2014.03.02
統合化推進プログラム 統合データ解析トライアル
研究成果報告会

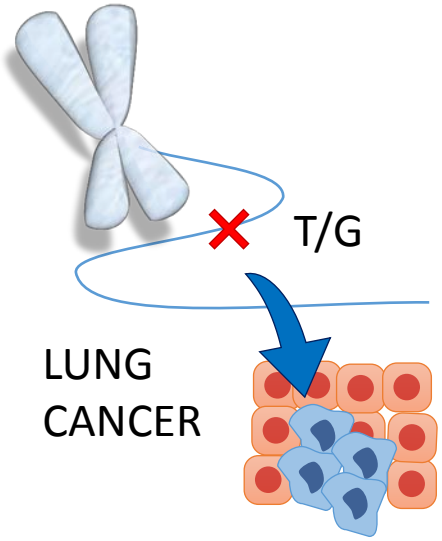
共起関係解析によるタンパク質の機能モジュール探索法の開発

九州工業大学・情報工・生命情報

藤井　聡

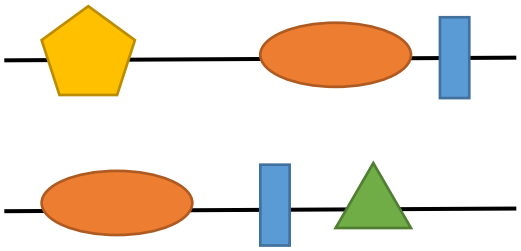
背景

疾病関連遺伝子



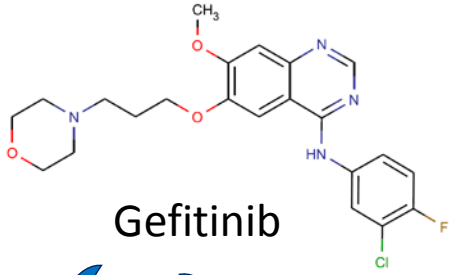
OMIM
NHGRI GWAS Catalog
Human Variation DB
etc...

ドメイン・モチーフ



PROSITE
Pfam
InterPro
CATH
SCOP
etc..

