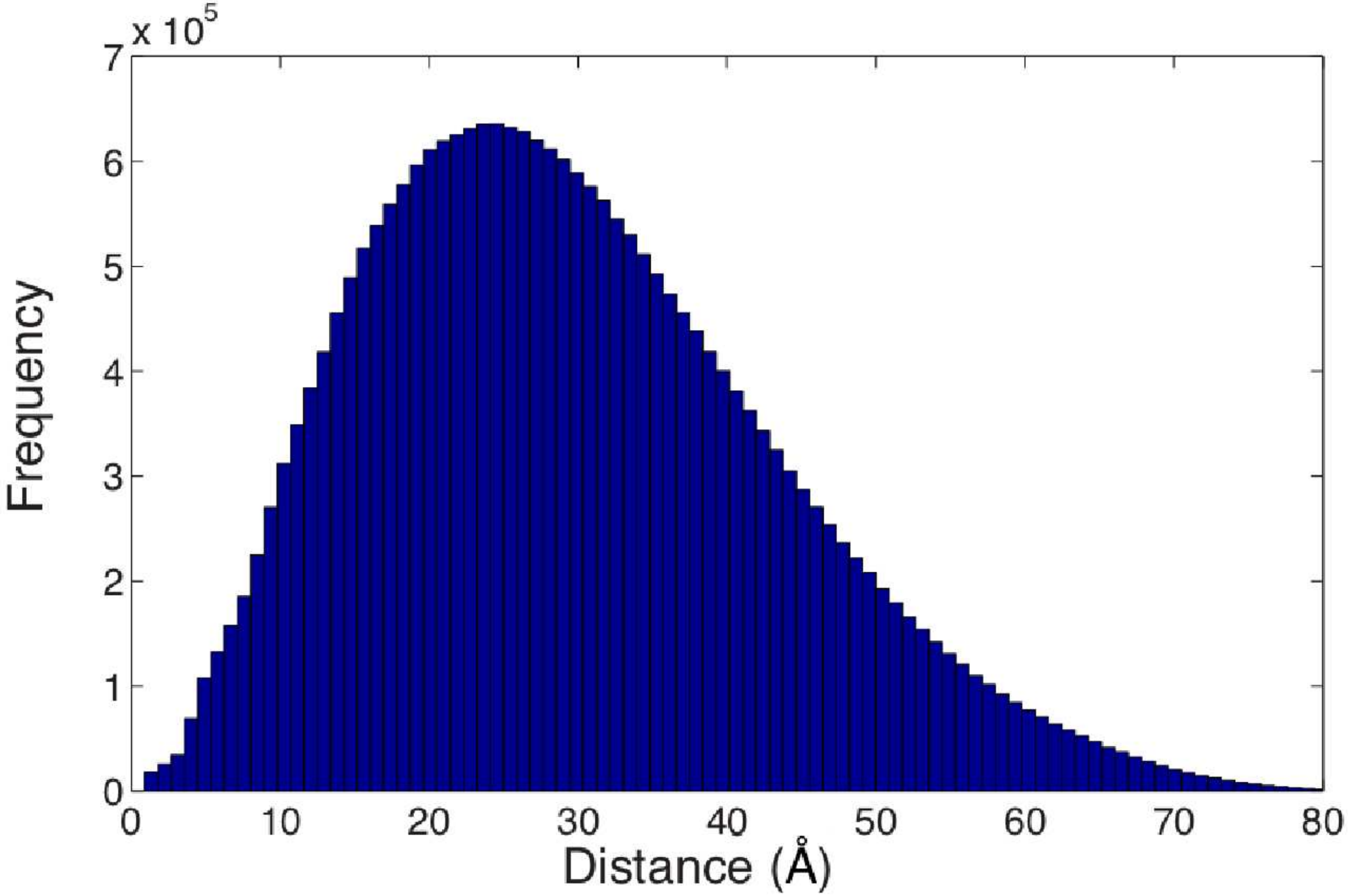


Does it work?

