

PDBjを利用した構造からの機能予測

木下賢吾

東大医科研

PDBj講習会2008

@大阪大学中之島センター

kinosita@hgc.jp

<http://www.hgc.jp/~kinosita>

Today's Topic++

- 背景
- 立体構造情報を得るまで
 - blastp, disorder, homology modeling
- 立体構造情報を得た後
 - 表面構造を利用する・基質結合部位の予測
 - 保存残基に着目する
 - DNA結合部位の予測
 - タンパク質間相互作用

増えるゲノム配列

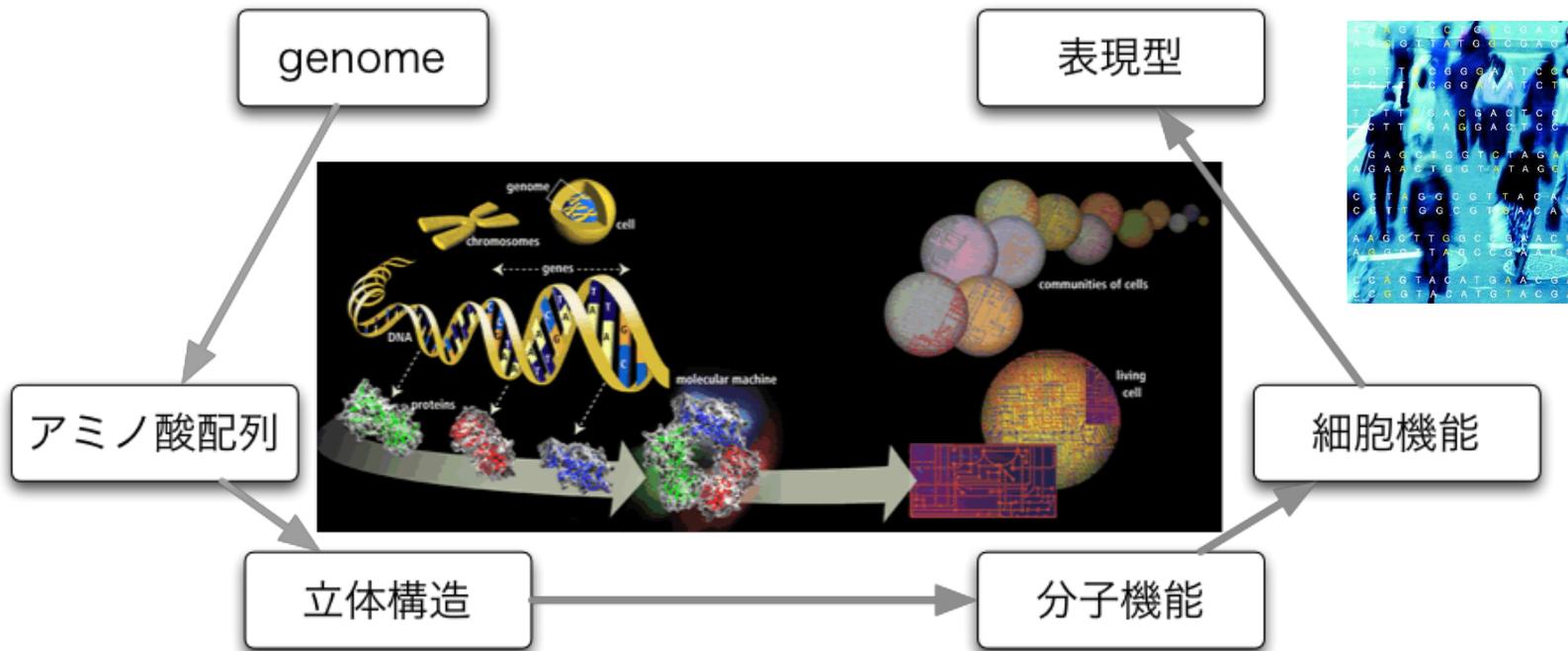
Genome sequencing projects statistics

Organism	Complete	Draft assembly	In progress	total
Prokaryotes	594	403	468	1465
Archaea	47	4	31	82
Bacteria	547	399	437	1383
Eukaryotes	23	129	186	338
Animals	4	53	90	147
Mammals	2	21	26	49
Birds		1	2	3
Fishes		3	6	9
Insects	1	19	20	40
Flatworms		1	3	4
Roundworms	1	3	13	17
Amphibians			2	2
Reptiles			2	2
Other animals		6	19	25
Plants	3	3	34	40
Land plants	2	2	27	31
Green Algae	1	1	7	9
Fungi	10	52	31	93
Ascomycetes	8	46	21	75
Basidiomycetes	1	4	6	11
Other fungi	1	2	4	7
Protists	6	19	27	52
Apicomplexans	1	10	6	17
Kinetoplasts	1	2	6	9
Other protists	4	7	14	25
total:	617	532	654	1803

Revised: Oct 22, 2007

約半分の遺伝子産物（蛋白質）の機能が分からない

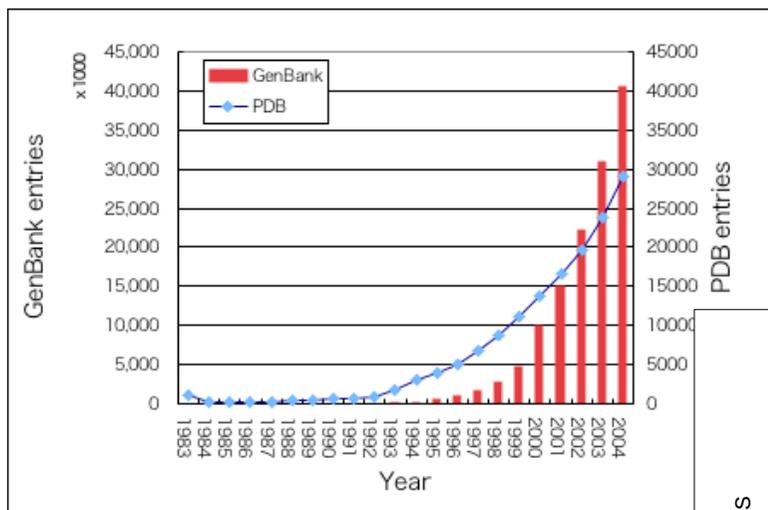
生物における情報の流れ



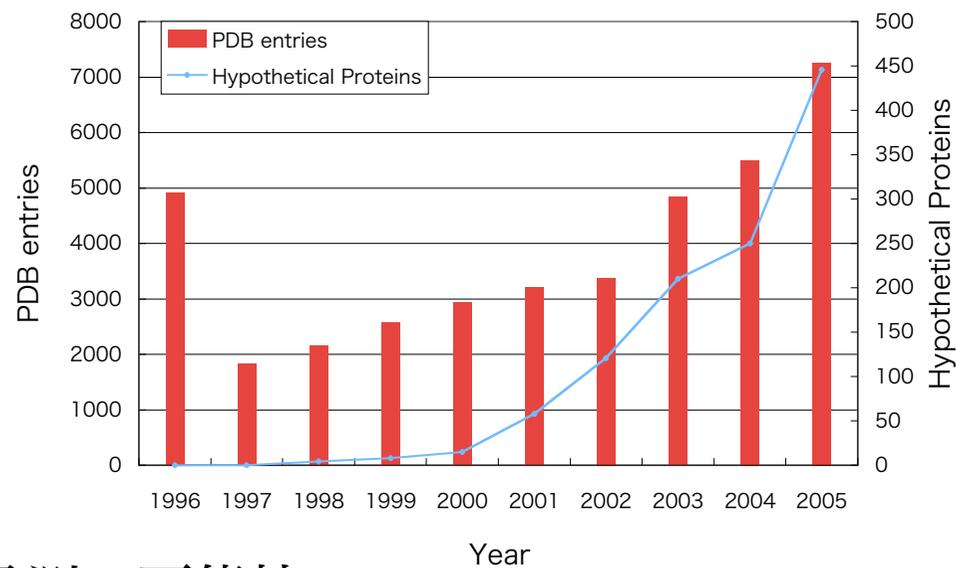
- ゲノムプロジェクト→ゲノムの配列を決める
- 構造ゲノムプロジェクト→タンパク質の立体構造を決める

立体構造情報の急激な増加

PDBエントリー数の急激な増加



機能未知たんぱく質の立体構造数の増加



➡ 立体構造からの機能予測の可能性

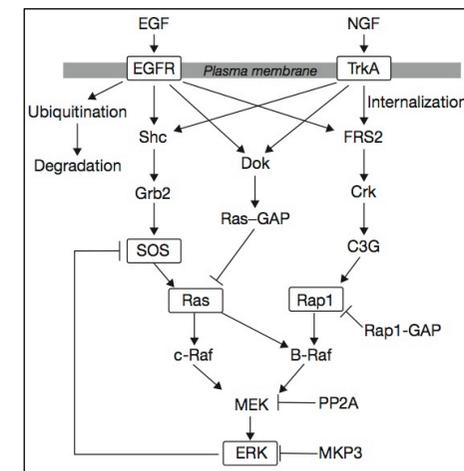
機能予測の対象とする「機能」

- 対象とする機能: 分子機能
 - タンパク質1つで決まる機能
 - 文脈によらない機能



構造を見ると分かるはず

- 分子機能
 - 基質認識
 - 低分子
 - 高分子→タンパク質間相互作用
 - 酵素反応



文脈による機能の例
シグナル伝達系でのクロストーク

立体構造情報を得るまで

BLASTPを利用して構造情報の有無を調べる

- 立体構造の有無を調べる時はDatabaseをPDBにする

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

MPVRRGHVAPONTFLDTHIRKFEQGSRKFIANARVENCAVIYCNDGFCELCG
YSRAEVMQRPTCDFLHGPRTQRRAAAQIAQALLGAERKVEIAFYRKDGSFC
LCLVDVVPKNEGDGAVIMFILNFEVMEKDMVCGSPAHDTNHRCPPTSWLAPGR
AKTFRLLKALLALTARESSVRSRGAGAGAPAVVDVLTTPAAPSSSLAL
DEVTAMDNHVAGLGAERRALVPGSPPRSAPGQLPSRAHSLNPDASGSSC

Query subrange

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database

Organism

Enter organism name or id—completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query

Enter an Entrez query to limit search

Program Selection

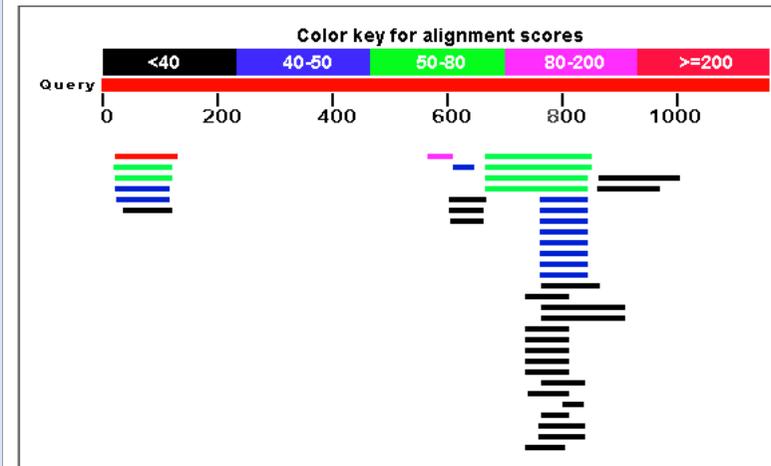
Algorithm blastp (protein-protein BLAST)
 PSI-BLAST (Position-Specific Iterated BLAST)
 PHI-BLAST (Pattern Hit Initiated BLAST)
Choose a BLAST algorithm

**実行開始
ボタン**

Search database **pdb** using **Blastp (protein-protein BLAST)**
 Show results in a new window

[Algorithm parameters](#) Note: Parameter values that differ from the default are highlighted in yellow

(例) hERG channelでの出力例



- 似た配列を持つ構造が有る部分を色つきのバーで表示
- 色の意味は上のカラーバー参照
- クリックすると詳しいアライメントが見れる

URLはNCBI BLASTPで検索

構造が解けていない部分

- ホモログの構造が
 - ある
 - ホモロジーモデリング
 - 構造情報を利用して機能情報を探る
 - ない
 - 基本的にあきらめる
 - 構造予測が出来ることは無いではないが信頼度は低い
 - Disorderを疑ってみる
 - より遠い類縁関係を探す

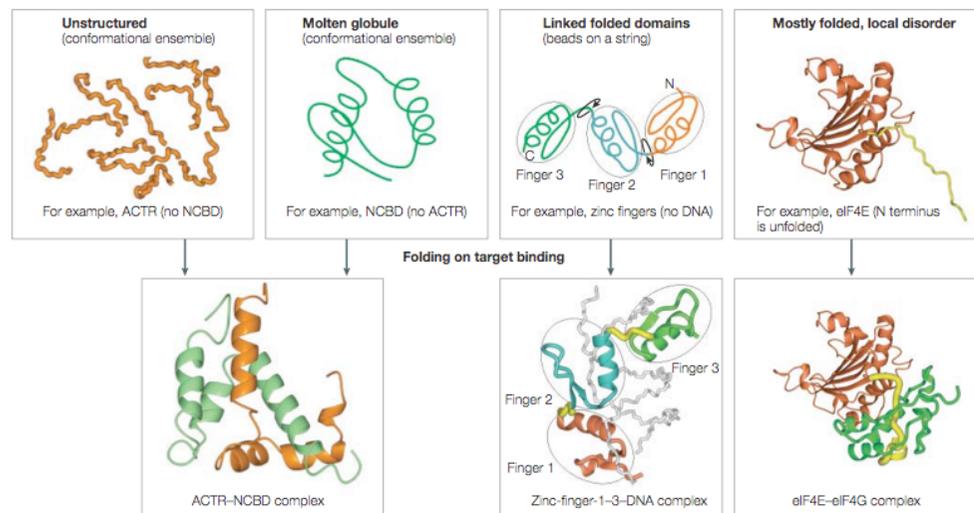
先に無い場合の話

Protein Disorder とは何か？

Intrinsically Disordered Protein

- 天然の状態では一定の構造をとらないタンパク質、及びその領域
 - Natively unstructured proteins とも呼ぶ
 - 定量的な定義は存在せず、いくつかの実験的手法により決定される
 - X線結晶構造解析、NMR、CD、プロテアーゼ消化、etc.
- パートナー分子と相互作用すると決まった構造を取る
- 様々なタイプが知られている

