Protein docking by sequence and structure similarity

Petras Kundrotas

Center for Computational Biology
The University of Kansas
Lawrence, Kansas, USA

pkundro@ku.edu

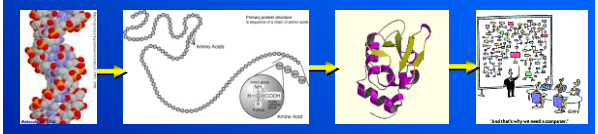
Introduction to Bioinformatics course, Stockholm University, February 2, 2017

Protein docking by sequence and structure similarity

- Introduction and overview
- Docking by sequence similarity
  - Availability of templates
  - Benchmark on X-ray structures
- Docking by structure similarity
  - Availability of templates
  - Benchmark on X-ray structures
  - Benchmark on model structures
- Docking enhancements
  - GO-score
  - Text mining

Where are we now?

Post-genomic era



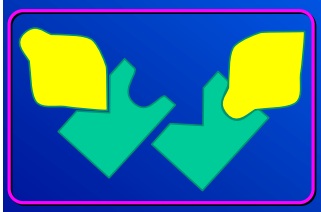
DNA sequence	protein amino acid sequences (primary structure)	3D structures of proteins (secondary and tertiary structures)	3D structures of protein complexes (quaternary structure)
completed	completed	will be completed soon	?

Predicting protein-protein interactions

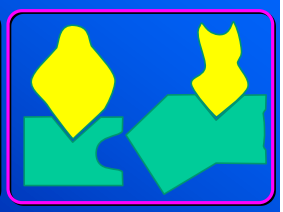
- Do sequences A and B interact ?
- Given that sequences A and B interact what is 3D structure of AB complex (docking)?

Modeling protein-protein interactions

Structure of monomers vs structure of a complex (binding mode)



- Structures of the monomers similar, binding mode different



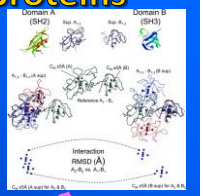
- Structures of the monomers different, binding mode similar

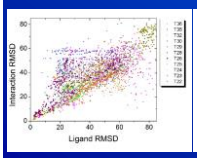
Comparing binding modes of two proteins

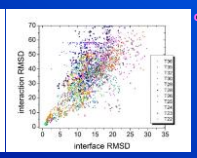
- Dissimilar proteins
  - Interaction RMSD, *i*RMSD

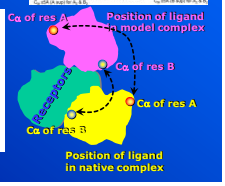
From: Aloy et al., JMB (2003), 332, 989

- Similar proteins (model vs native)
  - Ligand RMSD, *L*-RMSD
  - Interface RMSD, *i*-RMSD









Structure comparison

- Root-mean-square distance (RMSD)

Protein A: $a_1, a_2, \ldots, a_N$ Protein B: $b_1, b_2, \ldots, b_M$

$$\mathrm{RMSD}(A,B) = \sqrt{\frac{\sum_{i=1}^{N} d^2\left(a_i, f\left(R(b_k \rightarrow a_i)\right)\right)}{N}}$$

Spatial transformation

Correspondence between A and B

$$\frac{2+2}{2} = 2$$

$$\sqrt{\frac{2^2+2^2}{2}} = 2$$

$$\frac{3+1}{2} = 2$$

$$\sqrt{\frac{3^2+1^2}{2}} \approx 2.2$$

Alignment of sequences

- Rearranging sequences to infer regions of similarity (1D problem)

```
ADRLYGGTWQQAGW
ADGGAATQQWKL
```

- Different lengthes
- Some are more equal than others
  - Not only exact matches
  - Certain amino acid pairs are preferable
- Extended regions of evolutionary deletion/insertion

```
ADRLYGGTW-QQW--AGW
AD---GGAATQQWKL---
```

Alignment of structures

- Structure is more conserved than sequences
- Rearranging structures to infer regions of similarity (3D problem)

1LH1, leghemoglobin from yellow lupin

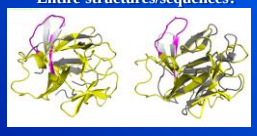
1URV, cytoglobin from human

- Can be used for functional assignment

Key questions are the same

- What do we want to align?

Entire structures/sequences?



Most relevant parts of structures/sequences?

- How to find the best alignment?
  - Structure/sequence representation and feature extraction
  - Structure/sequence comparison and alignment optimization
- How do we score an alignment?