- assume: DI pairs are physically close

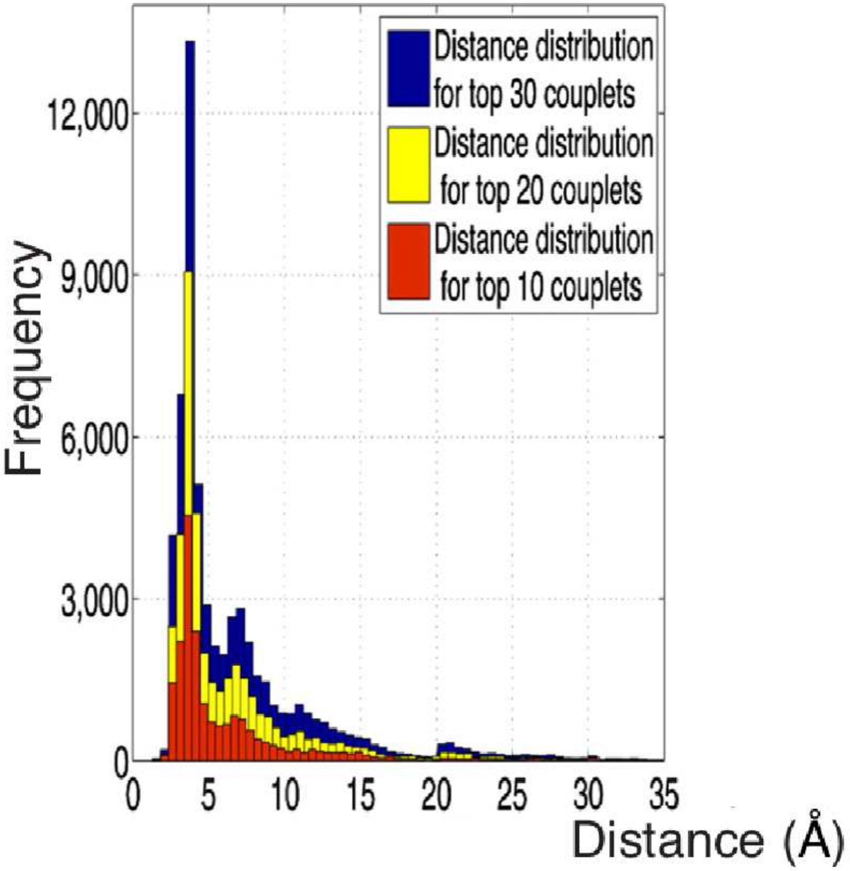
Does it work?

- assume: DI pairs are physically close



Does it work?

- assume: DI pairs are physically close

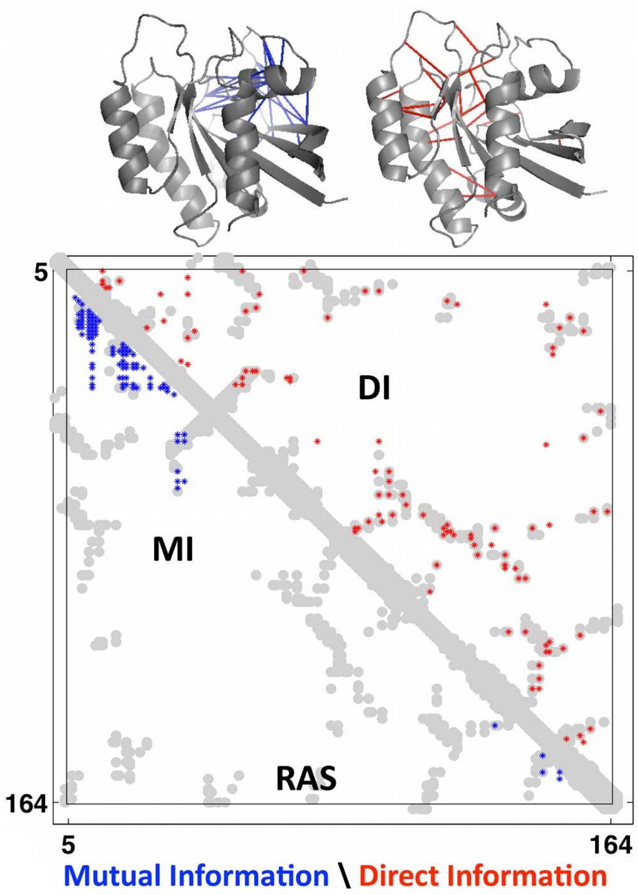


What does it detect?

- identifies interdependent sites

What does it detect?