(Dyson, HJ and Wright, PE, *Nat Rev Mol Cell Biol*, 2005)



Statistics of disordered region

高等生物になるほどDisordered 領域が多い

Kingdom	# of proteins	Disorder freq. (% of aa)	Length > 30 (% of chains)	Length > 50 (% of chains)
Archaea	11,742	3.8	2.0	0.7
Bacteria	35,389	5.7	4.2	1.6
Eukaryota	88,531	18.9	33.0	19.6

Ward *et al*, JMB, **337**, 635-645, 2004

Estimation by DISOPRED2 (Jones *et al*)



タンパク質 disorder 領域とその機能

- タンパク質の機能に関わる disorder 領域が存在する

– Dunker et .al, *Biochemistry*, 2002

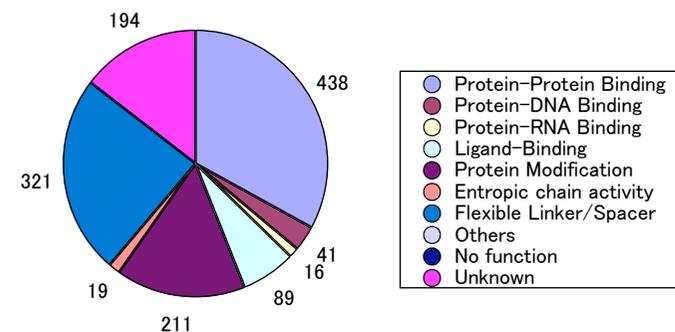
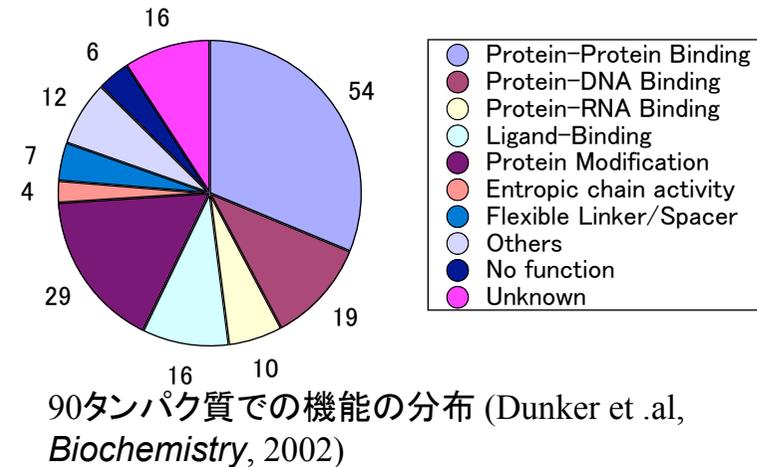
1. 分子認識(Molecular recognition)

- タンパク質-タンパク質結合
- タンパク質-DNA結合
- タンパク質-RNA結合 (t, r, m)
- リガンド結合

2. 分子集合体(Molecular assembly)

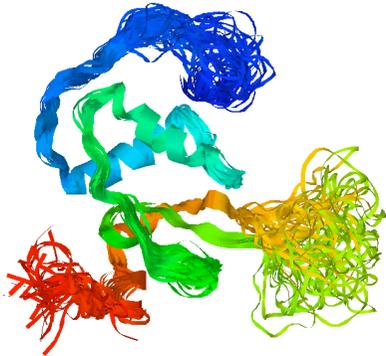
3. タンパク質修飾(Protein modification)

- リン酸化
- アセチル化
- グリコシル化

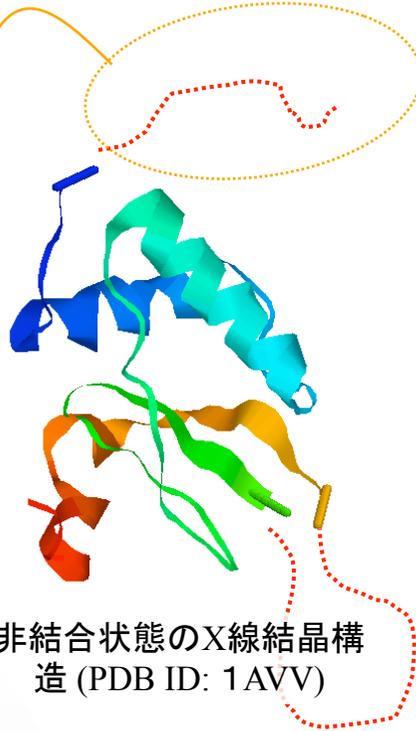


タンパク質-タンパク質結合での例

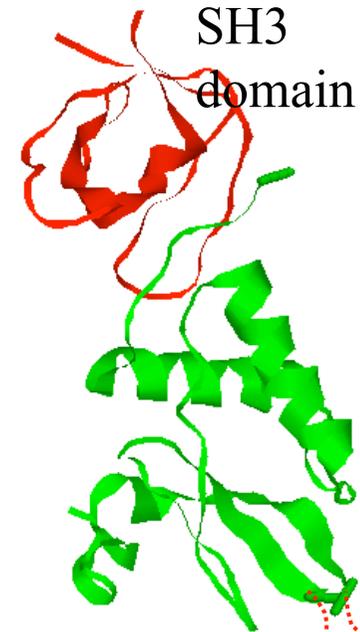
Negative Factor, HIV1



非結合状態のNMR構造 (PDB ID: 2NEF)



非結合状態のX線結晶構造 (PDB ID: 1AVV)



SH3 domain

結合状態のX線結晶構造 (PDB ID: 1AVZ)



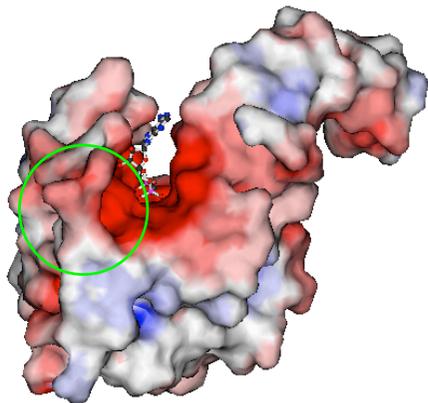
1qa4, HIV-1 NEF ANCHOR DOMAIN

非結合状態では1-73番目の残基が disorder の状態にあるが、70-75の残基がFYNのSH3ドメインと結合して安定化する

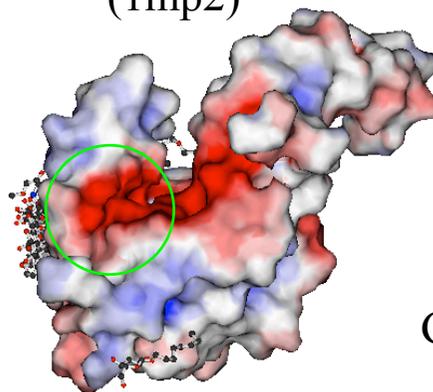
Kengo Kinoshita
IMS Tokyo University

タンパク質リガンドでの例

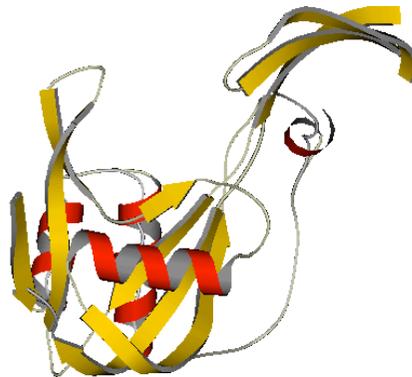
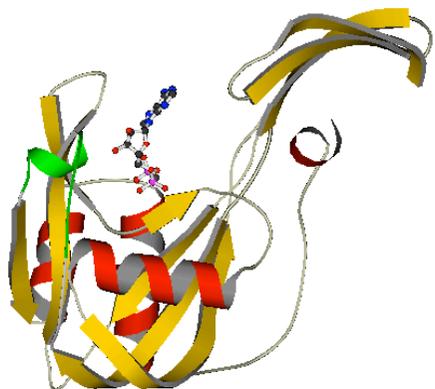
Complex form (1mqw)



Free form
(1mp2)



Green parts is flexible and invisible in the free form of the structure, which could prevent our method to predict the binding site.



IDPの予測: PrDOS

- アミノ酸配列で特徴がある
 - 荷電性残基 (Gluなど) が多い、配列が単調 ... etc
- 配列から高い精度で予測できる

<http://prdos.hgc.jp>

PrDOS Protein DisOrder prediction System

[Top](#) | [About](#) | [Usage](#) | [Contact](#)

PrDOS is a server to predict natively disordered regions of a protein chain from its amino acid sequence. PrDOS returns disorder probability of each residue as prediction results.

Submission

Input protein amino acid sequences in plain text or FASTA format. Multiple FASTA formatted inputs are acceptable. The server accepts [single-letter standard amino acid codes](#) and the code 'X' for non-standard amino acids. The codes for [ambiguous amino acids](#) and [particular non-standard amino acids](#) are automatically replaced by 'X'. If you want to obtain prediction results by e-mail, check "Receive prediction results by e-mail" and input your e-mail address.

Query amino acid sequence ([HELP](#))

Title (optional)

Prediction false positive rate: ([HELP](#))

Do not use template-based prediction ([HELP](#))

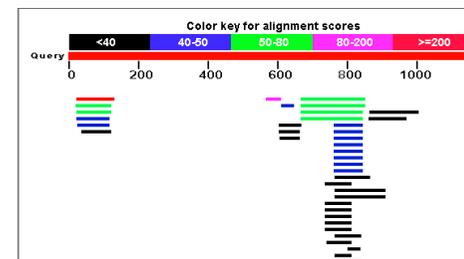
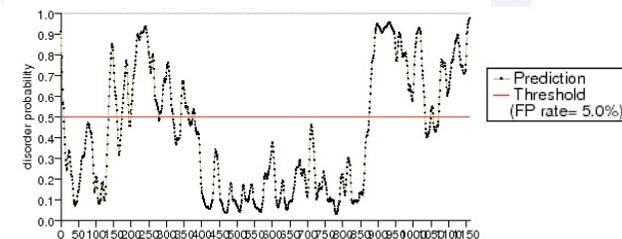
Receive prediction results by e-mail

E-mail address

配列を入力してPredictボタンを押すだけ
結果をメールで受け取りたいときのみメールアドレスを入れる

(例) hERG

1	MPVRRGHVAP	QNTFFLDIIR	KFEGQSRKFI	IANARVENCA	VIYVNDGFCE	50
51	LCGYSRAEVM	QRPCCTDFLH	GPRTQRRAAA	QIAQALLGAE	ERKVEIAFYR	100
101	KDGSFCFLCLV	DVVPVKMEDG	AVIMFILNFE	VVMEKDMVGS	FAHDTNHRGP	150
151	PTAALPAGRA	KTRFLKLPAL	LALTARESSV	RSGGAGGAGA	PGAVVVDVLD	200
201	TPAAPSESL	ALDEVMTAMN	HVAGLGAEE	RRALVPGSP	PRAPGQLFS	250
251	PRAHSLNPD	SGSSCSLART	RSRESCASVR	RASSADDIEA	MRAGVLPFP	300
301	RHASTGAMHP	LRSGLLNSTS	DSDLVRYRTI	SKIPQITLNF	VDLKGDPFLA	350
351	SPTSDREIA	PKIKERTHNV	TEKVTQVLSL	GADVLPEYKL	QAPRIHRWTI	400
401	LHYSFPAVM	DWLLLLIVY	TAVFTPYSA	FLKETEELP	PATECGYACQ	450
451	PLAVVDLIVD	IMFIVDILIN	FRTTYVWANE	EVVSHPGRIA	VHYFKGWFLI	500
501	DMVAIPFDL	LIFSGSSEEL	IQLLKTARLL	RLVRVARKLD	RYSEYGAAYL	550
551	FLLMCTFALI	AHWLACIWA	IGNMEQPHMD	SRIGWLHNLG	DQIGKPYNSS	600
601	GLGSPKDK	YVTALYFTFS	SLTSVQFQNV	SPNTNSEKIF	SICVMLIGSL	650
651	MYASIFGNVS	AIQRLVSGT	ARYHTQMLRV	REPFRHQIP	NPLRQRLEBY	700
701	FQHAWSYTNG	IDMNAVLGKF	PECLQADICL	HLNRSLLQHC	KPFRGATKGC	750
751	LRALAMKFKT	THAPPGDTLV	HAGDLLTALY	FISRGSIELL	RGDVVVALIG	800
801	KNDIFGEPLN	LYARPGKNSG	DVRALTYCDL	HKIHRDLDLE	VLDMPPEFSD	850
851	HFWSLEITF	NLRDTNMI	PGSPTELEGG	FSRQRKRKLS	FRRTDKDTE	900
901	QPGEVSLGPF	GRACAGPSSR	GAPOGFWGES	PSSGPFSSPES	SEDEGPGRSS	950
951	SPLRLVFPFS	PRPFGFPFG	EPLMEDCEKS	SDTCNPLSGA	FSGVSNIFSF	1000
1001	WDSRGRQYQ	ELPRCPATP	SLNIPLSSP	GRRPRGDVES	RLDALQRQLN	1050
1051	RLETRLSADM	ATVLQLLRQ	MLLVPPAYSA	VTPPGPGPTS	TSPLLPVSL	1100
1101	FTLLDLSLQ	VSQFMACEEL	PPGAPQLPQ	GPTRRLSLPG	QLGALTSQPL	1150
1151	HRHGSDFGS					1200



ita
arsity

より遠くの類縁関係を検出する

BLASTよりも高感度な配列検索

- Profile alignment: **PSI-BLAST**
- Profile-Profile: **FORTE**
- HMM-HMM: **HHPred**