Optimal alignment

- Maximize value of an objective function
- Sequence alignment
  - Similarity of amino acids

    Physical and chemical properties, sterical properties, etc.
- Structure alignment
  - Inverse distance between atoms of two molecules

Dynamical programming for structural alignment

Structure 1: HSRRRHVF Structure 2: GQVGMAC

- Scoring matrix

Sequence 1

Sequence 2	H	S	R	R	R	H	V	F
G								
Q								
V								
G								
M								
A								
C								

Distance between two residues is dependent on the alignment of other residues

Double dynamical programing

Structure 1: HSRRRHVF Structure 2: GQVGMAC

- **Lower level dynamical programming**

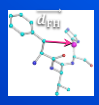
$C_\beta - C_\beta$ vectors from F to

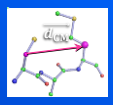
$C_\beta - C_\beta$ vectors from C to

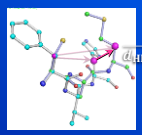
	H	S	R	R	R	H	V	F
G	12	2	3					
Q	1	1	10	1				
V		0	2	1	0			
G			1	23	1	0		
M				1	7	4	1	
A					0	2	14	1
C						0	1	25

$$s_{ij}(\mathrm{FC}) = \frac{a}{d_{ij}^2(\mathrm{FC}) + b}$$

$$\overrightarrow{d_{ij}}(\mathrm{FC}) = \overrightarrow{d_{\mathrm{F}i}} - \overrightarrow{d_{\mathrm{C}j}}$$







Taylor & Orengo, JMB (1989) 208, 1

Double dynamical programing

Structure 1: HSRRRHVF Structure 2: GQVGMAC

- **Lower level dynamical programming**

$C_\beta - C_\beta$ vectors from F to

$C_\beta - C_\beta$ vectors from C to

	H	S	R	R	R	H	V	F
G	12	2	3					
Q	1	1	10	1				
V		0	2	1	0			
G			1	23	1	0		
M				1	7	4	1	
A					0	2	14	1
C						0	1	25

$C_\beta - C_\beta$ vectors from X to

$C_\beta - C_\beta$ vectors from C to

	H	S	R	R	R	H	V	F
G	16	1	2					
Q	1	21	1	1				
V		1	4	0	0			
G			5	4	1	1		
M				4	5	1	1	
A					2	15	1	0
C						1	25	1

...

$$s_{ij}(\mathrm{XZ}) = \frac{a}{d_{ij}^2(\mathrm{XZ}) + b}$$

$$\overrightarrow{d_{ij}}(\mathrm{XZ}) = \overrightarrow{d_{\mathrm{X}i}} - \overrightarrow{d_{\mathrm{Z}j}}$$

Σ

Sequence 2

Sequence 1

	H	S	R	R	R	H	V	F
G	28							
Q		21	10					
V			4					
G				27				
M					12			
A						15	14	
C							25	25

- **Higher level dynamical programming**

Everything is about the scoring matrix!

Protein docking

- **Template free (ab initio) docking**
  - Shape complementarity (Lennard-Jones potential)
  - Enhancements (physics- and knowledge-based)
- **Template-based (homology) docking**
  - Sequence similarity
  - Structure similarity
    - Full structures
    - Local structures (interfaces)

Template-based docking

- **Detection of templates**
  - Threshold of global/local sequence/structure similarity
- **Quality of models**
  - Ligand/Interface RMSD
  - Predicted contacts/residues
  - CAPRI criteria
    - Acceptable quality : F_{nat} > 0.1 *and* (L-RMSD < 10 Å *or* *i*-RMSD < 4 Å)
    - Medium quality : F_{nat} > 0.3 *and* (L-RMSD < 5 Å *or* *i*-RMSD < 2 Å)
    - High quality : F_{nat} > 0.3 *and* (L-RMSD < 5 Å *or* *i*-RMSD < 2 Å)
- **Ranking of templates/models**
  - By global/local sequence/structure similarity
  - By adding "enhancements"
- **Success criteria for a target**
  - At least one "good" model in top N predictions

Docking by sequence similarity

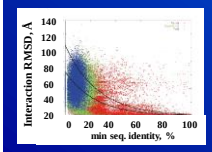
Main idea

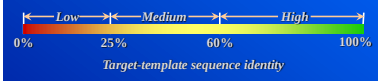
- **Monomer proteins**

Two sequences are similar (homologous) → Their structures are similar too

- **Protein-protein complexes**

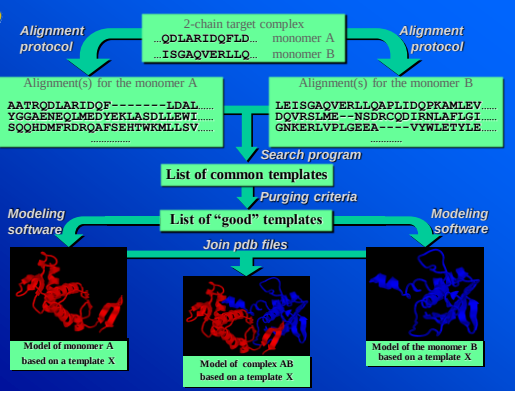
Two interacting sequences A and B are similar to two other sequences A′ and B′ forming a complex → Structure of AB complex is similar to structure of A′ B′ complex.