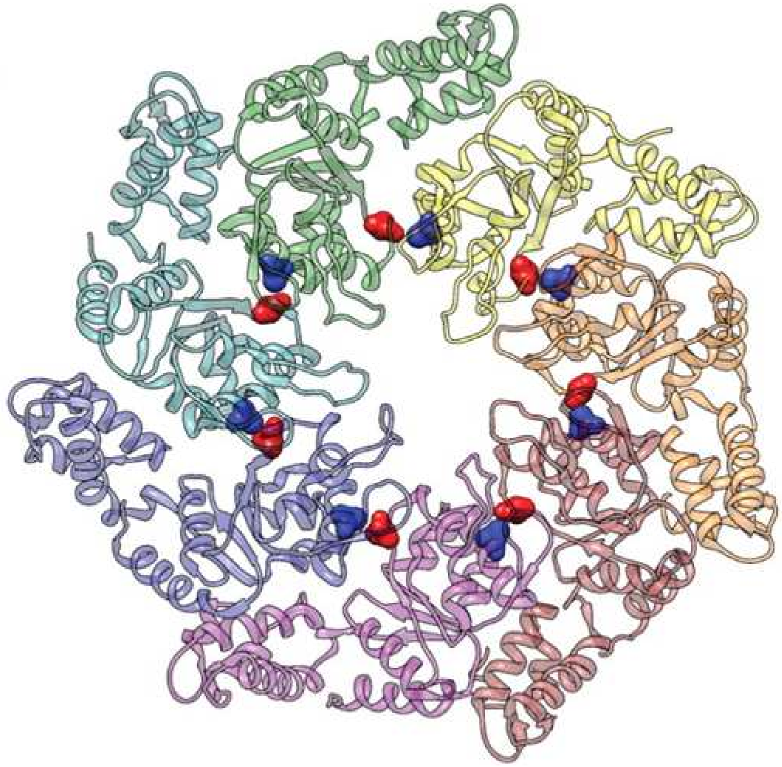
- identifies interdependent sites
- many of the best DI pairs are native-state contacts



Marks *et al.* 2011, PDB 1NY6, 1JE8, 3C3W – p. 9

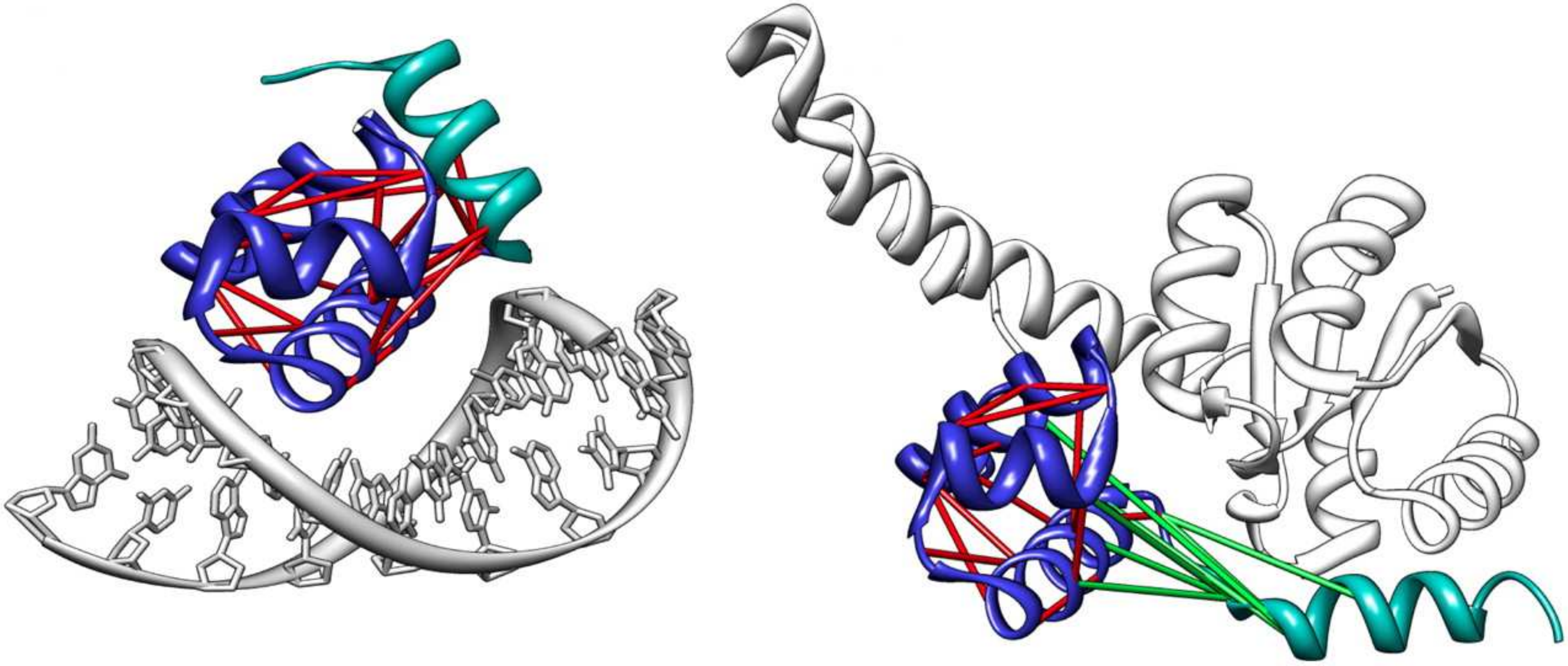
What does it detect?