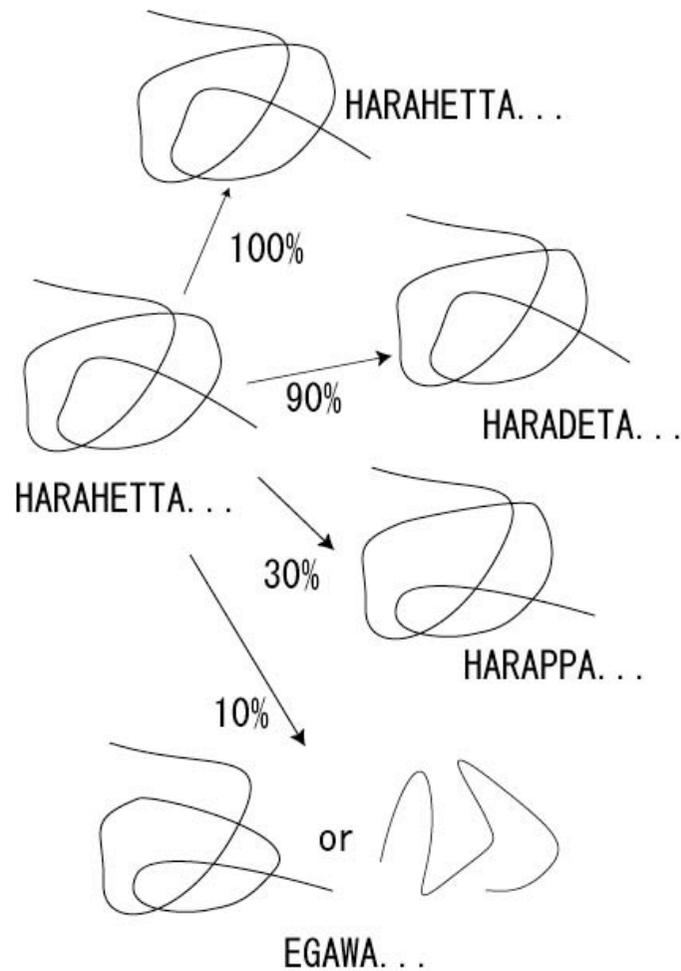
(下に行くほど高感度な事が多い＝計算時間も掛かる)

遠い類縁関係を使う場合の注意

- 全体構造は似ていても、違う部分も多くなる
- 機能が似ていないこともある
 - 機能のコアな部分は似ていても基質は違うことが多い
(ref: Todd et al, JMB, **307**, 1113-1143, 2001など)

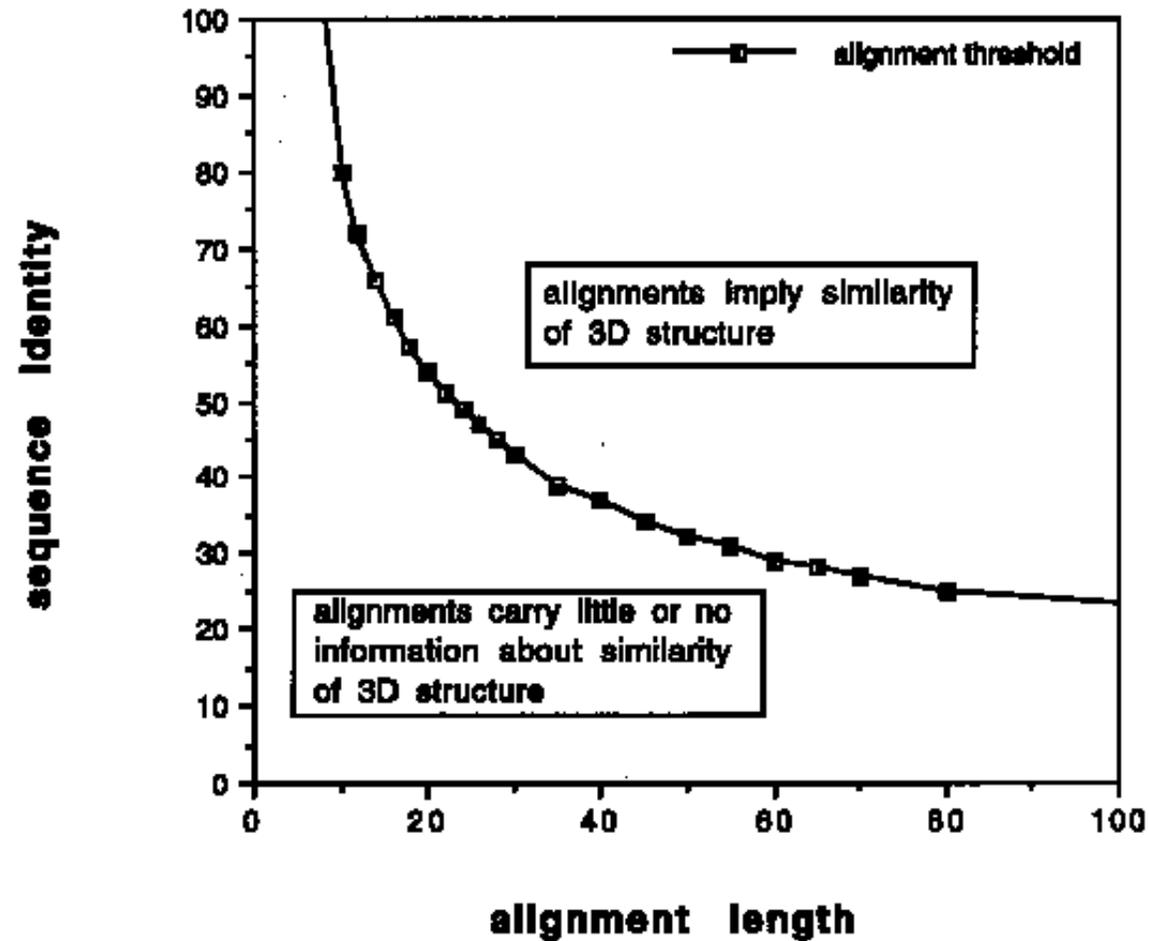


Introduction to Homology Modeling



- 足りない立体構造情報を補填する技術
- 原則
 - 配列が似ていれば構造は似ている
 - 構造未知でも、配列の似た蛋白質で構造既知のものがあれば、その構造を参考に構造を予想できる
- やること
 - 似た配列を探す
 - 配列のalignmentを作る
 - 配列の違う部分のつじつまを合わせる

Threshold of structural homology



Homology modelling の適用範囲

range of sequence
similarity in %
identical residues

key limiting factor
in model building
by homology

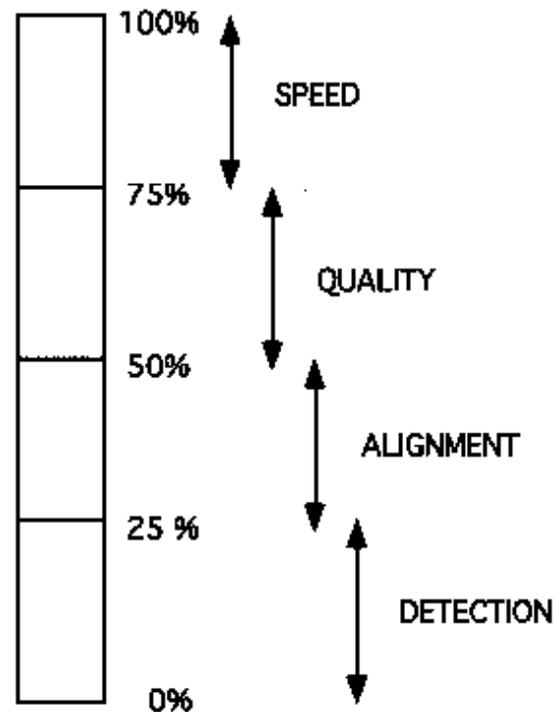
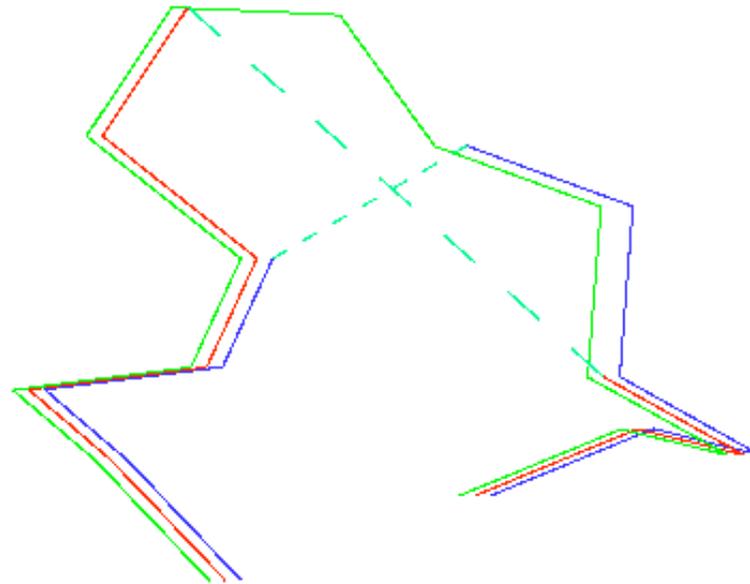


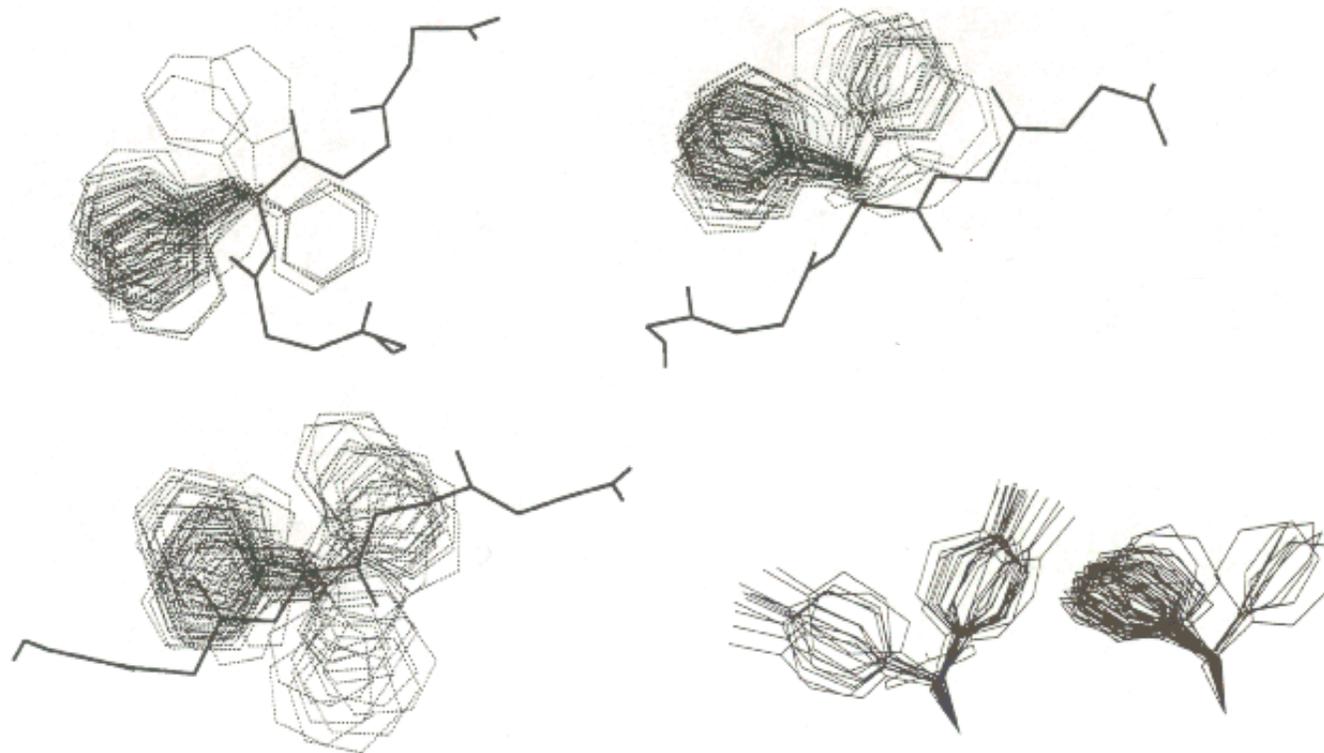
Figure 1. The main limiting steps for model building by homology as function of the percentage sequence identity between the structure and the model.