Aloy et al., JMB (2003), 332, 989; *Curr Opin Struct Biol, (2002) 15, 15-22*

Docking by sequence similarity

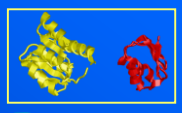


Docking by sequence similarity

Purging list of common templates

Initial list of common templates

```
1eer, chains A, T
1a8u, chains C, D
1acb, chains A, A
2dfa, chains B, B
```



Purging

```
1eer, chains A, T
1a8u, chains C, D
1acb, chains A, A
2dfa, chains B, B
```

Chains do not form a complex

Large overlap of alignments

```
1 AATRQDLARIDQF---
  YGGAENEQLMEDYERL

2 RDRQAFSEHTWKMLLSV
  QLMEDYERLASDLLEWI
```

List of "good" templates

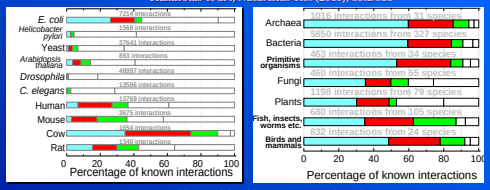
```
1a8u, chains C, D
2dfa, chains B, B
```

Docking by sequence similarity

Availability of templates

- **GWIDD structural coverage**

Kundrotas et al., Nucl.Acids Res (2010), 38:D513



PDB structures

Docking models with sequence templates from the same organism

Docking models with sequence templates from a different organism

http://gwidd.compbio.ku.edu

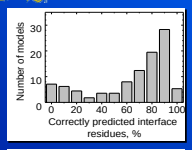
- **Benchmark studies**

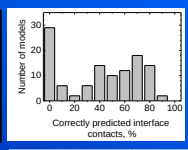
Kundrotas & Alexov (2006) BBA, 1764: 1498 ; (2008) Int. J Biolog. Macromol. 43:198

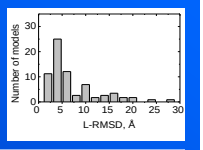
- 74 out of 463 (~16%) complexes with templates

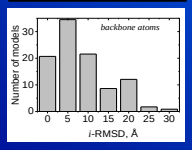
Docking by sequence similarity

Quality of resulting 116 models for 74 complexes



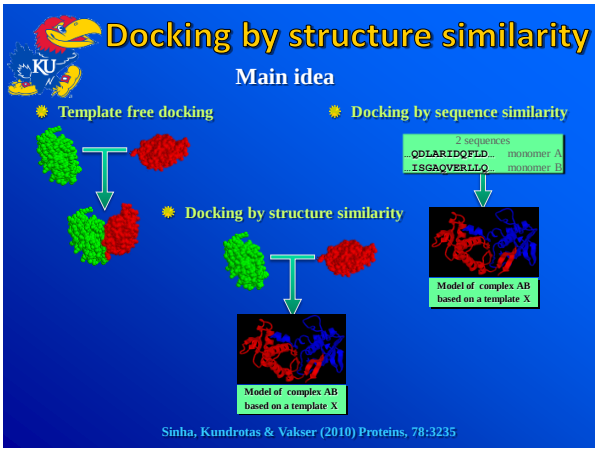


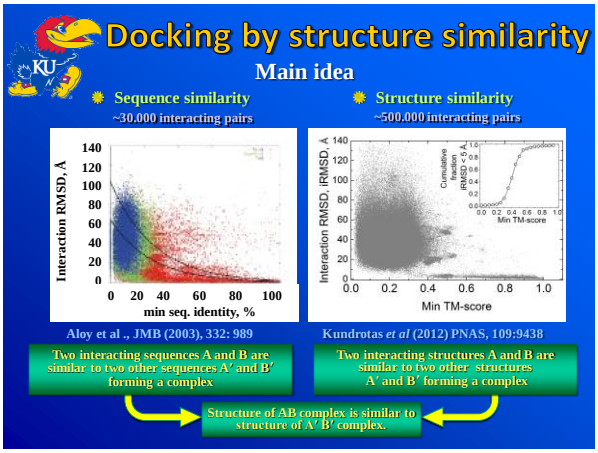


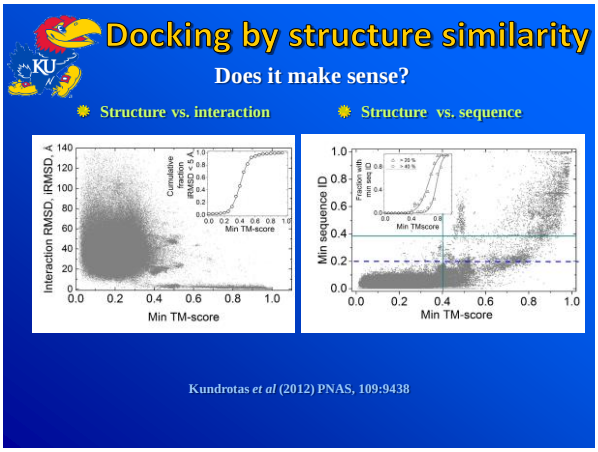


CAPRI classification	% of total number of models	
	BLAST	PROFILES
Medium	20.6	18.1
Acceptable	45.8	45.7
Incorrect	33.6	36.2