- identifies interdependent sites
- many of the best DI pairs are native-state contacts
- some pairs are homomeric contacts



What does it detect?

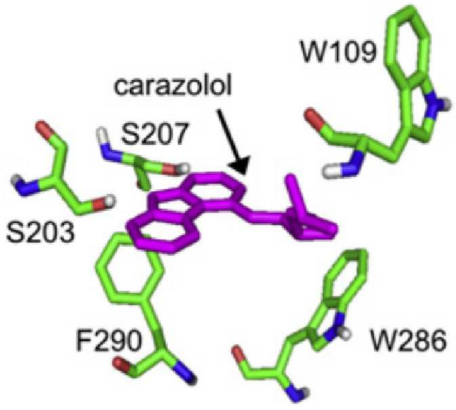
- identifies interdependent sites
- many of the best DI pairs are native-state contacts
- some pairs are non-native state contacts

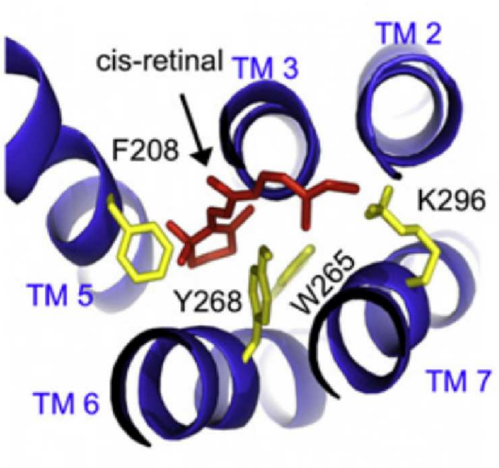


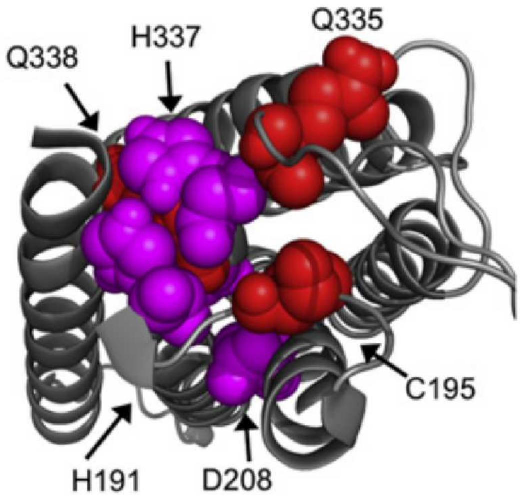
Marks *et al.* 2011, PDB 1NY6, 1JE8, 3C3W

What does it detect?

- identifies interdependent sites
- many of the best DI pairs are native-state contacts
- some pairs are binding interfaces







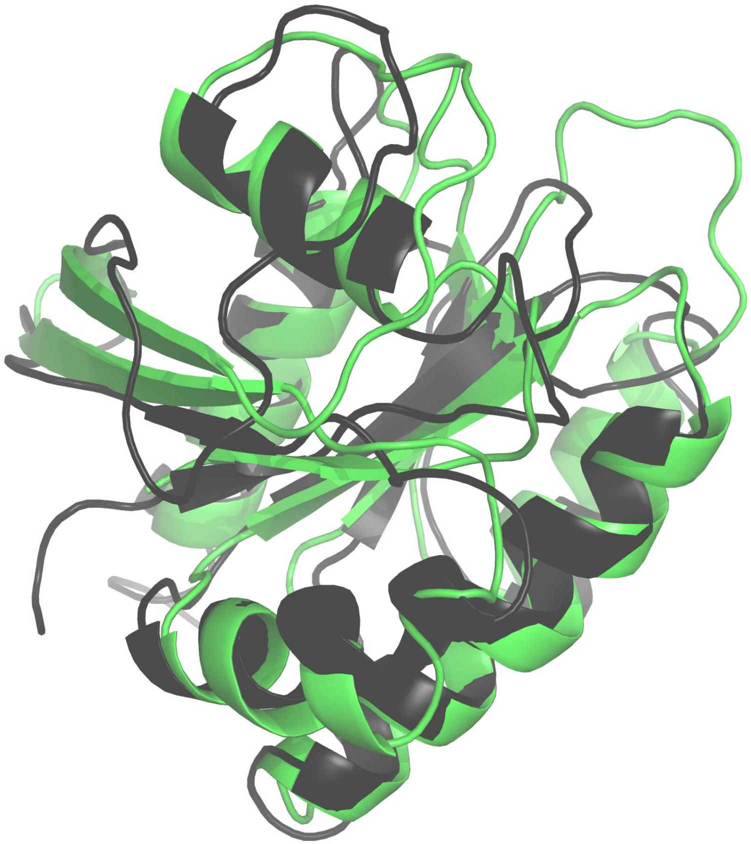
Marks *et al.* 2011, PDB 1NY6, 1JE8, 3C3W – p. 9