Alignmentの難しさ～Gapの入れ方の問題



1	2	3	4	5	6	7	8	9	10	11	12	13	14	
PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL	CYS	template
PHE	ASN	VAL	CYS	ARG	THR	PRO	---	---	---	GLU	ALA	ILE	CYS	alignment-1
PHE	ASN	VAL	CYS	ARG	---	---	---	THR	PRO	GLU	ALA	ILE	CYS	alignment-2

側鎖のモデリングの難しさ



同じ種類の残基でも
場所によって様々な
方向を持ちうる

隣接する残基の方向と同
時に決めないといけない

ホモロジーモデリングの実際

- web serverだとSWISS-MODELなどが利用可能
 - <http://swissmodel.expasy.org//SWISS-MODEL.html>
 - 簡便だが性能面で物足りない
- 余力があれば
 - MODELLERを入手して利用するのが良い
 - <http://www.salilab.org/modeller/>
- MODELLERをオンラインで使う手もある
(ライセンスコードは別途必要)
 - <https://genesilico.pl/toolkit/unimod?method=Modeller>

構造情報が手に入ったら？

- 機能部位の予測
 - p-cats
 - evolutionary trace analysis
- 分子表面の絵を眺める
 - eF-site/eF-surf
- リガンド結合部位の予測
 - eF-seek
- 蛋白質複合体の予測
 - docking
- DNAの結合部位の予測

配列解析ベースの方法

- P-cats
 - 保存している残基を、構造上の特徴を考慮して
 - 構造的保存残基
 - 機能性保存残基 (Active Site)
に区別する
 - 詳しくはJMB, 327, 1053-1064 (2003)
- Evolutionary Trace法
 - 詳しくは後述

P-cats server: <http://p-cats.hgc.jp/p-cats>



Ist Step
Input your query

ABOUT P-cats:

The catalytic or functionally important residues of a protein are known to exist in evolutionarily constrained regions. However, the patterns of residue conservation alone are sometimes not very informative, depending on the homologous sequences available for a given query protein. Here, we present an integrated method, named **P-cats**, to locate the catalytic residues in a protein from its sequence and structure. Mutations of functional residues usually decrease the activity, but concurrently often increase in stability. Also, catalytic residues tend to occupy partially buried sites in holes or clefts on the molecular surface. **P-cats** takes all of these analyses, i.e., the stability profile and surface geometry, into account as well as sequence conservation. A user, who would like to ask the location of catalytic, or functionally important residues to **P-cats**, should only prepare the coordinate of query structure in [PDB format](#). It will take some time to make an analysis so the result will be sent by E-mail.

REFERENCE: Ota M., Kinoshita K. and Nishikawa K. (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.* **327**, 1053-1064
[PMID: 12662930 \[PubMed - indexed for MEDLINE\]](#)

P-cats Job
submission starts
from here!

SUBMISSION STEP-1:

1. Upload your query structure in PDB format: 3chy.pdb
2. Enter your E-mail address:
3. Enter the e-value for the homologous sequence search by BLAST:
4. Select the sequence database:
5. Proceed to the next step or Reset

Result will be sent by
email

If you have any question or suggestion, please do not hesitate to contact us!
mail: p-cats@hgc.jp

M. Ota & K. Kinoshita, May 2003

Kengo Kinoshita
IMS Tokyo University

An example of return mail: 2nd Part

Prediction Results:

N	A	H	L	C1	C2	C3	S	R	DS	G	P	D
16	T	1	a	1.00	0.70	0.69	-0.25	6	0.55	1	0.09	
59	N	2	e	1.00	0.65	1.00	-0.21	11	0.50	1	0.68	*
87	T	5	b	1.00	0.93	0.91	-1.00	1	0.00	1	0.09	
106	Y	8	b	1.00	0.90	0.83	-0.09	10	1.36	1	0.82	*

Prediction results are shown in a table

In this example, 59th and 106th residues are predicted as "catalytic" residue

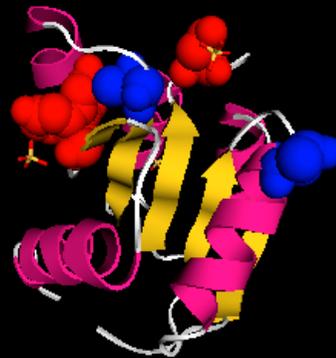
N : Residue number
 A : Amino acid
 H : Hydration class
 L : Local structure
 C1: Conservation number
 C2: Local conservation number
 C3: Spatial conservation number
 S : Score
 R : Rank position
 DS: Score difference
 G : Geometry of the site (1: surface, 2: cleft, 3: hole)
 P : Probability of the catalytic residues
 D : Final decision (* catalytic)

An explanation of the symbols in the result table.
 (see the original paper for detail:
 JMB, 327,1053-1064, 2003)

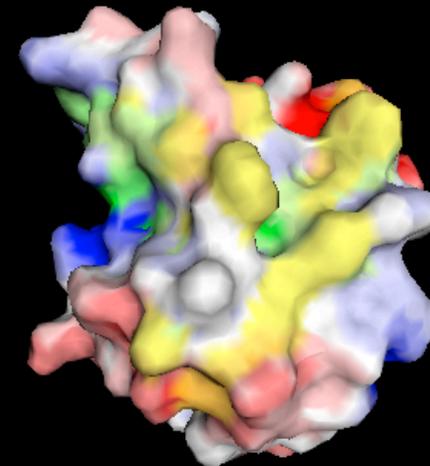
An example of interactive view of the prediction results.

P-cats results for the request-Thu Feb 10 14:02:36

In the left panel, ribbon model of your protein is shown with **RED** or **BLUE** CPKs for catalytic residues and conserved residues, respectively.



In the right panel, molecular surface with electrostatic potential is shown from the same direction.



```
107 atoms selected.  
jv>  
27 Atoms Selected.  
jv>  
14 Atoms Selected.  
jv>
```

Alignment

Active Sites

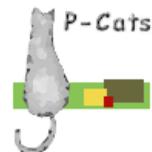
- 1) **59: Asn**
- 2) **106: Tyr**

Conserved Sites

- 1) **16: Thr**
- 2) **87: Thr**

electrostatic potential
-0.1V +0.1V
hydrophilic
hydrophobic

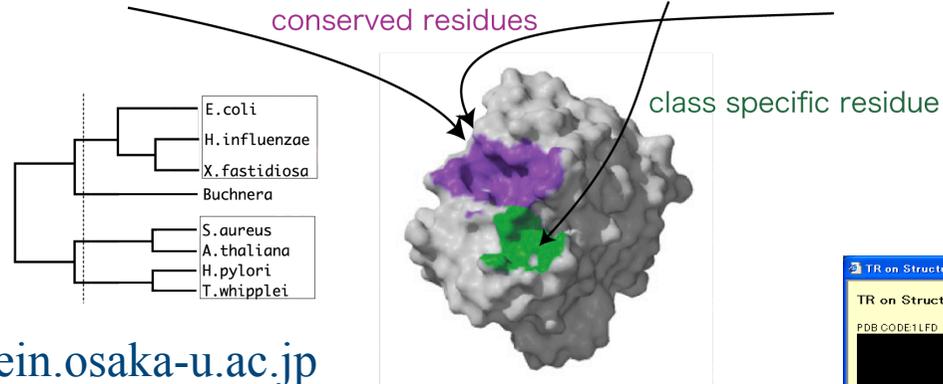
Color scheme of the molecular surface.



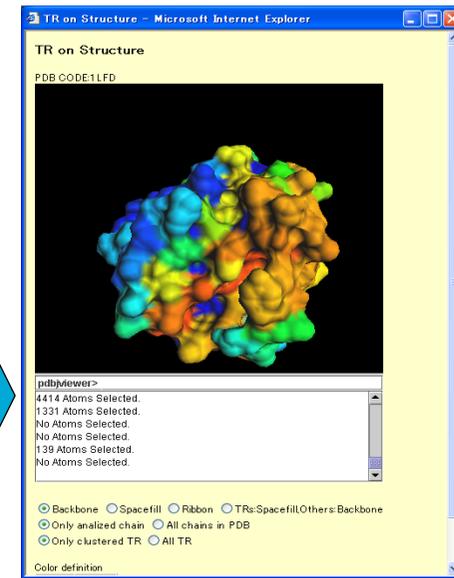
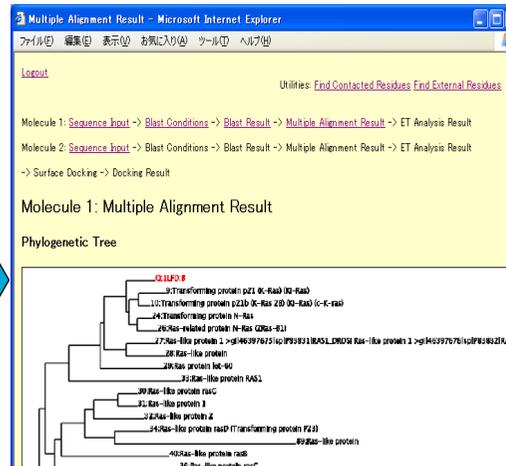
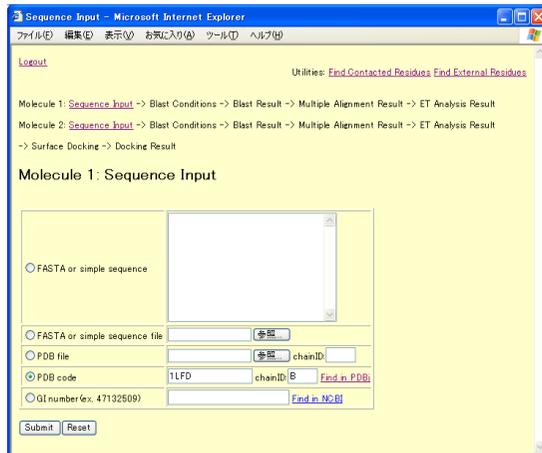
p-cats@hgc.jp