Kundrotas & Alexov (2008) Int. J Biolog. Macromol. 43, 198


Docking by structure similarity
Main idea
Template free docking
Docking by sequence similarity
2 sequences
..QDLARIDQFLD.. monomer A
..ISGAQVERLLQ.. monomer B
Docking by structure similarity
Model of complex AB based on a template X
Model of complex AB based on a template X
Sinha, Kundrotas & Vakser (2010) Proteins, 78:3235

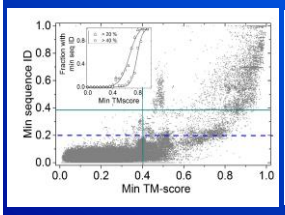

Docking by structure similarity
Main idea
Sequence similarity
~30,000 interacting pairs
Structure similarity
~500,000 interacting pairs
Interaction RMSD, Å
min seq. identity, %
Interaction RMSD, iRMSD, Å
Min TM-score
Aloy et al ., JMB (2003), 332: 989
Kundrotas et al (2012) PNAS, 109:9438
Two interacting sequences A and B are similar to two other sequences A′ and B′ forming a complex
Two interacting structures A and B are similar to two other structures A′ and B′ forming a complex
Structure of AB complex is similar to structure of A′ B′ complex.

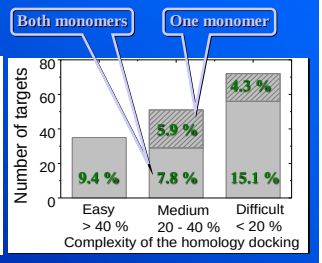

Docking by structure similarity
Does it make sense?
Structure vs. interaction
Structure vs. sequence
Interaction RMSD, iRMSD, Å
Min TM-score
Min sequence ID
Kundrotas et al (2012) PNAS, 109:9438

Docking by structure similarity

Does it make sense?

- Structure vs. sequence
- Benchmark studies





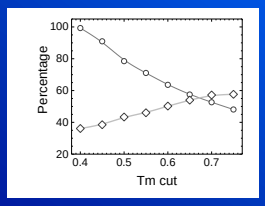
Kundrotas *et al* (2012) PNAS, 109:9438

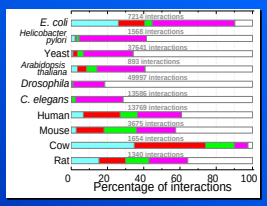
Sinha, Kundrotas & Vakser (2010) Proteins, 78:3235

Docking by structure similarity

Availability of templates

- Structure vs. interaction
- GWIDD structural coverage





Kundrotas *et al* (2012) PNAS, 109:9438

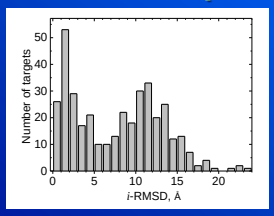
Kundrotas et al ., Human Genomics (2012), 6:7

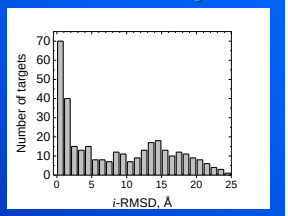
Templates are available for *almost all* protein-protein complexes provided that structures of interactors are known or can be modeled

Docking by structure similarity

Quality of resulting models

- Hetero-dimeric complexes
- Homo-dimeric complexes





- 372 targets
- 11932 templates (similar excluded)
- 145 successful targets (39%)

best model with *i*-RMSD < 5 Å

- 352 targets (2011 – 2012)
- 2148 templates (before 2011)
- 153 successful targets (43.5%)

best model with *i*-RMSD < 5 Å

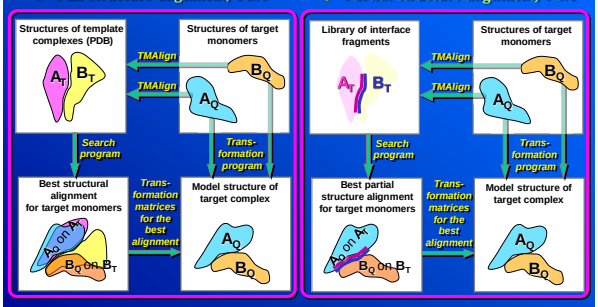
Sinha, Kundrotas & Vakser (2010) Proteins, 78:3235

Kundrotas, Vakser & Janin (2013) Protein Sci., 22:1655

Docking by structure similarity

Two flavors

- Full-structure alignment, FSA
- Partial-structure alignment, PSA



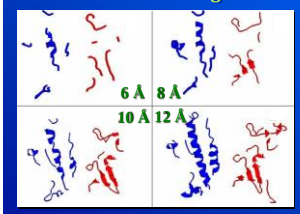
Sinha, Kundrotas & Vakser (2010) Proteins, 78:3235

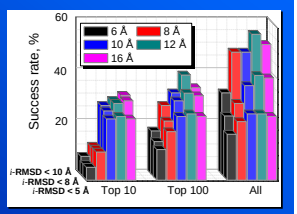
Docking by structure similarity

How much structure is needed?