Is it useful?

- associated structure prediction algorithm EVfold

Is it useful?

- associated structure prediction algorithm EVfold
- constrained Crystallography & NMR Systems suite (CNS)



Is it useful?

- associated structure prediction algorithm EVfold
- constrained Crystallography & NMR Systems suite (CNS)
- use constraints derived from:
  - DI pairs

Is it useful?

- associated structure prediction algorithm EVfold
- constrained Crystallography & NMR Systems suite (CNS)
- use constraints derived from:
  - DI pairs
  - secondary structure prediction

Is it useful?

- associated structure prediction algorithm EVfold
- constrained Crystallography & NMR Systems suite (CNS)
- use constraints derived from:
  - DI pairs
  - secondary structure prediction
  - topology prediction (for transmembrane proteins)

Is it useful?

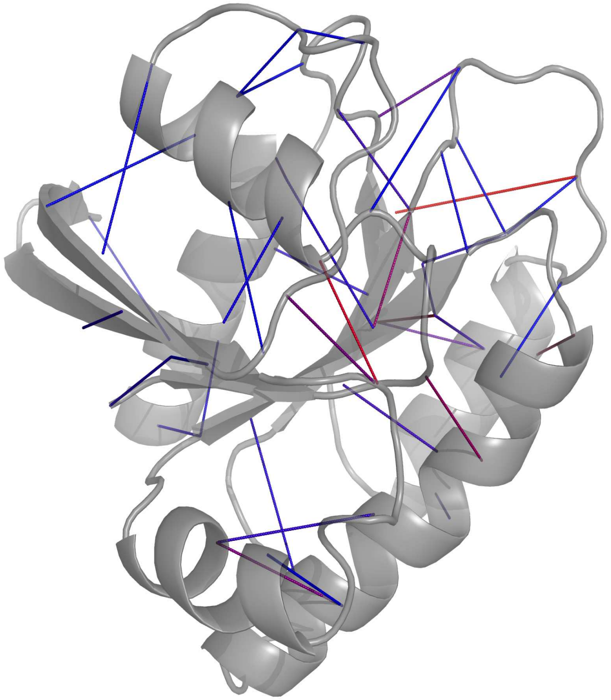
- associated structure prediction algorithm EVfold
- constrained Crystallography & NMR Systems suite (CNS)
- use constraints derived from:
  - DI pairs
  - secondary structure prediction
  - topology prediction (for transmembrane proteins)
- complicated...

How do we use it?

- use Rosetta AtomPairConstraints

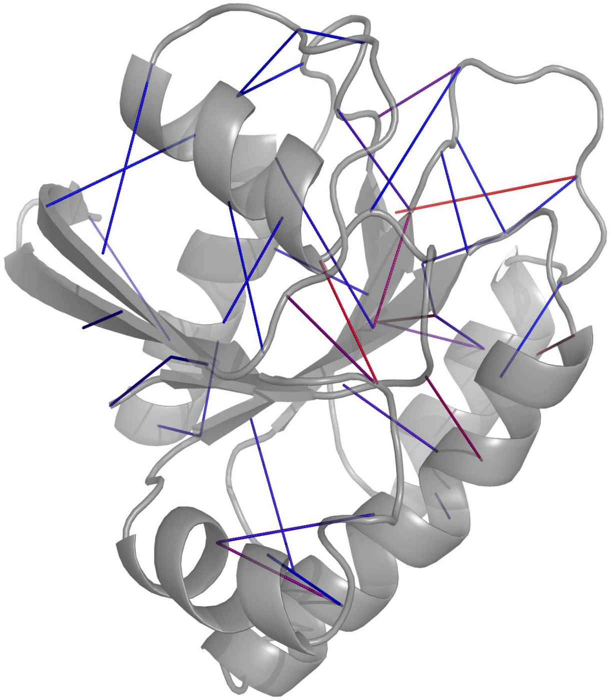
How do we use it?