ドラッグターゲット



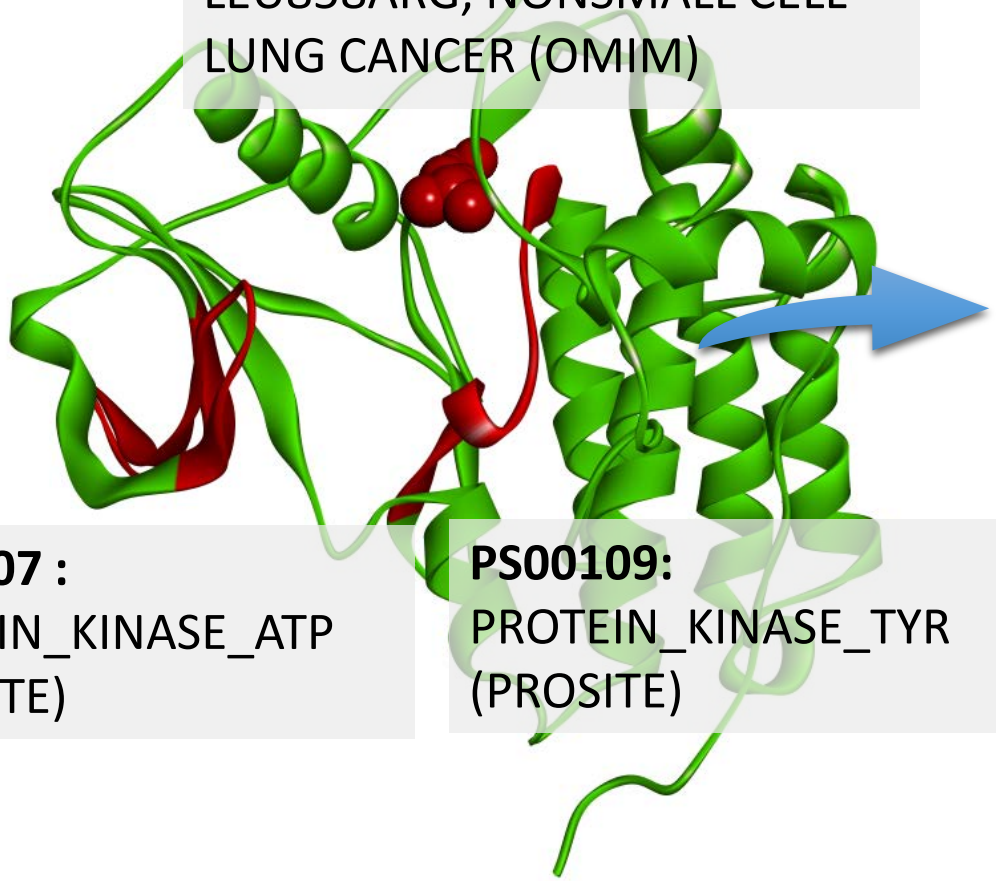
EGFR　CYP2D6　ABCG2

DrugBank
PubChem
KEGG DRUG
etc..

etc..

EGFR tyrosine kinase domainの結晶構造

LEU858ARG, NONSMALL CELL LUNG CANCER (OMIM)



機能モジュール:
3次元構造中で近傍に存在しており関係性が高い。

PS00107 :
PROTEIN_KINASE_ATP
(PROSITE)

PS00109:
PROTEIN_KINASE_TYR
(PROSITE)

目的

- 非常に多くのゲノム・プロテオームに関する情報の集積体（データベース）が構築されている。
- 疾病関連遺伝子やタンパク質の機能を示すドメインやモチーフ、薬剤などの相互作用部位を現すリガンド相互作用サイト、タンパク質−タンパク質相互作用サイトなどが挙げられる。
- しかし、単独では価値を理解することが難しいものも多い。

疾病情報やドメインなどの機能情報同士の間に浮かび上がる共起関係に注目し、構造と機能の有機的な結び付きを現す機能モジュールを探索する手法を開発することを目的とする。

方法

- 研究項目として、検出するターゲットはPROSITE, Pfamから得ることのできる機能ドメインと機能モチーフに対象を絞る。
- 共起関係は1対1の関係に絞って**解析手法を確立**を目指す。

1. データの取得と生成、データの整形
2. 共起関係の解析手法の確立
3. データベース作成ならび検索サイトの作成

- 最終的にその得られた共起関係のリストを、空間的な距離やその出現数、統計的な有意性を含めてデータベースとして公開するまでを第1目標とする。

データの取得と生成、データの整形

タンパク質の3次元構造データ

- PDBjより全PDB構造を取得した。

ドメイン・モチーフの情報

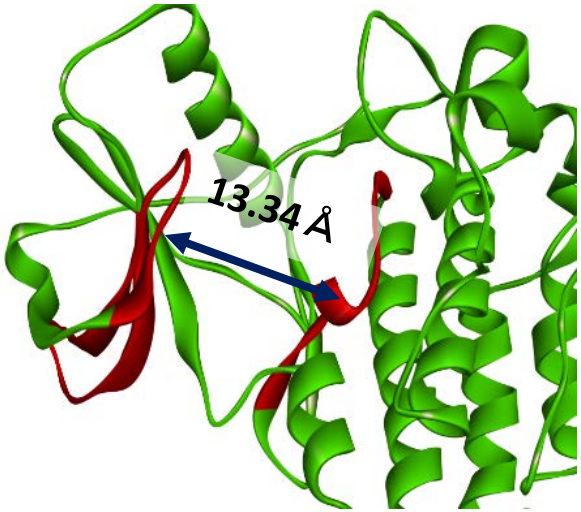
- タンパク質に存在するドメイン・モチーフの情報はPROSITEから得た。
  - 元々価値の高いドメインやファミリー同士の共起関係だけでなく、情報として価値の低い機能サイトとの共起関係についてなども同時に検証することを想定してPROSITEを最初のターゲットにした。
- ドメイン・モチーフの位置は、配列に対してPROSITEのps_scanにより配列に対して予測計算を行い求めた。