11,932 fragments from PDB biological units

- 1BP3 interface fragment
- Success rates





12 Å cut-off distance is an optimum for interface definition in partial-structure alignment modeling

Sinha, Kundrotas & Vakser (2012) PLoS ONE, 7:e031349

Docking by structure similarity

Performance of PSA12 and FSA

372 two-chain complexes from the DOCKGROUND resource

- Best models (lowest *i*-RMSD)

Model *i*-RMSD	Number of targets modeled by		
	both PSA12* and FSA**	PSA12 only	FSA only
0 – 5 Å	130 (124)***	13	15
5 – 10 Å	38 (2)***	73	16

*PSA12 – partial-structure alignment with 12 Å library of interface fragments

**FSA – full-structure alignment

***Models build by both protocols using the same template

Sinha, Kundrotas & Vakser (2010) Proteins, 78:3235

Docking by structure similarity

Comparison of PSA12 and FSA

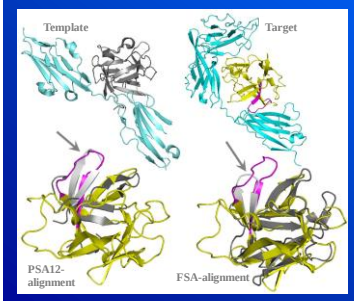
- Best (lowest *i*-RMSD) models are ranked Nr.1
  - 108 out of 143 PSA12 models
  - 116 out of 145 FSA models
  - 102 out of 130 common models
- PSA12 models are better compared to the FSA models
  - 17 PSA12 models have *i*-RMSD lower by 1 Å and more
  - 75 PSA12 models have *i*-RMSD lower by up to 1 Å
  - 19 PSA12 models have the same *i*-RMSD
  - 15 PSA12 models have *i*-RMSD higher by up to 1 Å
  - 4 PSA12 models have *i*-RMSD higher by 1 Å and more

Kundrotas & Vakser (2013) Proteins, 81:2137

Docking by structure similarity

Better performance of full-structure alignment

- Flexible interface loop



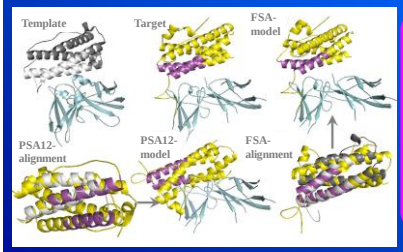
Target:	1ITB(A,B) Human interleukin-1beta
Template:	1CVS(A,C) Human fibroplast growth factor 2
Seq. identity:	16 % (14%)
i-RMSD (FSA)	4.8 Å
i-RMSD (PSA12)	7.3 Å

Kundrotas & Vakser (2013) Proteins, 81:2137

Docking by structure similarity

Better performance of full-structure alignment

- 4-helix bundle structure



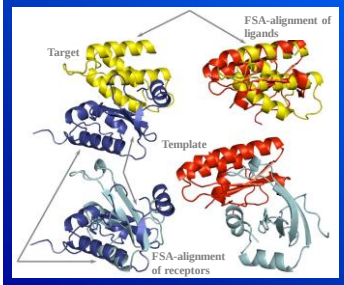
Target:	1F6F(A,B) *Ovis aries* placental lactogen
Template:	1PVH(A,B) Human leukemia inhibitor factor
Seq. identity:	14 % (23%)
i-RMSD (FSA)	4.5 Å
i-RMSD (PSA12)	15.0 Å

Kundrotas & Vakser (2013) Proteins, 81:2137

Docking by structure similarity

Better performance of full-structure alignment

- **Local structure similarity away from interface**



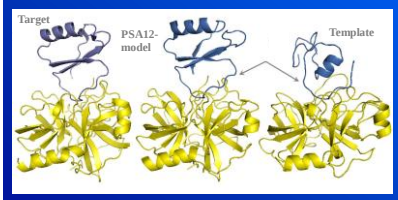
Target:	1V74(A,B) Colicin D with immunity protein
Template:	2FHZ(A,B) Colicin E5 with immunity protein
Seq. identity:	16 % (13%)
i-RMSD (FSA)	5.8 Å
i-RMSD (PSA12)	8.1 Å