- use Rosetta AtomPairConstraints
- can coevolution constraints improve Abinitio?



How do we use it?

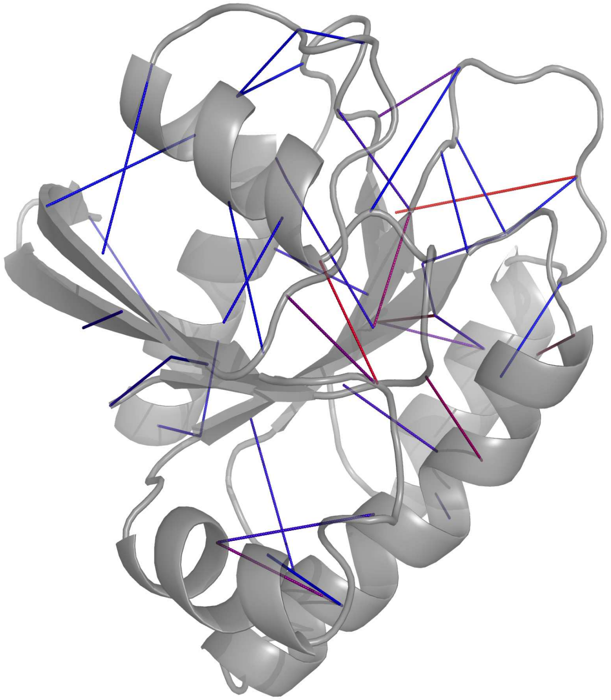
- use Rosetta AtomPairConstraints
- can coevolution constraints improve Abinitio?



- challenge: some DI pairs are not physically close

How do we use it?

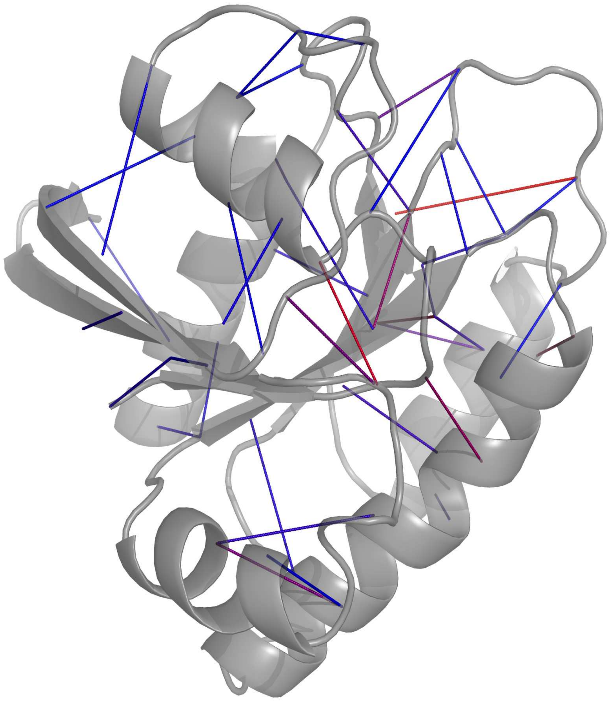
- use Rosetta AtomPairConstraints
- can coevolution constraints improve Abinitio?



- challenge: may require many constraints

How do we use it?

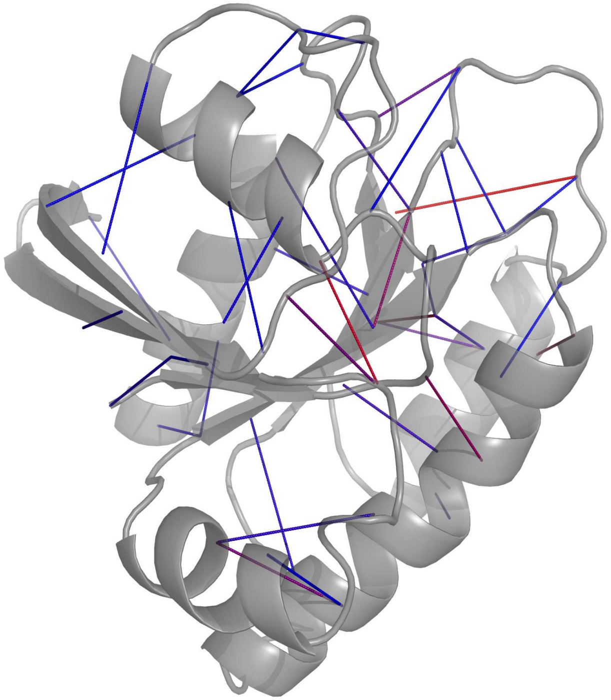
- use Rosetta AtomPairConstraints
- can coevolution constraints improve Abinitio?