タンパク質構造の冗長化

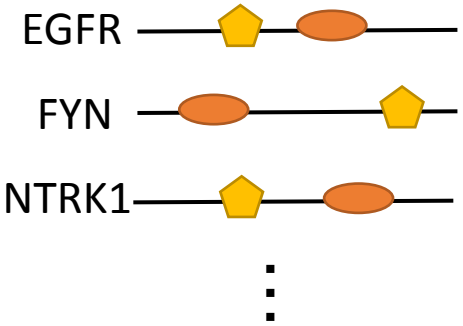
- タンパク質構造情報の冗長化はEMBL-EBI/SIFTSを利用して行ってた。
  - タンパク質の構造は同じタンパク質から複数得られていたり、タンパク質の一部分のみの構造が得られていたりしているので冗長化を行う必要がある。

共起関係の検出

① タンパク質構造中で近傍に存在する共起関係の検出



② タンパク質全体で高頻度に見られる共起関係の検出



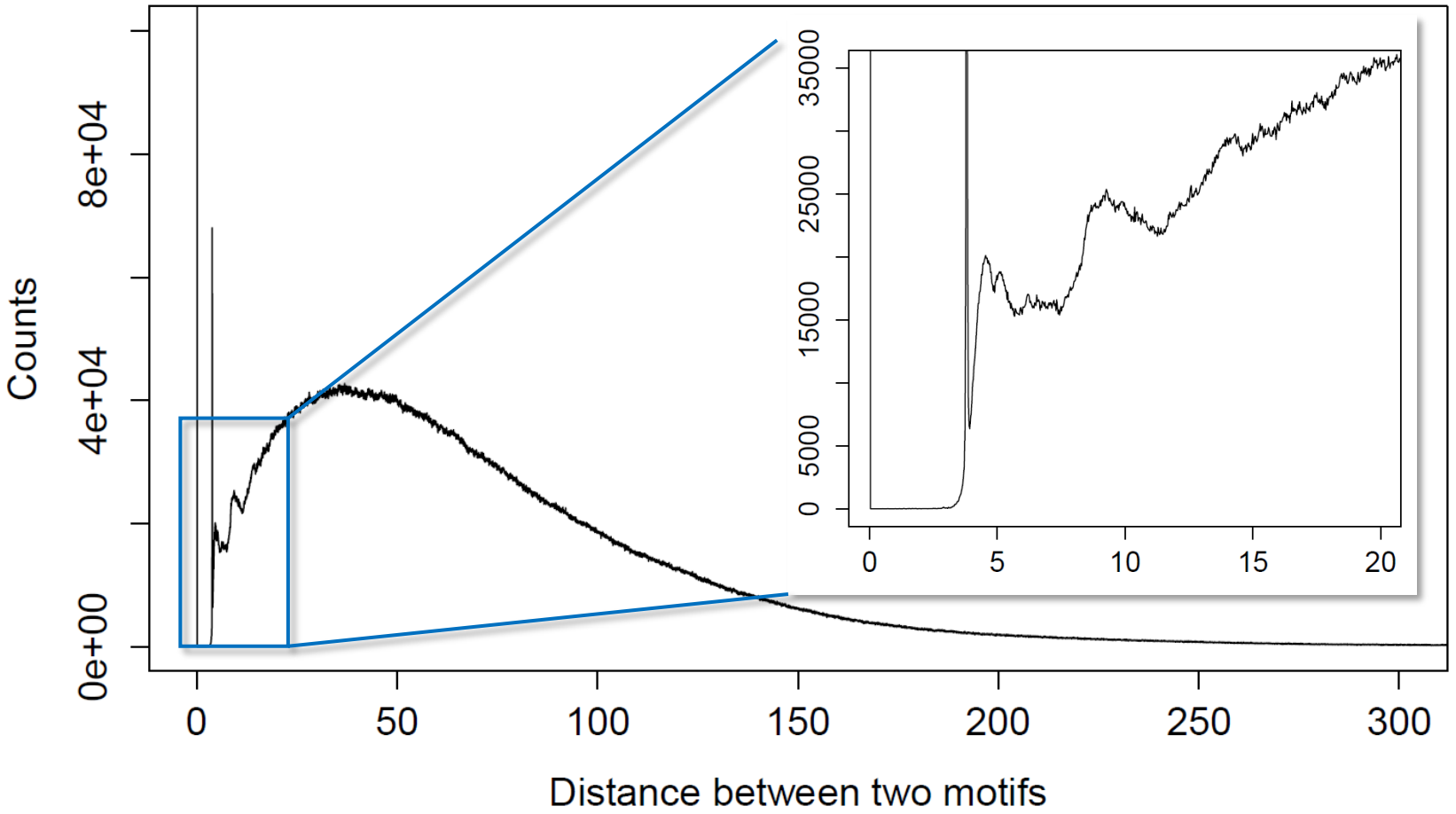
③ ①＋②両方の条件に合致する共起関係の検出

結果:
2つのモチーフ同士の距離

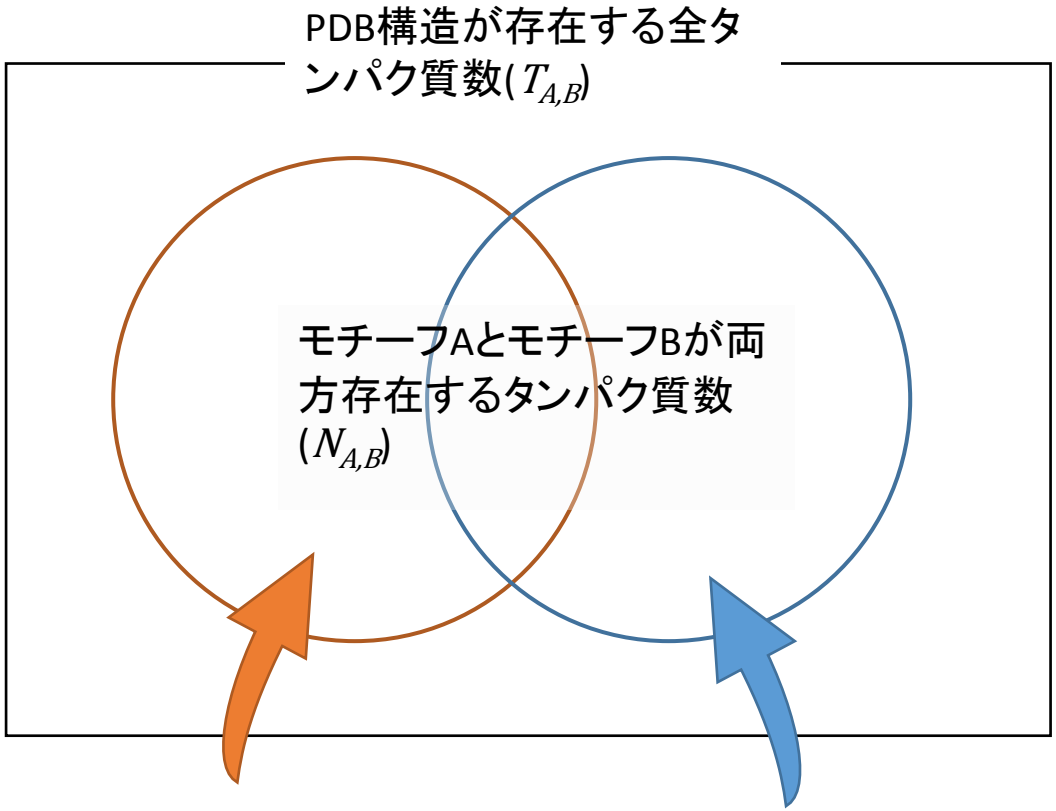
Uniprot.ID_A	Prosite_A	Start_A	End_A	Uniprot.ID_B	Prosite_B	Start_B	End_B	N_pdb	Cα.distance (min)	Cα.distance (ave.)	Cα.distance (s.d.)
O87988	PS00005	66	68	O87988	PS00006	211	214	12	15.91	34.11	18.88
P16932	PS00008	143	148	P16932	PS00009	152	155	16	7.15	7.6	0.24
P96110	PS00005	189	191	P96110	PS00008	192	197	108	3.77	49.43	25.95
O66608	PS00006	266	269	O66608	PS00009	17	20	4	7.94	11.41	3.94
D2YW38	PS00005	73	75	D2YW38	PS00008	90	95	3	7.55	7.61	0.06
P24183	PS00006	63	66	P24183	PS00008	67	72	2	3.79	3.82	0.03
Q9XG81	PS00008	64	69	Q9XG81	PS00118	82	89	17	6.94	14.86	11.03