Kundrotas & Vakser (2013) Proteins, 81:2137

Docking by structure similarity

Better performance of interface-structure alignment

- **Multiple binding**



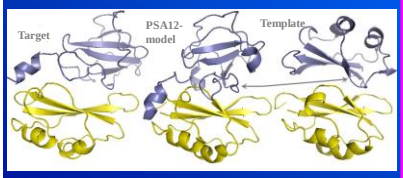
Target:	1ACB(E,I) Bovine chymo-trypsin with Eglin C
Template:	1LDT(T,L) Pig trypsin with its inhibitor
Seq. identity:	41 % (9%)
i-RMSD (FSA)	--
i-RMSD (PSA12)	1.3 Å

Kundrotas & Vakser (2013) Proteins, 81:2137

Docking by structure similarity

Better performance of interface-structure alignment

- **Biological vs. crystallographic interface**



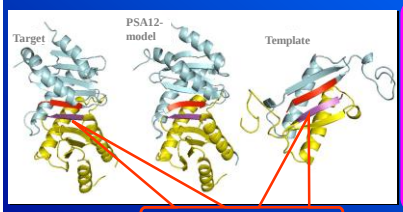
Target:	1E44(E,I) Colicin E3 with immunity protein
Template:	3EIP(A,B) Colicin E3 homodimer
Seq. identity:	99 % (17%)
i-RMSD (FSA)	--
i-RMSD (PSA12)	7.3 Å

Kundrotas & Vakser (2013) Proteins, 81:2137

Docking by structure similarity

Better performance of interface-structure alignment

- **Common structural motif at the interface**



Target: 1VET(A,B) Mice protein signaling complex

Template: 4OTC(B,C) RUVA protein from *E.coli*

Seq. identity: 8 % (11%)

i-RMSD (FSA) --

i-RMSD (PSA12) 6.0 Å

6 (out of 45/53) residues

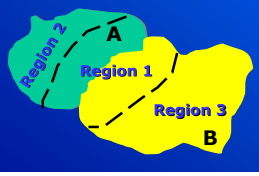
Kundrotas & Vakser (2013) Proteins, 81:2137

Are homology models accurate enough for docking?

Only interfacial part is needed

$$X(A:B) = X(AB) - (X(A) + X(B))$$

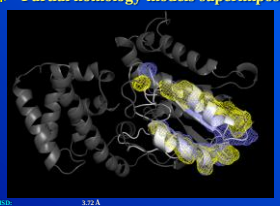
$$X_1(AB) + \cancel{X_2(AB)} + \cancel{X_3(AB)} \qquad X_1(A) + \cancel{X_2(A)} \qquad X_1(B) + \cancel{X_2(B)}$$

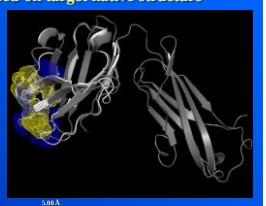


Are homology models accurate enough for docking?

"Bad" local Ψ-BLAST alignments

- **Partial homology models superimposed on target native structure**





Interface RMSD:	3.72 Å	5.00 Å
Target:	1bgg, chain A RGS DOMAIN OF RGS9 AND GTII CHIMERA ALPHA SUBUNIT (BOVINE + RAT)	1e1g, chain H COMPLEX BETWEEN PVS25 AND FAB FRAGMENT OF MALARIA TRANSMISSION BLOCKING ANTIBODY 2A8 (MOUSE)
Template:	2q77, chain B GYP1 TBC DOMAIN IN COMPLEX WITH RAB33 GTPASE (YEAST)	1fo6, chain A BM3.3 T CELL RECEPTOR ALPHA-CHAIN (MOUSE)
e-value:	0.859	0.37
Alignment identity: 17.95 %	28.87 %	
Interface identity	15.79 %	35.71 %
Target coverage:	34.46 %	38.89 %

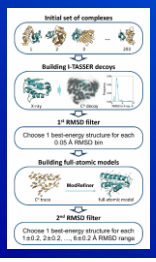
Kundrotas & Vakser (2010) PLoS Comp Biol, 6:e10000727

Docking by structure similarity

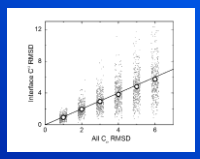
Benchmark on model-model complexes

- Set of protein models with pre-defined inaccuracy level
  - 165 complexes (from bound DOCKGROUND set)

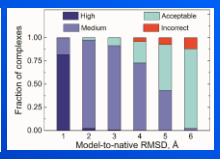
Algorithm



Local and global distortions



Quality of "ideal" complexes



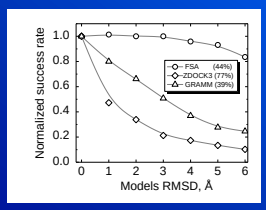
Anischenko *et.al.* (2015) Proteins, 83: 891