- challenge: undetermined weight in Rosetta scoring

How do we use it?

- use Rosetta AtomPairConstraints
- can coevolution constraints improve Abinitio?



- challenge: many constraint scores are quadratic

Can we keep it simple?

- original authors wanted the “distance between Cα atoms...less than 7Å, set as a harmonic constraint at 4Å”

Can we keep it simple?

- original authors wanted the “distance between Cα atoms...less than 7Å, set as a harmonic constraint at 4Å”
- try 10 best scoring DI pairs using the Rosetta HARMONIC scoring

Can we keep it simple?

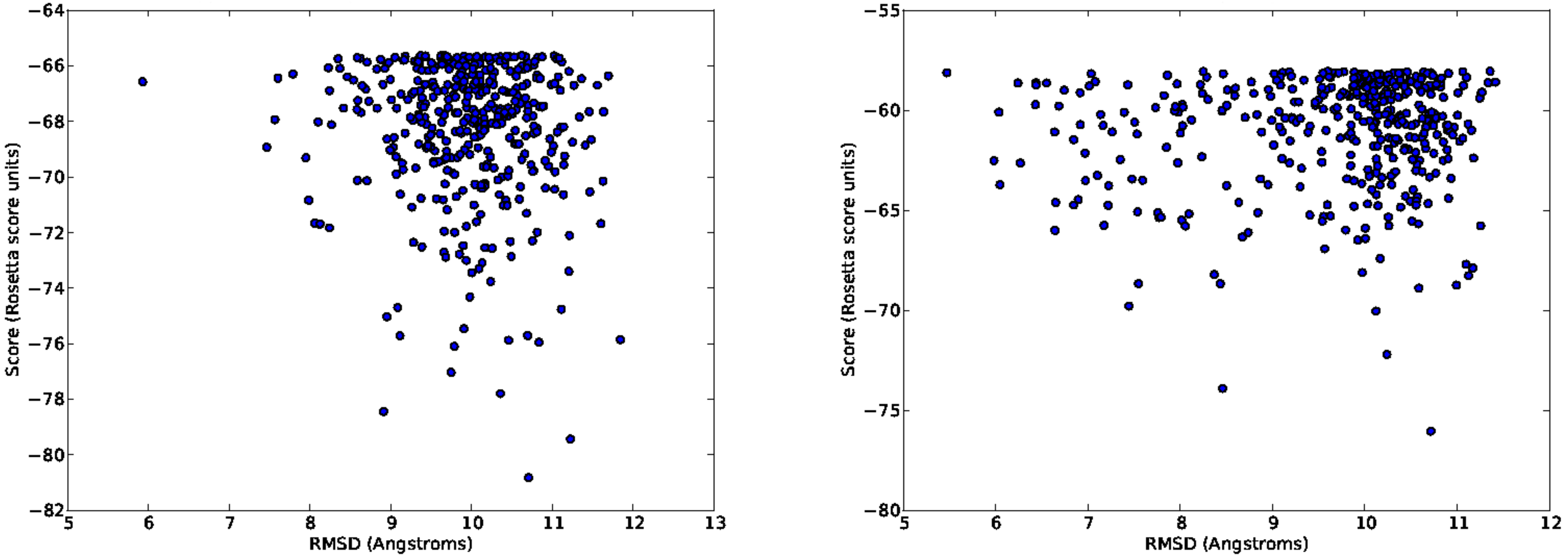
- original authors wanted the “distance between Cα atoms...less than 7Å, set as a harmonic constraint at 4Å”
- try 10 best scoring DI pairs using the Rosetta HARMONIC scoring
- apply constraints during centroid stages of AbinitioRelax (20000 decoys)

Can we keep it simple?