系統関係を考慮した保存度の解析 Evolutionary Trace

E. coli	QGA ¹ A ² F ³ E ⁴ G ⁵ V ⁶ I ⁷ A ⁸ Y ⁹ E ¹⁰ P ¹¹ V ¹² W ¹³ A ¹⁴ I ¹⁵ G ¹⁶ T ¹⁷ G ¹⁸ K ¹⁹ S ²⁰ A ²¹ T ²² P ²³ A ²⁴ Q ²⁵	Q ¹ A ² V ³ H ⁴ K ⁵ F ⁶ I ⁷ R ⁸ D ⁹ H ¹⁰ I ¹¹ A ¹² K ¹³ V ¹⁴ D ¹⁵ A ¹⁶ N ¹⁷ -I ¹⁸ A ¹⁹ E ²⁰ Q ²¹ V ²² I ²³ I ²⁴ Q ²⁵ Y ²⁶ G ²⁷ G ²⁸ S ²⁹ V ³⁰ N ³¹ A ³² S ³³ A ³⁴ E ³⁵ L ³⁶ F ³⁷ A ³⁸ Q ³⁹ P ⁴⁰ D ⁴¹ I ⁴² D ⁴³ G ⁴⁴ L ⁴⁵ V ⁴⁶ G ⁴⁷ G ⁴⁸ A ⁴⁹ S ⁵⁰ L ⁵¹ K ⁵² A ⁵³ D ⁵⁴ F
H. influenzae	L ¹ G ² V ³ E ⁴ A ⁵ F ⁶ N ⁷ G ⁸ V ⁹ I ¹⁰ A ¹¹ Y ¹² E ¹³ P ¹⁴ I ¹⁵ W ¹⁶ A ¹⁷ I ¹⁸ G ¹⁹ T ²⁰ G ²¹ K ²² S ²³ A ²⁴ T ²⁵ P ²⁶ A ²⁷ Q ²⁸	Q ¹ A ² V ³ H ⁴ A ⁵ F ⁶ I ⁷ R ⁸ G ⁹ H ¹⁰ I ¹¹ A ¹² A ¹³ K ¹⁴ S ¹⁵ Q ¹⁶ A ¹⁷ -V ¹⁸ A ¹⁹ E ²⁰ Q ²¹ V ²² I ²³ I ²⁴ Q ²⁵ Y ²⁶ G ²⁷ G ²⁸ S ²⁹ V ³⁰ N ³¹ D ³² A ³³ A ³⁴ E ³⁵ L ³⁶ F ³⁷ T ³⁸ Q ³⁹ P ⁴⁰ D ⁴¹ I ⁴² D ⁴³ G ⁴⁴ L ⁴⁵ V ⁴⁶ G ⁴⁷ G ⁴⁸ A ⁴⁹ S ⁵⁰ L ⁵¹ K ⁵² A ⁵³ P ⁵⁴ A ⁵⁵ F
X. fastidiosa	V ¹ G ² S ³ A ⁴ G ⁵ F ⁶ A ⁷ R ⁸ A ⁹ V ¹⁰ W ¹¹ A ¹² Y ¹³ E ¹⁴ P ¹⁵ I ¹⁶ W ¹⁷ A ¹⁸ I ¹⁹ G ²⁰ T ²¹ G ²² R ²³ T ²⁴ A ²⁵ T ²⁶ P ²⁷ D ²⁸ Q ²⁹	Q ¹ A ² V ³ H ⁴ A ⁵ F ⁶ I ⁷ R ⁸ G ⁹ E ¹⁰ V ¹¹ A ¹² K ¹³ A ¹⁴ D ¹⁵ A ¹⁶ R ¹⁷ -I ¹⁸ A ¹⁹ D ²⁰ S ²¹ L ²² P ²³ I ²⁴ L ²⁵ Y ²⁶ G ²⁷ G ²⁸ S ²⁹ V ³⁰ K ³¹ P ³² D ³³ N ³⁴ A ³⁵ E ³⁶ L ³⁷ F ³⁸ S ³⁹ Q ⁴⁰ P ⁴¹ D ⁴² V ⁴³ D ⁴⁴ G ⁴⁵ L ⁴⁶ V ⁴⁷ G ⁴⁸ G ⁴⁹ A ⁵⁰ S ⁵¹ L ⁵² V ⁵³ A ⁵⁴ E ⁵⁵ D ⁵⁶ F
Buchnera	L ¹ G ² T ³ S ⁴ A ⁵ F ⁶ K ⁷ N ⁸ I ⁹ I ¹⁰ I ¹¹ A ¹² Y ¹³ T ¹⁴ P ¹⁵ I ¹⁶ W ¹⁷ A ¹⁸ I ¹⁹ G ²⁰ T ²¹ G ²² V ²³ S ²⁴ A ²⁵ D ²⁶ P ²⁷ E ²⁸ H ²⁹ V	Q ¹ L ² I ³ H ⁴ V ⁵ F ⁶ I ⁷ K ⁸ N ⁹ Y ¹⁰ L ¹¹ K ¹² Y ¹³ S ¹⁴ S ¹⁵ I ¹⁶ -N ¹⁷ R ¹⁸ N ¹⁹ D ²⁰ I ²¹ I ²² Q ²³ Y ²⁴ G ²⁵ G ²⁶ I ²⁷ N ²⁸ H ²⁹ T ³⁰ N ³¹ V ³² K ³³ F ³⁴ I ³⁵ E ³⁶ Q ³⁷ P ³⁸ D ³⁹ I ⁴⁰ N ⁴¹ L ⁴² L ⁴³ I ⁴⁴ G ⁴⁵ N ⁴⁶ S ⁴⁷ L ⁴⁸ S ⁴⁹ A ⁵⁰ K ⁵¹ E ⁵² F
S. aureus	L ¹ S ² E ³ D ⁴ Q ⁵ L ⁶ K ⁷ S ⁸ V ⁹ V ¹⁰ I ¹¹ A ¹² Y ¹³ E ¹⁴ P ¹⁵ I ¹⁶ W ¹⁷ A ¹⁸ I ¹⁹ G ²⁰ T ²¹ G ²² K ²³ S ²⁴ T ²⁵ S ²⁶ E ²⁷ D ²⁸ A	N ¹ E ² M ³ C ⁴ A ⁵ F ⁶ V ⁷ R ⁸ Q ⁹ T ¹⁰ I ¹¹ A ¹² D ¹³ L ¹⁴ S ¹⁵ S ¹⁶ K ¹⁷ E ¹⁸ V ¹⁹ S ²⁰ E ²¹ A ²² T ²³ R ²⁴ I ²⁵ Y ²⁶ G ²⁷ G ²⁸ Y ²⁹ V ³⁰ K ³¹ P ³² N ³³ N ³⁴ I ³⁵ K ³⁶ E ³⁷ Y ³⁸ M ³⁹ A ⁴⁰ Q ⁴¹ T ⁴² I ⁴³ D ⁴⁴ G ⁴⁵ L ⁴⁶ V ⁴⁷ G ⁴⁸ G ⁴⁹ A ⁵⁰ S ⁵¹ L ⁵² K ⁵³ ---V
A. thaliana	V ¹ T ² N ³ --W ⁴ S ⁵ N ⁶ V ⁷ V ⁸ I ⁹ A ¹⁰ Y ¹¹ E ¹² P ¹³ W ¹⁴ A ¹⁵ I ¹⁶ G ¹⁷ T ¹⁸ G ¹⁹ K ²⁰ V ²¹ S ²² A ²³ P ²⁴ A ²⁵ Q ²⁶	Q ¹ Y ² H ³ V ⁴ H ⁵ E ⁶ L ⁷ R ⁸ K ⁹ W ¹⁰ L ¹¹ A ¹² K ¹³ N ¹⁴ S ¹⁵ A ¹⁶ D ¹⁷ V ¹⁸ A ¹⁹ T ²⁰ T ²¹ R ²² I ²³ Y ²⁴ G ²⁵ Y ²⁶ V ²⁷ G ²⁸ N ²⁹ C ³⁰ K ³¹ E ³² L ³³ G ³⁴ Q ³⁵ A ³⁶ D ³⁷ V ³⁸ D ³⁹ G ⁴⁰ F ⁴¹ L ⁴² V ⁴³ G ⁴⁴ G ⁴⁵ A ⁴⁶ S ⁴⁷ L ⁴⁸ K ⁴⁹ P ⁵⁰ -E ⁵¹ F
H. pylori	I ¹ D ² L ³ N ⁴ -Y ⁵ P ⁶ N ⁷ L ⁸ W ⁹ A ¹⁰ Y ¹¹ E ¹² P ¹³ I ¹⁴ W ¹⁵ A ¹⁶ I ¹⁷ G ¹⁸ T ¹⁹ G ²⁰ K ²¹ S ²² A ²³ S ²⁴ L ²⁵ E ²⁶ I ²⁷ D	Y ¹ L ² T ³ H ⁴ G ⁵ F ⁶ L ⁷ K ⁸ Q ⁹ I ¹⁰ L ¹¹ N ¹² -----Q ¹³ K ¹⁴ T ¹⁵ P ¹⁶ L ¹⁷ L ¹⁸ Y ¹⁹ G ²⁰ Y ²¹ V ²² N ²³ T ²⁴ Q ²⁵ N ²⁶ A ²⁷ K ²⁸ E ²⁹ L ³⁰ G ³¹ I ³² D ³³ S ³⁴ V ³⁵ D ³⁶ G ³⁷ L ³⁸ L ³⁹ I ⁴⁰ G ⁴¹ S ⁴² A ⁴³ S ⁴⁴ W ⁴⁵ E ⁴⁶ L ⁴⁷ N ⁴⁸ F
T. whipplei	F ¹ L ² D ³ S ⁴ Q ⁵ L ⁶ H ⁷ M ⁸ L ⁹ V ¹⁰ A ¹¹ Y ¹² E ¹³ P ¹⁴ S ¹⁵ S ¹⁶ A ¹⁷ T ¹⁸ N ¹⁹ S ²⁰ G ²¹ N ²² C ²³ A ²⁴ N ²⁵ S ²⁶ G ²⁷ D ²⁸ I	V ¹ R ² M ³ A ⁴ A ⁵ I ⁶ K ⁷ D ⁸ I ⁹ V ¹⁰ N ¹¹ -----V ¹² R ¹³ V ¹⁴ L ¹⁵ Y ¹⁶ G ¹⁷ G ¹⁸ Y ¹⁹ N ²⁰ L ²¹ F ²² N ²³ A ²⁴ S ²⁵ A ²⁶ V ²⁷ F ²⁸ N ²⁹ E ³⁰ D ³¹ L ³² L ³³ D ³⁴ G ³⁵ I ³⁶ L ³⁷ V ³⁸ G ³⁹ R ⁴⁰ A ⁴¹ S ⁴² N ⁴³ A ⁴⁴ S ⁴⁵ D ⁴⁶ F



<http://pdbjets.protein.osaka-u.ac.jp>



分子表面を眺める

eF-site database: <http://ef-site.hgc.jp>



[About eF-site](#) | [Tools](#) | [References](#) | [Links](#) | [Acknowledgements](#) | [Feedback](#)

190750 Entries, Last Update: 25-Dec-2005

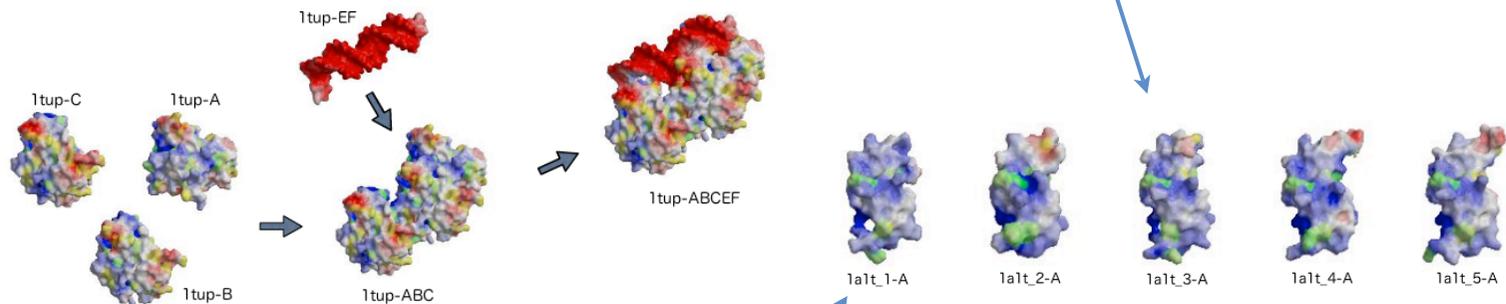
Keyword Search

 PDB code only and or

Category Search

- [Antibody](#)
- [Prosite](#)
- [Active Site](#)
- [Membrane](#)
- [Binding Site](#)

Examples of molecular surface



For all pdb entries, molecular surfaces are generated for individual subunits of proteins and the complex of proteins. In addition, when double strands DNA is included in the entry, a molecular surface for each dsDNA is also stored.

ほぼすべてのPDBエントリーを計算
サブユニット毎に計算
NMRはすべてのモデルを計算

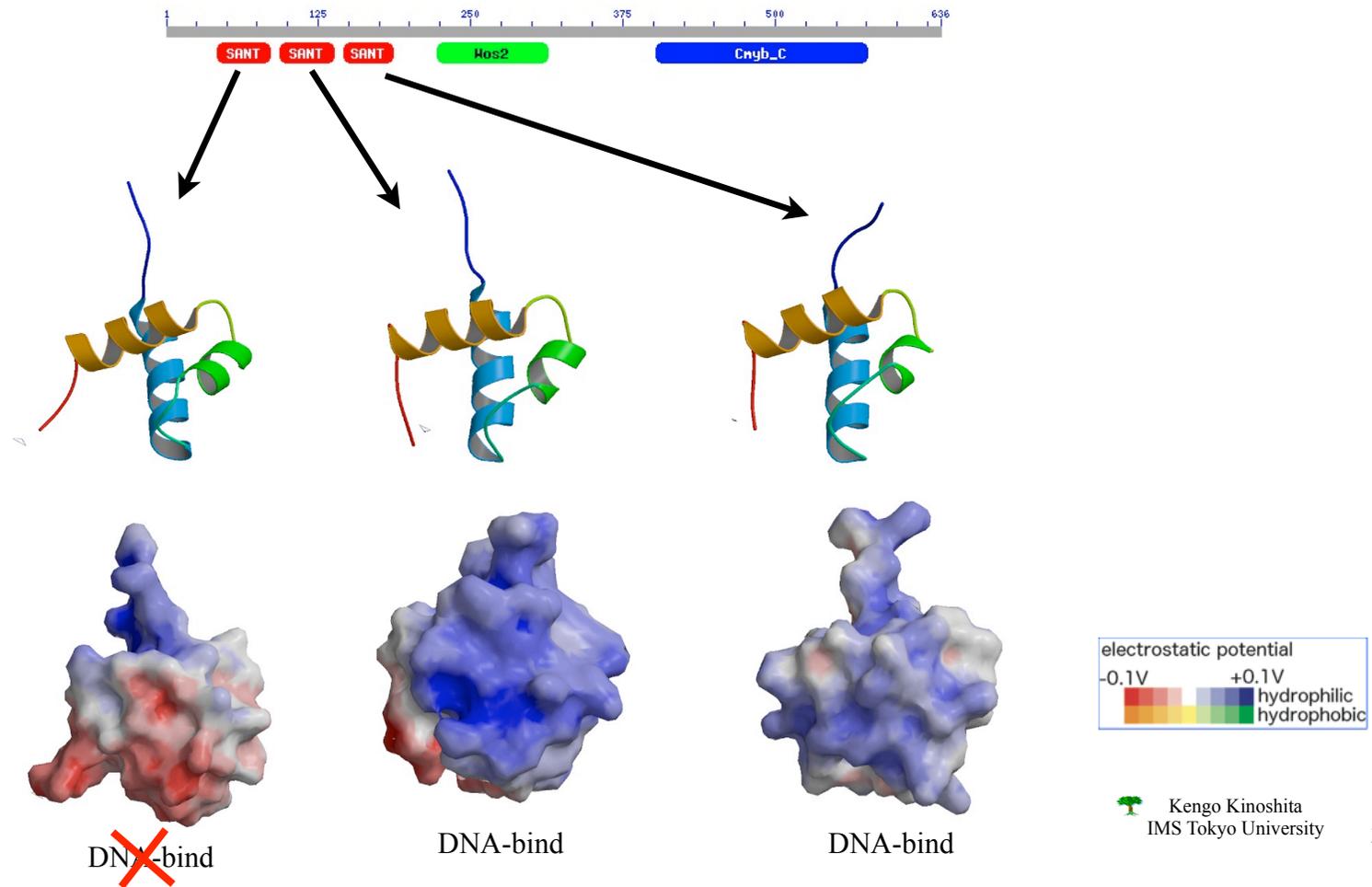
eF-site ID

When multiple models are stored in PDB, molecular surfaces for all the models are generated.

Constructed by Kinoshita, Kengo (The Institute of Medical Science, The University of Tokyo) and Nakamura, Haruki (Institute for Protein Research, Osaka University), collaborated with Information and Mathematical Science Laboratory, Inc. The development of this database is supported by BIRD-JST.
email: eF-site@protein.osaka-u.ac.jp

分子表面を眺めればそれで分かることもある

- (例) Myb proto-oncogene protein



低分子結合部位の予測

eF-seek @ <http://ef-site.hgc.jp/eF-seek>

- eF-siteに対する類似性検索による機能部位の予測
- 代表結合部位に対して検索
- 2006年12月から運用開始

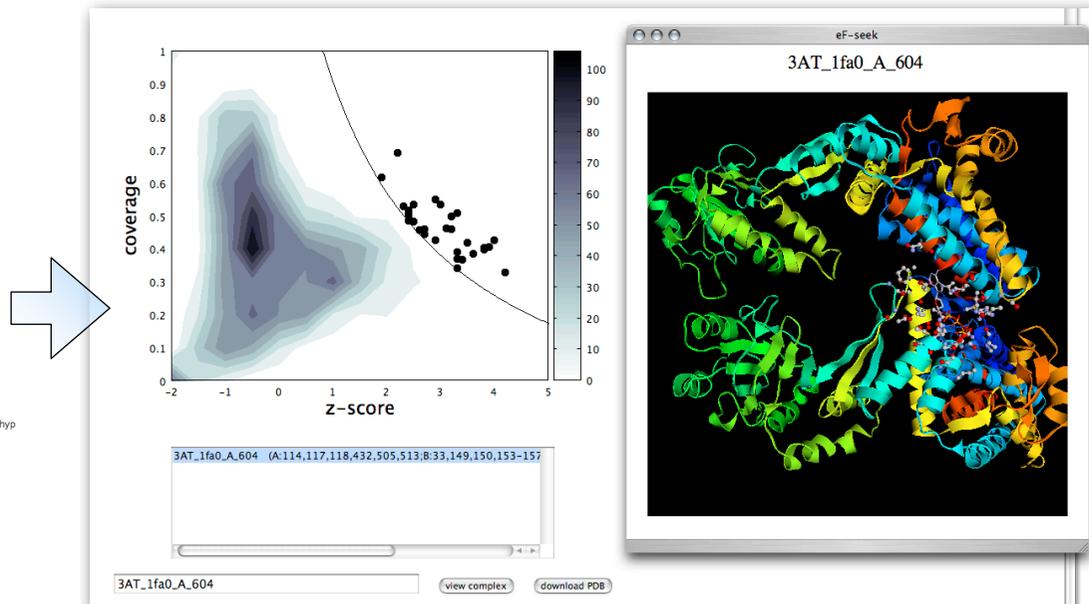
PDB eF-seek
<http://www.pdbj.org/index.html>

ABOUT eF-seek:
Molecular function of proteins are determined by their three dimensional structures, thus the similarity of protein structure can give some clues to infer their functions. In many cases, the molecular function are begun with the molecular interaction with small molecules (ligands). eF-seek is a web server to search for the similar ligand binding sites for the uploaded coordinate file with PDB format. The representative binding sites in eF-site database are search by our own algorithm based on the clique search algorithm.