Docking by structure similarity

Benchmark on model-model complexes

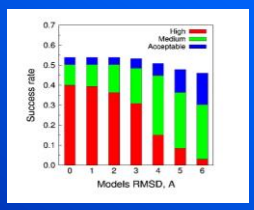
- Docking success

FSA and template-free docking, set 2



- Quality by CAPRI criteria

FSA models only, set 2

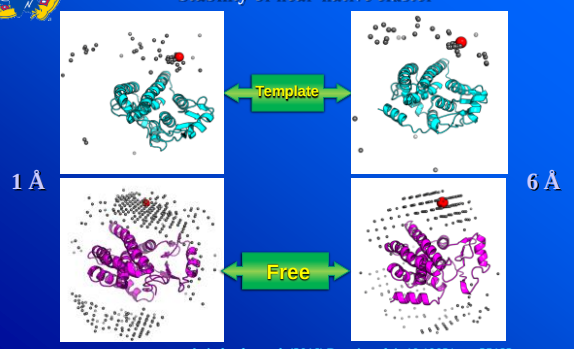


FSA docking is considerably less sensitive to inaccuracies of protein models compared to *ab initio* docking

Anischenko *et.al.* (2016) Proteins, doi: 10.1002/prot.25183

Docking by structure similarity

Stability of near-native cluster

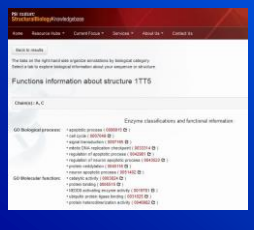


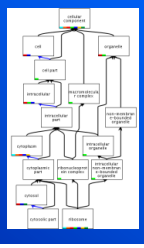
Anischenko *et.al.* (2016) Proteins, doi: 10.1002/prot.25183

Enhancements of Docking

GO-terms

- Ontology of pre-defined terms representing gene product properties
- Covers three domains
  - Cellular component, molecular functions and biological process
- Organized in directed acyclic graphs
- Accessible on Web





Enhancements of Docking

GO-terms

Technical details

- GO-score between terms t_1 and t_2

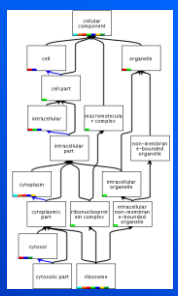
$$GO(t_1,t_2)=\frac{1}{1+IC(t_1)+IC(t_2)-2SH(t_1,t_2)}$$

- Information content of a term t in a database (UNIPROT)

$$IC(t)=-\log\frac{N(t)}{N}$$

- Shared information between t_1 and t_2

$$SH(t_1,t_2)=\max\{IC(t),t\in ComAnc(t_1,t_2)\}$$



F. Couto et al, Data & Knowledge Eng (2007), 61: pp.137 – 152

Enhancements of Docking

GO-terms

Technical details

- GO-score between proteins A and B

$$GO(A,B)=\frac{GO(A,T\{B\})+GO(B,T\{A\})}{2}$$

$$GO(A,T_B)=\frac{1}{N_{T\{A\}}}\sum_{i=1}^{N_{T\{A\}}}GO(t_i\{A\},T\{B\})$$

$$GO(t_i\{A\},T\{B\})=\max\left\{GO(t_i\{A\},t_1\{B\}),\cdots,GO\left(t_i\{A\},t_{N_{T(B)}}\{B\}\right)\right\}$$

F. Couto et al, Data & Knowledge Eng (2007), 61: pp.137 – 152

Enhancements of Docking

GO-terms

Technical details

- **GO-score between complexes AB and CD**

$$GO(AB,CD) = \max\left\{\frac{GO(A,C)+GO(B,D)}{2}, \frac{GO(A,D)+GO(B,C)}{2}\right\}$$

- **Cumulative GO-score**

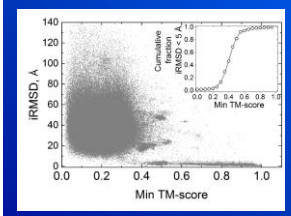
$$\text{GO}-\text{score} = 0.50 \times GO_{\text{mol.fun}} + 0.23 \times GO_{\text{bio.proc.}} + 0.27 \times GO_{\text{cel.comp.}}$$