Prosite モチーフ数:	2,006		
総PDBchain数:	221,581	総タンパク質数:	32,042
モチーフHit数:	3,163,170		1,116,766
モチーフ組み合わせ数:	164,122,109		7,945,374

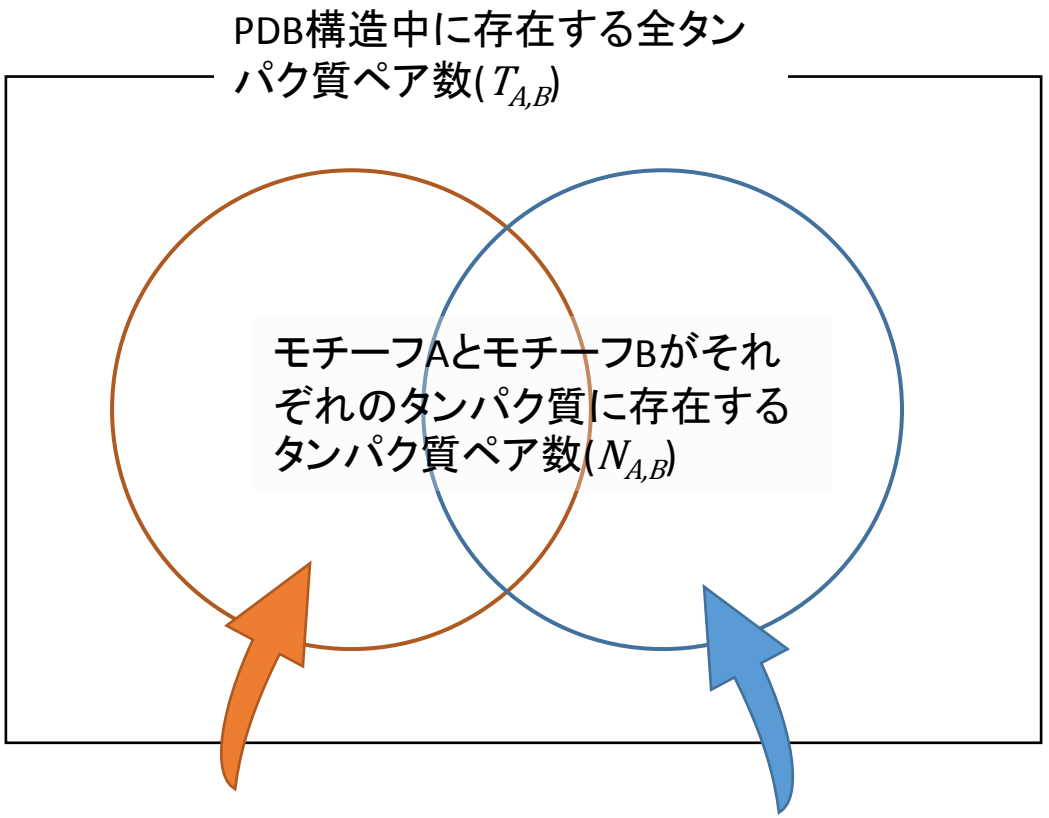
モチーフ間のCα最短距離の分布

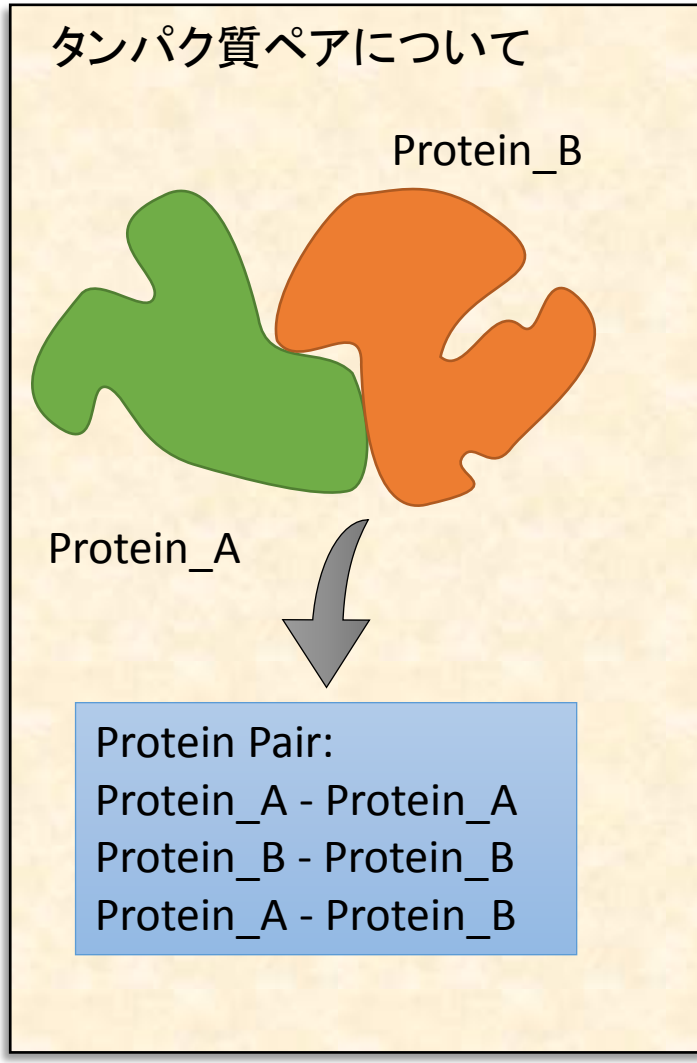


共起関係の検出法 ~Intra-molecule~



共起関係の検出法 ~Inter-molecule~





結果：
共起関係のEnrichment

Motif_combination	$N_{A,B}$	N_A	N_B	$T_{A,B}$	FE	p-value	FDR
PS00029_PS01281	1	676	5	32042	9.48	0.00426	0.007658
PS00008_PS01201	3	24311	3	32042	1.32	0	0
PS00004_PS00433	6	7689	9	32042	2.78	0.001029	0.001979
PS00006_PS01132	23	26003	23	32042	1.23	0	0
PS00006_PS50995	37	26003	39	32042	1.17	0.002908	0.005322