Submission STEP-1:
Specify a PDB format file: ファイルが選択されていません
E-mail address:
Keyword: *1
Title: (optional)

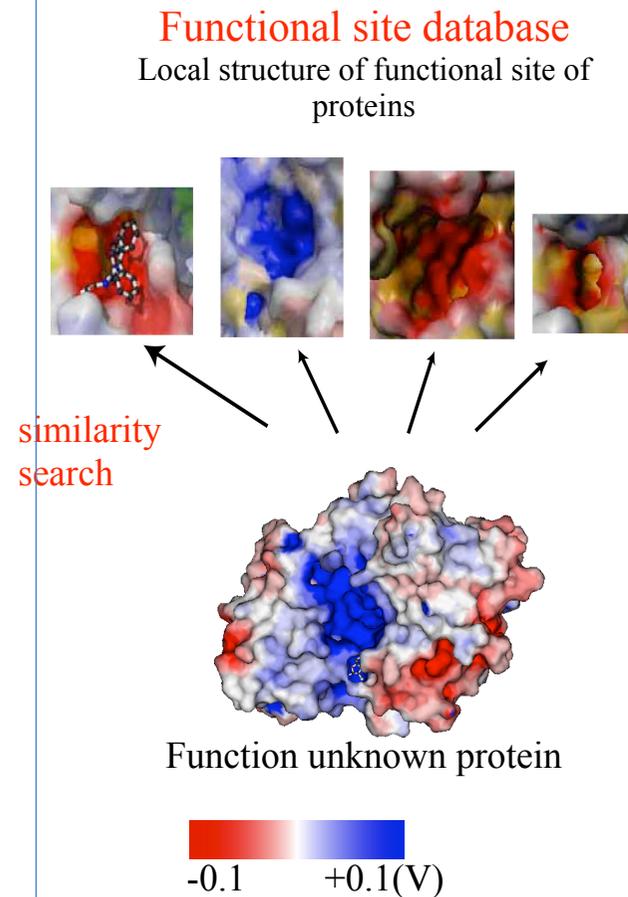
*1: Keyword will be used as the directory name of the output files, so only the alphabets, numbers, hyp and underscore are acceptable. And the length of the keyword should be from 4 to 16.

アップロードされたPDBファイルに対して機能部位を予測し、複合体の構造を返す



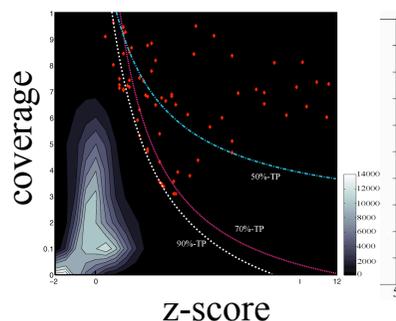
Structure based function prediction

- Goal
 - To predict a molecular function of proteins from their 3D structures
- Approach
 - To search for similar structures against the functional site database (local structure)
- **Structural information**
 - Molecular surface generated by Connolly's algorithm
 - Electrostatic potential obtained by solving Poisson-Boltzmann equations numerically

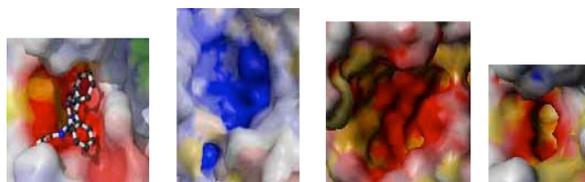


Normalization of similarity score

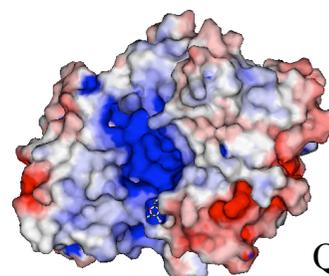
22,747 hetero compound binding sites appeared in PDB



Functional site patches



similarity search

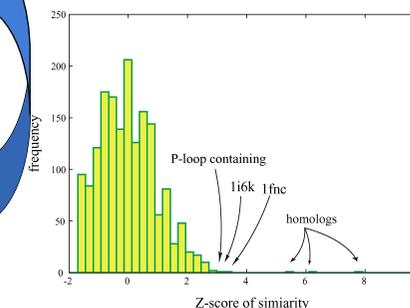


Query protein

Normalization from query protein's view.

$$Z\text{-score} = (\text{score} - \text{mean}) / \text{std}$$

Larger patches would get larger Z-score.



Normalization from functional sites' view

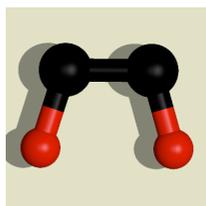
$$\text{coverage} = \frac{\# \text{ of corresponding vertices}}{\# \text{ of vertices in each patch}}$$

Results will be shown in **coverage vs. Z-score plot**.

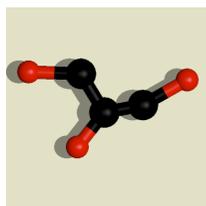
The number of corresponding vertexes is used as similarity score.

Threshold line determination

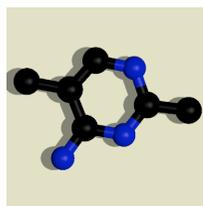
- 10 randomly selected representative with free and complex structures.
 - Homologous proteins with similar ligands are considered to be “correct”.



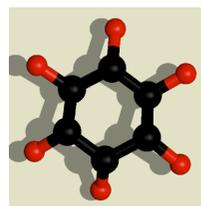
Ethylene glycol



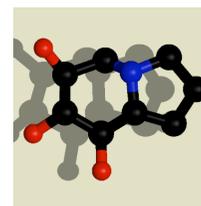
Glycerol



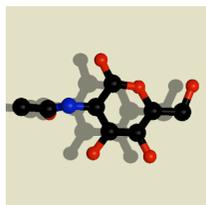
2,5-dimethyl-
pyrimidin-4-
ylamine



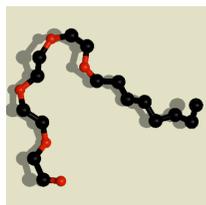
Myo-inositol



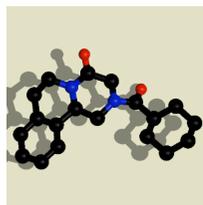
Castanospermine



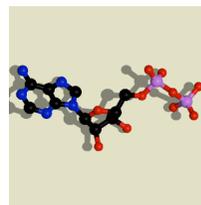
N-acetyl-d-
galactosamine



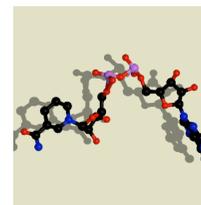
(Hydroxy
ethyloxy)tri
(ethyloxy)octane



Praziquantel

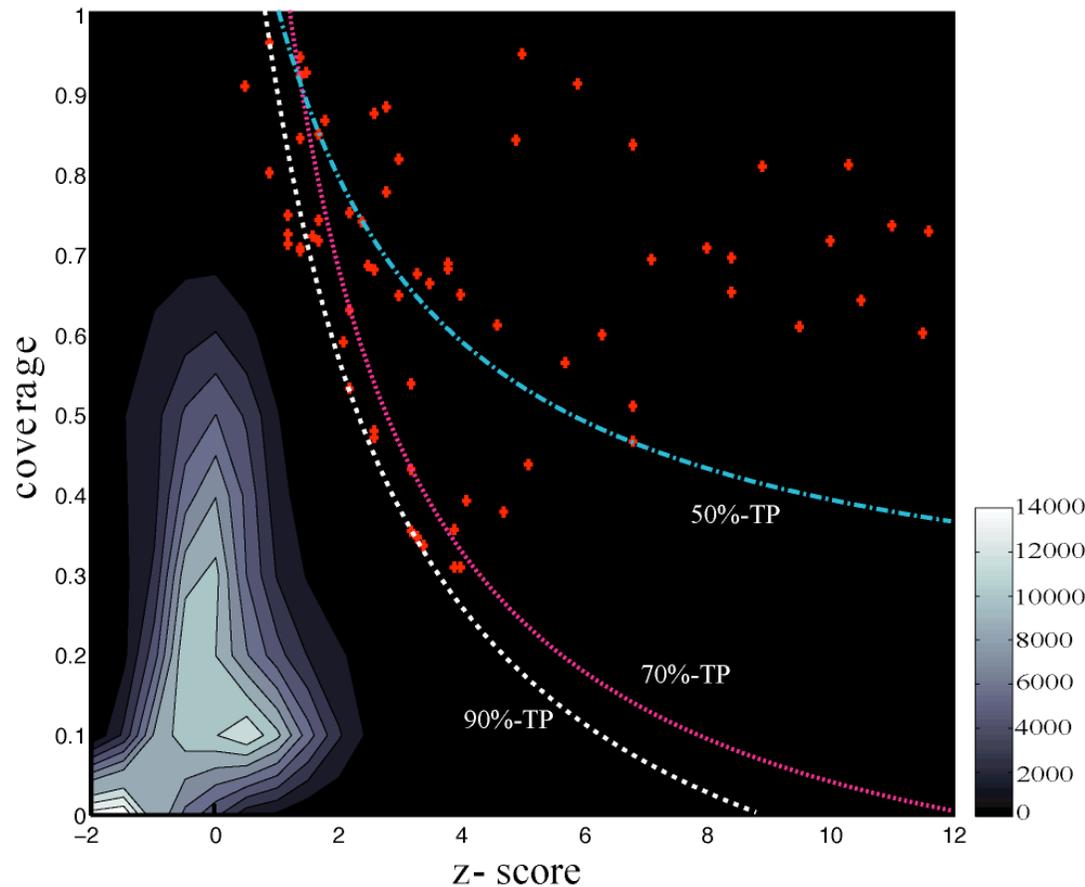


ADP



NAD

Threshold line determination



Maximize CC
 with a threshold line

$$\frac{a}{z+b} + c$$

under a constraint that
 fraction of TP exceed 90%,
 70% or 50%

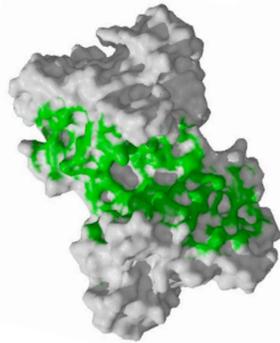
90%-TP line will be used
 hereafter.

$$CC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

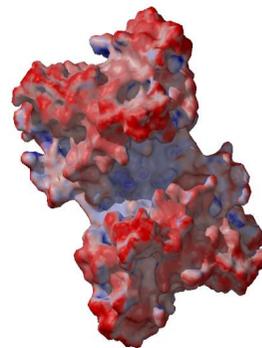
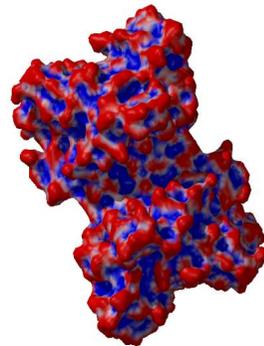
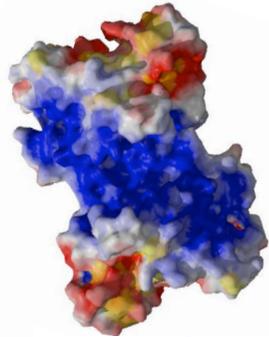
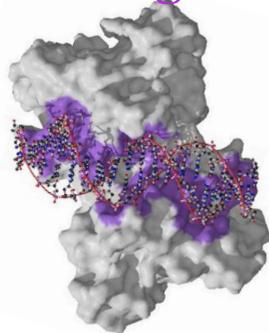
DNA結合部位の予測

Methionine Repressor
(1MJQ)