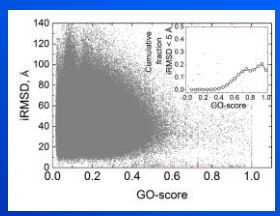
Enhancements of Docking

GO-terms

Interaction modes

- **TM-score**
- **GO-score**





GO-score can be used only as a compliment to TM-score

Kundrotas et al (2012) PNAS, 109, 9438

Hadarovic et al, in preparation

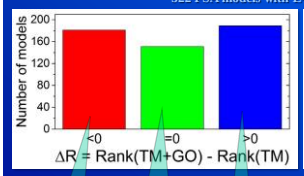
Enhancements of Docking

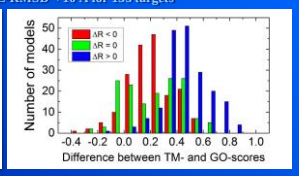
GO-terms

Change in model ranking

- **Generic view**
- **Detailed view**

522 FSA models with L-RMSD < 10 Å for 135 targets





181 models for 66 targets

150 models for 98 targets

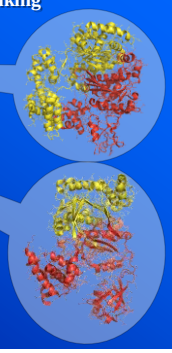
191 models for 66 targets

Enhancements of Docking

GO-terms

Change in model ranking

- **Example**



- **Target: 1tt5 AB**

  human amyloid protein-binding protein 1 + ubiquitin-activating enzyme E1C isoform 1

- **Template: 3kyd AB**

  human SUMO-activating enzyme subunits 1 and 2

- **Model parameters**

  TM-score1 – 0.438; TM-score2 – 0.604;
  GO-score – 0.604; L-RMSD – 3.05 Å
  Number of models: 523

- **Model ranking**

  By TM-score - 235

  By sum of TM- and GO-scores - 6

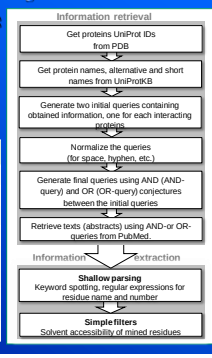
Enhancements of Docking

Text mining

Main idea and algorithm

- **Text mining was previously used for:**
  - Identifying interacting proteins
  - Small molecule binding sites
- **Information retrieval**
  - protein1 AND protein2 (AND-query)
  - protein1 OR protein2 (OR-query)
- **Information extraction**

Parameter	Value
Number	[1-9][0-9]*
Amino acid (AA)	{Ala, …, Val} OR {ala, …, val} OR {ALA, …, VAL}**
Three_letter_residue	AA(no space)Number OR AA(space)Number OR AA – Number | AA(Number)
Full_AA	{Alanine,…,Valine} OR {alanine,…,valine}***
Full_word_residue	Full_AA(no space)Number OR | Full_AA(space)Number OR Full_AA – Number OR Full_AA(Number)
Single_AA	{A,…,V}****
Single letter mutation	Single_AA(no space)Number(no space)Single_AA
Three letter mutation	AA(no space)Number(no space)AA OR AA-Number(no space)AA



Enhancements of Docking

Text mining

Benchmark results

- **Dataset**

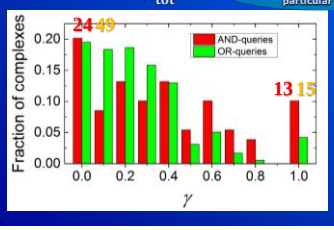
  587 binary protein complexes from DOCKGROUND
  Direct PDB citations excluded

- **Text mining performance**

$$\gamma = \frac{N_{\text{int}}}{N_{\text{tot}}}$$

Number of interface residues extracted from all abstracts for a particular complex

Total number of residues extracted from all abstracts for a particular complex



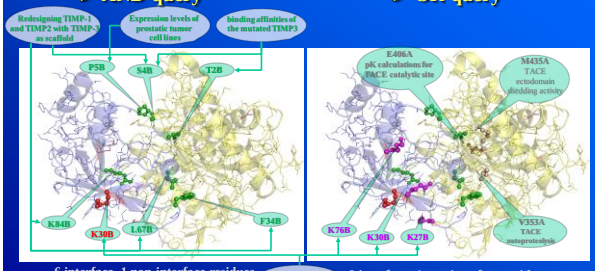
- **AND-queries**
  - 129 complexes (22%)
  - 1986 residues
  - 1859 abstracts
  - 725 interface residues (36%)
- **OR-queries**
  - 354 complexes (60%)
  - 13111 residues
  - 29483 abstracts
  - 3482 interface residues (26%)

Enhancements of Docking

Text mining

TACE-N-TIMP-3 complex (PDB 3cki)

- AND-query
- OR-query



6 interface, 1 non-interface residues
γ = 0.86

TIMP3-ECM binding site

9 interface, 4 non-interface residues
γ = 0.69

Some conclusions

Topics for discussion

- Docking of proteins by structure similarity outperforms docking by sequence similarity in terms of template availability and detection
- Docking of proteins by the interface structure similarity complements docking by the full-structure similarity
- Protein models can be docked by the structure similarity protocol with reasonable accuracy
- Docking protocols can be enhanced by the information-driven approaches

Protein docking by sequence and structure similarity

Petras Kundrotas

Center for Computational Biology/Bioinformatics
The University of Kansas
Lawrence, Kansas, USA

pkundro@ku.edu

Thank you for your attention!

Bioinformatics course, Stockholm University, February 2, 2017