- original authors wanted the “distance between Cα atoms...less than 7Å, set as a harmonic constraint at 4Å”
- try 10 best scoring DI pairs using the Rosetta HARMONIC scoring
- apply constraints during centroid stages of AbinitioRelax (20000 decoys)
- compare the score v. RMSD plots with and without constraints for cluster centers of the lowest 400 scoring decoys

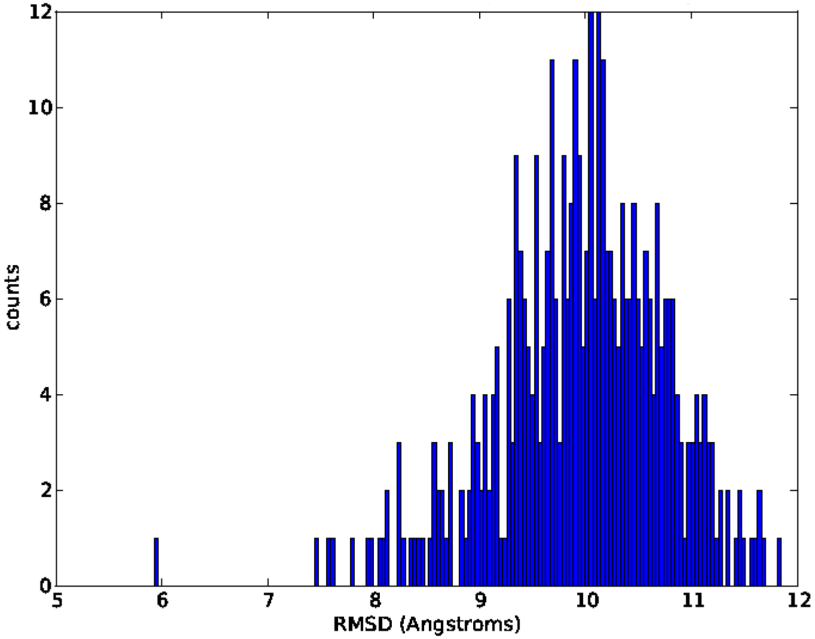
Can we keep it simple?

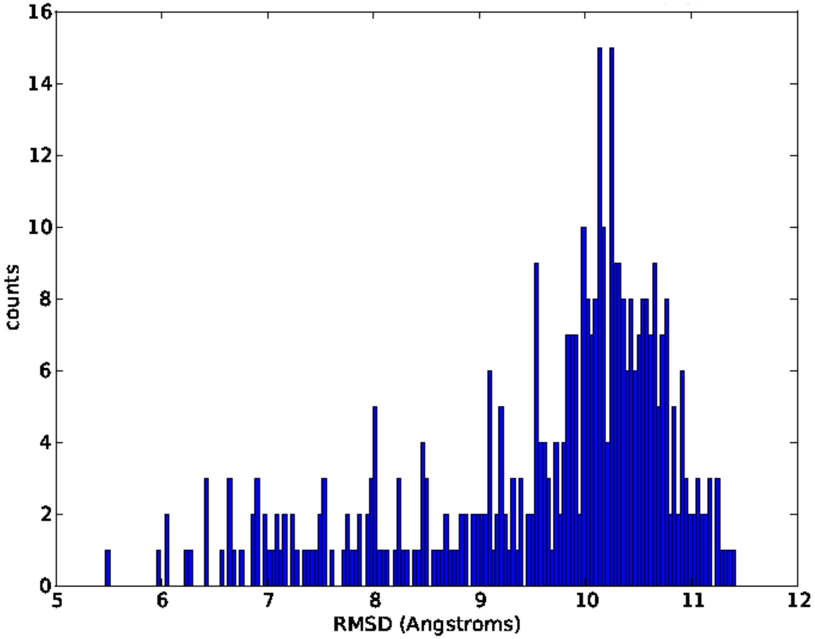
unconstrained constrained



Can we keep it simple?

unconstrained constrained





Where are we now?

- how do we properly score these constraints?
- can we classify the interactions identified by DI?
- can DI pairs indicate near-native decoys?
- do DI constraints improve AbinitioRelax?
- can DI constraints improve other protocols?

Thanks for listening!

Rich Bonneau

Kevin Drew

Doug Renfrew

Noah Youngs

Duncan Penfold-Brown

Glenn Butterfoss

Timothy Craven

Abba Leffler

Rebecca Alford

Leif Halvorson

Chris Poultney

The Bonneau Lab

Debora Marks - Harvard Medical School, Dept. of Systems Biology

Chris Sanders - Memorial Sloan-Kettering Cancer Center

Lucy Colwell - MRC Laboratory of Molecular Biology

The Rosetta Community

PyRosetta Team