Predicted site



Binding site



Electrostatic potential

Red: negative, blue: positive

Yellow: hydrophobic surface

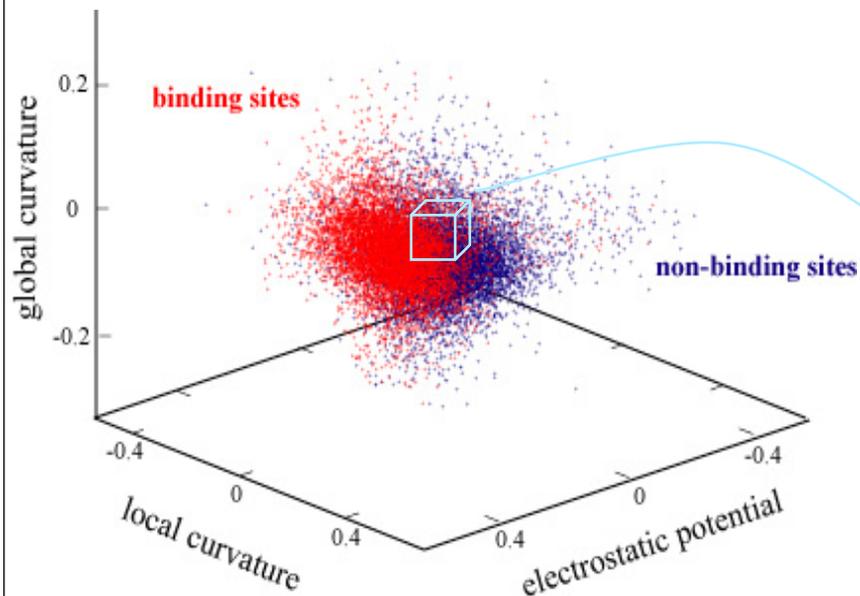
Local curvature relating to DNA-binding
directly: slightly protruded

Red: protrusion, Blue: concave

Global curvature relating to entire surface
geometry: largely concave

3つの特徴からの予測法

Prediction Scheme-1: Statistical Preference Measure



Distribution of electrostatic potential, local curvature and global curvature for all proteins in dataset-1

Relative Frequency

$$F_{bind}(\varphi_e, K_{local}, K_{global})$$

$$= N_{bind}(\varphi_e, K_{local}, K_{global}) / N_{bindtotal}$$

$$F_{non-bind}(\varphi_e, K_{local}, K_{global})$$

$$= N_{non-bind}(\varphi_e, K_{local}, K_{global}) / N_{non-bindtotal}$$

Statistical Preference Measure

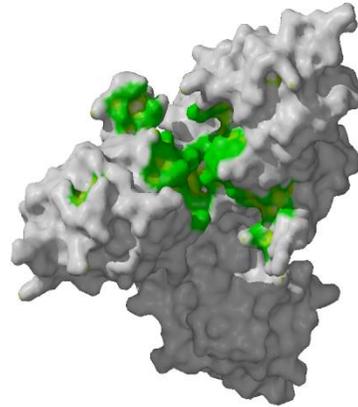
$$P_{bind} / P_{non-bind}$$

Prediction Scheme-2: Prediction Score (Pscore)

Statistical Preference Measure

$$P_{bind} / P_{non-bind} > 4.0$$

For each vertex, calculate the measure and colour it when the value exceed 4.0.



$$P_{bind} / P_{non-bind}$$



$$\text{Pscore} = \max (\text{Parea} / \text{Whole area})$$

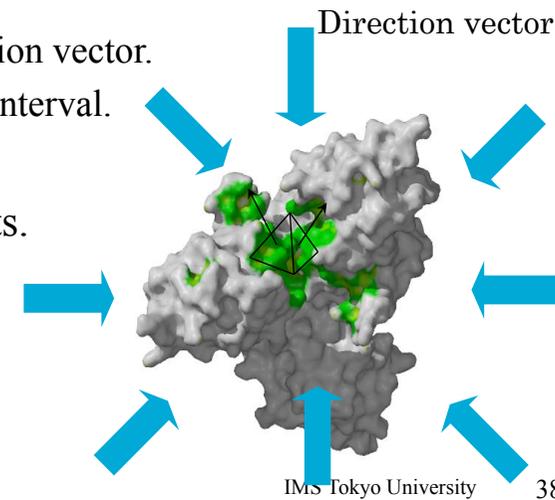
Parea : predicted DNA-binding *weighted area for a given direction*

Whole area: whole *weighted area for a given direction*.

Weights are calculated as inner product of normal vector and direction vector.

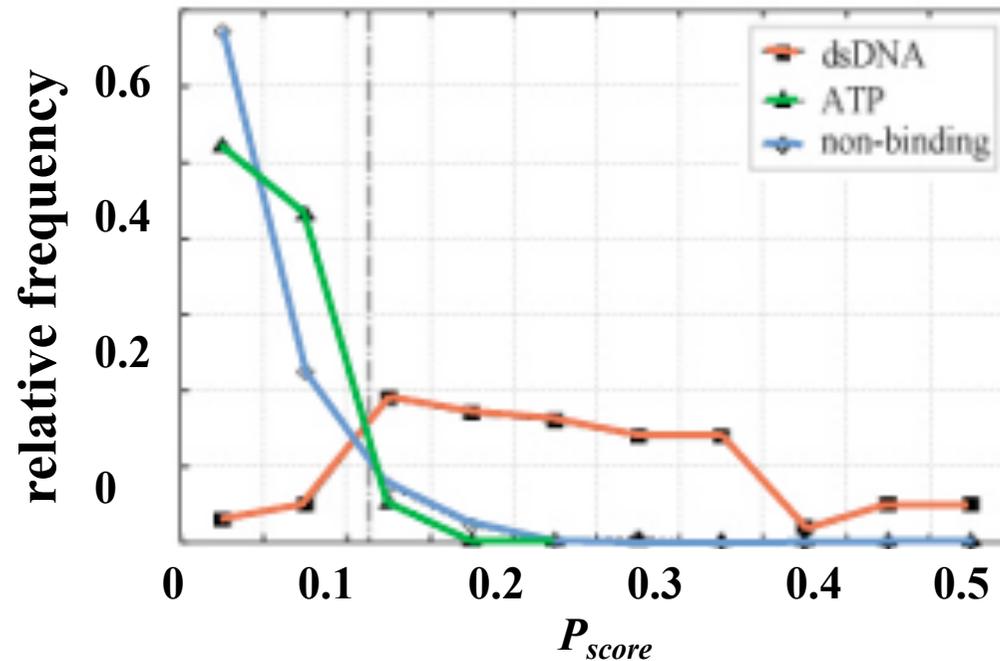
Maximization was done by searching all possible direction by 10° interval.

Pscore will be used as an indicator of the prediction results.



Prediction Results

Tsuchiya et al. (2004) *PROTEINS*, 55, 885-894.



Histogram of Prediction score for dsDNA-binding proteins (63), ATP-binding proteins (21), and non-dsDNA-binding proteins (406)

86% accuracy for predicting dsDNA-binding proteins, and

96% accuracy for predicting non-DNA-binding proteins including ATP-binding proteins.

PreDs: Prediction of DNA-binding site

PreDs

Prediction of DNA-binding site

PreDs is a server making the prediction of dsDNA-binding site on protein surfaces, according to our prediction method developed by focusing our attention on the shape of the molecular surface and the electrostatic potential on the surface.

Please input your E-mail address, your name and job title, and upload a **protein coordinate file with PDB format** of the query structure or **PDB ID**. E-mail address is essential since we will return a mail with a URL at which you will find the results.

E-mail Address	<input type="text"/>
Your Name	<input type="text"/>
Job Title	<input type="text"/>

Choose uploading a coordinate file or inputting PDB ID.
If both boxes are filled, the server chooses using the uploaded coordinates.

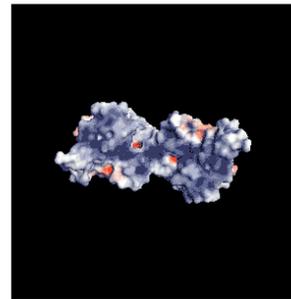
<input type="radio"/> Upload Coordinate File	<input type="text"/>	<input type="button" value="Browse..."/>
<input type="radio"/> Input PDB ID	<input type="text"/>	

<input type="button" value="SUBMIT"/>
<input type="button" value="RESET"/>

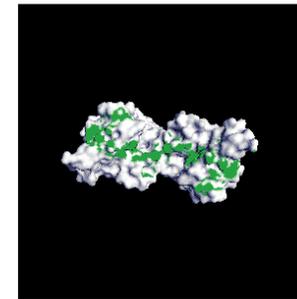
[HELP](#)

[Introduction](#)
[Usage](#)
[Prediction Method](#)
[Reference](#)

Electrostatic Potential

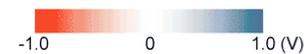


Predicted DNA-Binding Surface



Predicted DNA-Binding Region
(Residue Numbers (Chain ID))

A2(B) A47(A) A48(A) A48(B) A55(A) A55(B)
A96(A) A96(B) R61(A) R61(B) R66(A)
R66(B) R74(A) R74(B) R104(B) R140(A)
R140(B) N57(A) N57(B) N119(A) N119(B)
N123(A) N123(B) D117(A) D117(B) Q63(A)
Q63(B) Q70(B) Q141(A) G33(A) G51(A)
G51(B) G53(A) G53(B) G76(A) G76(B)
H27(B) H73(B) H78(A) H78(B) H98(A) I68(B)
L46(A) L46(B) L116(A) L116(B) K56(A)
K56(B) K65(A) K65(B) K92(A) K92(B)
K120(A) K120(B) F54(A) F54(B) P49(A)
P49(B) P59(A) P59(B) P126(A) P126(B)



<http://pre-s.protein.osaka-u.ac.jp/~preds/>

タンパク質間相互作用

- 何が
 - 最近の相互作用DBの整備
 - DNAマイクロアレーを使った相互作用DBの構築
- どのように
 - 表面の形の相補性&保存度に基づいた複合体予測法を開発
 - CAPRIに参加してみた
 - そこで見えてきた問題点&対応策

何が相互作用するのか？

タンパク質間相互作用情報の蓄積

- MINT, MIPS, DIP : 草分け的存在
- IntAct@EBI (最近はDNAやRNA,低分子も扱う)
- BioGrid : 統合DB。genetic interactionも扱う
- HPRD:ヒトに特化したDB。病気関連情報が豊富
refseq-IDがついていて便利

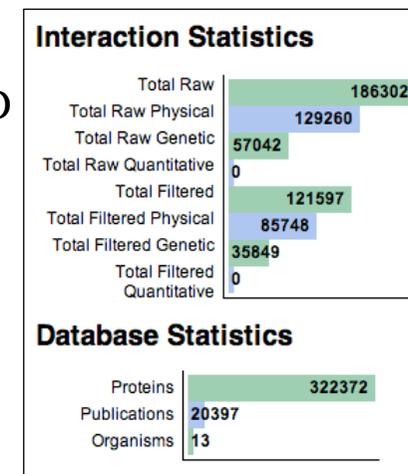


BioGRID

何が相互作用するのかという情報の蓄積



どのように相互作用するのか



何が相互作用するのか？ <http://COXPRESdb.hgc.jp>

(A) Top page

search
actin related protein 3
target: human mouse
 keyword (e.g. chloride)
 gene symbol (e.g. Grasp2)
 Entrez Gene ID (e.g. 66102)
 case sensitive
BLAST and other searches available. [more](#)

(B) Search result

Search result by keyword
Your request is **actin related protein 3** as keyword for **human**. (11 hits.)
No functional category is found.
11 loci are found.
Shaded loci do not have affy data.

Entrez Gene ID*	function ((symbol))*
10096	ARP3 actin-related protein 3 homolog (yeast) ((ACTR3))
57180	ARP3 actin-related protein 3 homolog B (yeast) ((ACTR3B))
399746	ARP3 actin-related protein 3 homolog B pseudogene ((FKSG74))
440888	ARP3 actin-related protein 3 homolog B pseudogene ((LOC440888))

(C) Locus page

COXPRESdb: locus page for ACTR3 (human)
Gene Symbol: **ACTR3 (human)**

functional annotation

function*	GO BP*	GO CC*	GO MF*
ARP3 actin-related protein 3 homolog (yeast)	GO:0006928 [list] [network] cell motility (347 loci) TAS	GO:0005885 [list] [network] Arp2/3 protein complex (14 loci) TAS	GO:0030027 [list] [network] lamellipodium (82 loci) IEA
	GO:0005198 [list] [network] structural molecule activity (1341 loci) IEA	GO:0005524 [list] [network] ATP binding (2594 loci) IEA	GO:0000166 [list] [network] nucleotide binding (3871 loci) IEA
		GO:0005515 [list] [network] protein binding (10915 loci) IPI	

KEGG*
ortholog [ortholog page] **Actr3**

subcellular localization*
cyto 7, cyto_nucl 4 (prediction for NP_005712.1)

coexpression

rank	cor	symbol
0	1.00	ACTR3
1	0.64	CAPZA1
2	0.63	ACTR2
3	0.63	ARPC5
4	0.62	MOBK1B
5	0.62	ARPC2
6	0.58	C2orf25
7	0.57	RAB1A
8	0.57	SNX6
9	0.57	ARPC3
10	0.57	RAB10
11	0.57	YME1L1
12	0.56	ABI1
13	0.56	TMOD3
14	0.55	ZC3H15
15	0.55	GNAI3
16	0.55	TMEM167
17	0.55	PREI3
18	0.55	TPM3
19	0.55	HRB
20	0.55	UBE2D3

network for coexpressed genes (Top 20)

network* in tissue
• Fetus
• **Neutrophil**

expression
all samples [expression pattern for all samples]

tissue-specific gene expression*

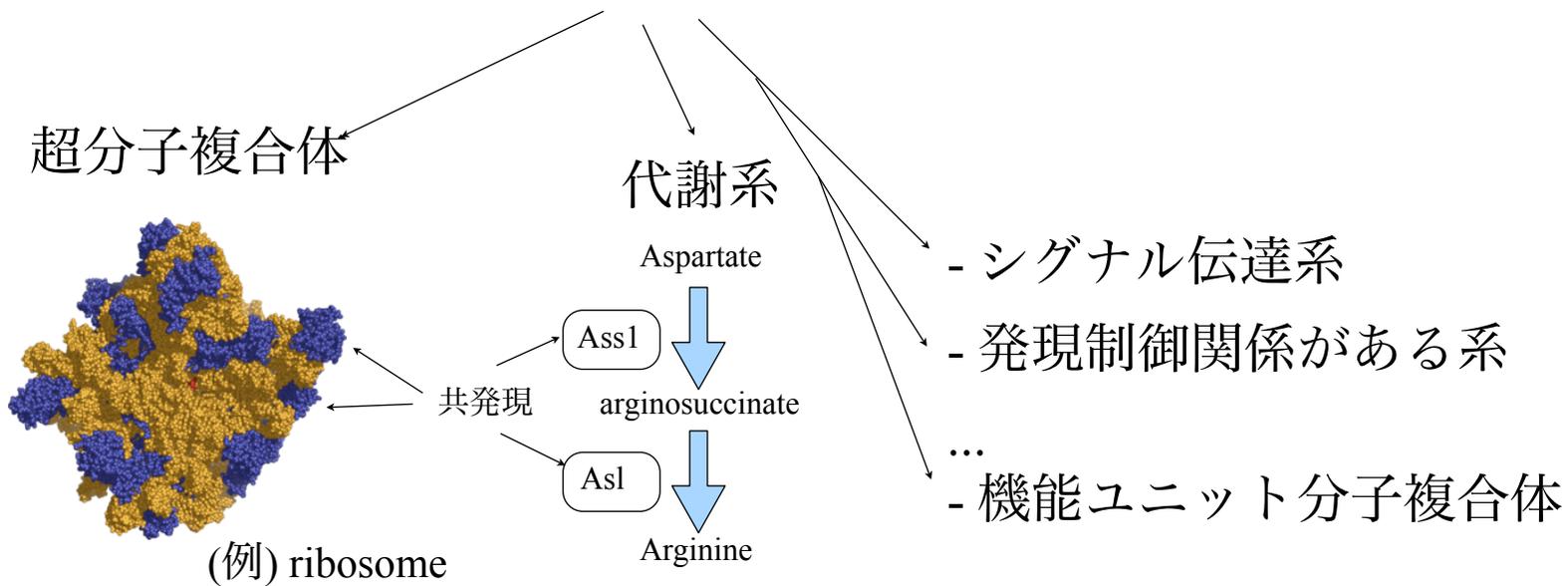
(D) Tissue-specific network page

Coexpressed gene network expressed in Neutrophil

DNAマイクロアレーデータを利用した相互作用ネットワークの構築

DNA Array Dataを利用用途

機能的に関係がある遺伝子は共発現することが多い

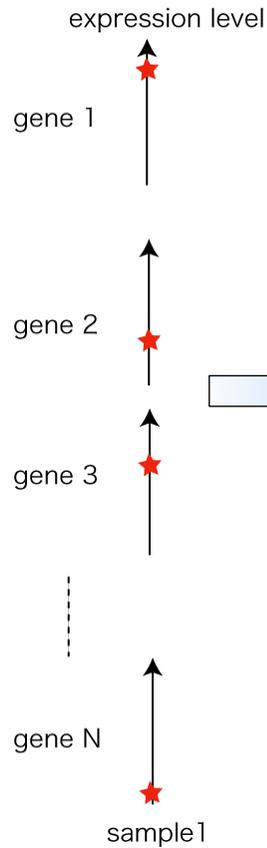


研究している遺伝子と共発現している遺伝子群を見つける
その中から新規（機能未知）でおもしろい遺伝子を同定する

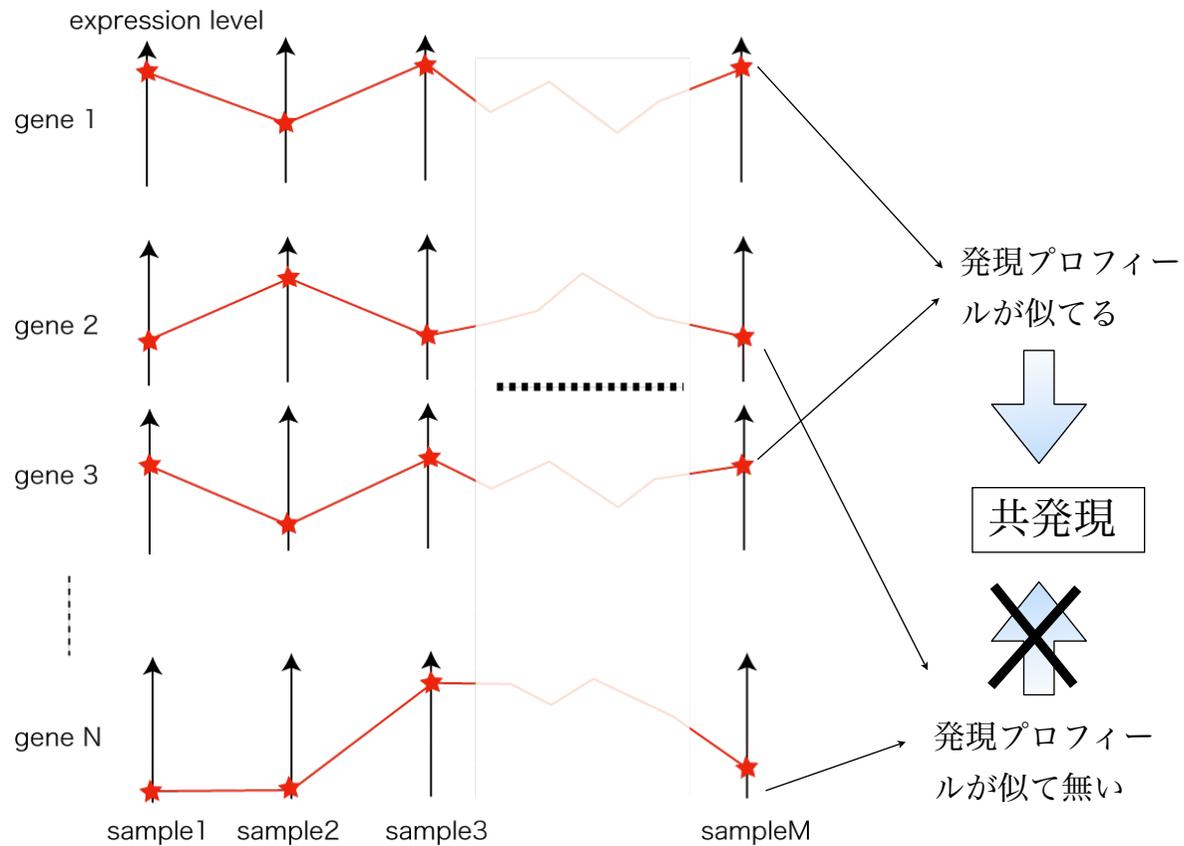


共発現度の定量化

1つのArray Data



多数のArray Data



発現プロフィールの相関係数 = 共発現度

Current status of COXPRES

Organism	Chip Name	Samples	# of loci
Homo sapiens	Human Genome U133 Plus 2.0 Array	5635	19777
Homo sapiens	Human Genome U133 Array Set HG-U133A	10839	12849
Homo sapiens	Human Genome U133 Array Set HG-U133B	2468	10269
Homo sapiens	Human Genome U95 Version [1 or 2] Set HG-U95A	3905	8878
Homo sapiens	Human Genome U133 Array Set HG-U133A	1218	-
Homo sapiens	Human HG-Focus Target Array	759	-
Mus musculus	Mouse Genome 430 2.0 Array	2200	21036
Mus musculus	Mouse Expression Array 430A and Mouse Genome 430A 2.0	2680	13225
Mus musculus	Murine Genome U74 Version 2 Set MG-U74A	4266	8912
Mus musculus	Mouse Expression Array 430B	780	-
Rattus norvegicus	Rat Genome 230 2.0 Array	886	11912
Rattus norvegicus	Rat Genome U34 Array Set RG-U34A	2371	-
Rattus norvegicus	Rat Expression Set 230 Array RAE230A	1166	-

★よく利用されているGeneChip (Sample数が多い) の中で、

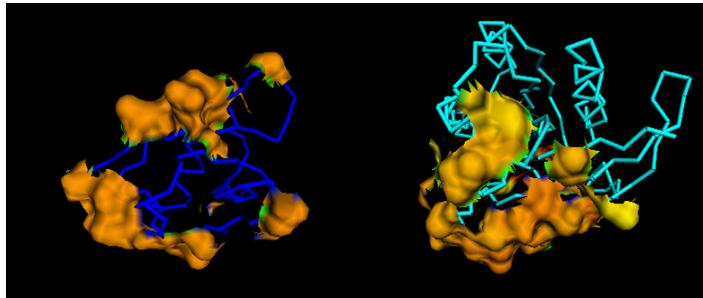
★locus数の多いGeneChipをターゲットとした

★Mouseとヒトの結果が利用可能

★Ratの計算はほぼ終了はしているので近日公開予定

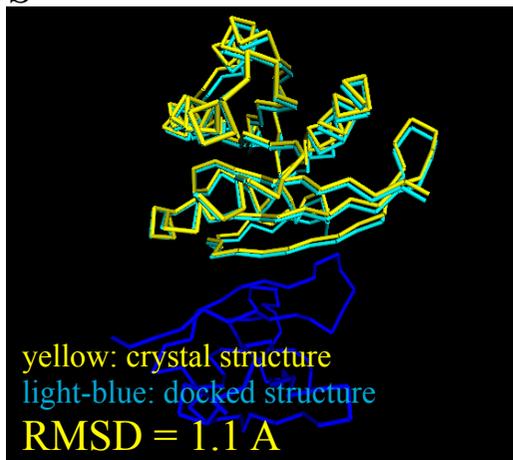
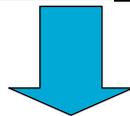


どのように相互作用するのか？ 複合体構造の予測



RID of
RalGD
S

Ras

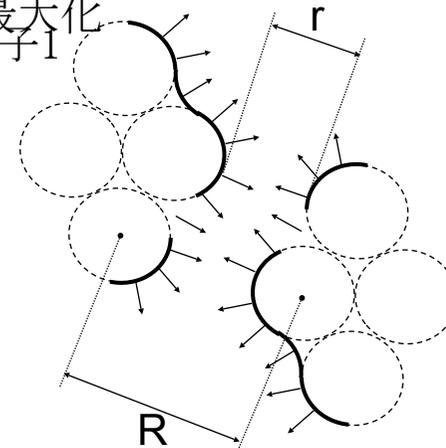


Ras / RID of aIGDS

保存度の重みをつけて表面の相補性の探索

$$F = aF_1 + bF_2$$

GA+MCで最大化
分子1



$$F_1 = \sum_{\text{all vertices pairs}} I^2 \times w_1^2 \times w_2^2 \times D$$

$$I = (-\vec{n}_1 \times \vec{n}_2 + 1) / 2$$

$$D = \begin{cases} 0.25 / r^2 & (r > 0.5) \\ 1.0 & (r \leq 0.5) \end{cases}$$

w_1, w_2 : weight of conservation, \vec{n}_1, \vec{n}_2 : normal vector 分子2

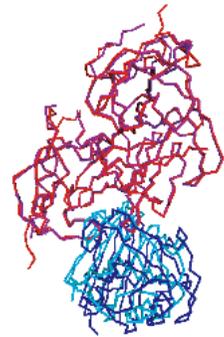
r : distance between vertices

$$F_2 = \sum_{\text{atom pairs}} 4.0((\sigma / R)^6 - (\sigma / R)^{12})$$

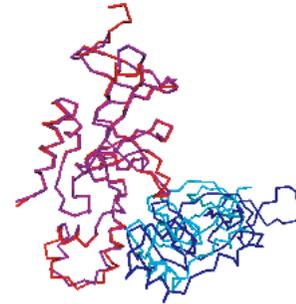
R : distance between atoms

CAPRIのmeetingにて招待講演

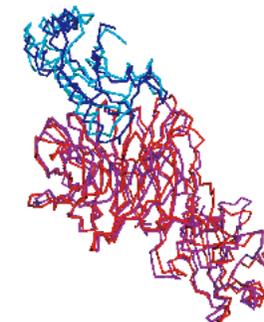
CAPRI (複合体の予測コンテスト) の結果



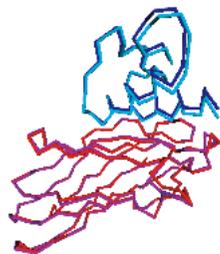
T18: xylanase inhibitor/xylanase



T21: Orc1/Sir1



T26: TolB/Pal



T12: cohesin/dockerin



T25: Arf1/ArfBD

赤と青：結晶構造、紫：予測構造

X線結晶構造に近いものを予測できるようになりつつある
候補構造の中から天然構造に近いものを選ぶ点に課題が残る

まとめ: 機能予測の様々なアプローチ

- 構造情報を利用した機能部位 (相互作用部位) の予測
 - 活性残基の予測
 - <http://p-cats.hgc.jp/p-cats>
 - 分子表面のDB
 - <http://eF-site.hgc.jp/eF-site>
 - 静電ポテンシャルの計算
 - <http://eF-site.hgc.jp/eF-surf>
 - 低分子結合部位の予測
 - <http://eF-surf.hgc.jp/eF-seek>
 - DNA相互作用部位
 - <http://pre-s.protein.osaka-u.ac.jp/~preds>
 - タンパク質相互作用部位
 - ホモ: 分類・解析
 - <http://pre-s.protein.osaka-u.ac.jp/~classppi>
 - ヘテロ: 表面保存度解析&docking
 - <http://pdbjets.protein.osaka-u.ac.jp>
- COXPRESdb: ヒトとマウスの共発現DB
 - <http://coxpresdb.hgc.jp>



eF-site
electrostatic surface of Functional-site

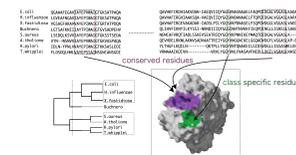
eF-surf

eF-seek

PreDs

Prediction of DNA-binding site

classPPI
classification of *homo* Protein-Protein Interfaces



COXPRESdb