PS00370_PS00742	5	6	7	32042	3814.52	0	0
PS00004_PS00975	3	7689	6	32042	2.084	0.032538	0.05088
PS00008_PS00128	16	24311	24	32042	0.88	0.796727	0.850721
PS00006_PS01028	4	26003	4	32042	1.23	0	0

*p-valueは超幾何分布により算出

FDR < 0.05: 11,885 entries

$$P(X = k) = 1 - \sum_{k}^{N_{A,B}} \frac{\binom{N_B}{k}\binom{T_{A,B} - N_B}{N_A - k}}{\binom{T_{A,B}}{N_A}}$$

FDRはBH法により計算

距離と共起のEnrichment両方での絞り込み

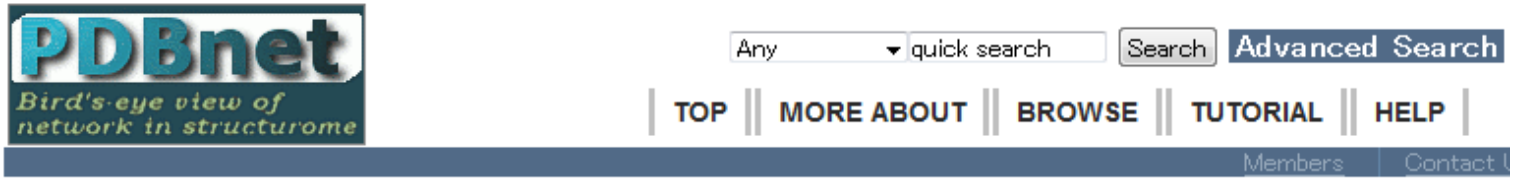
含まれているモチーフ同士のCα最短距離が3.5 - 6.0 Å内にある。

共起のEnrichmentにおける有意性がFDR< 0.05である。

$$\frac{\text{距離によって共起が検出されたタンパク質数}}{\text{Enrichmentのよって共起が検出されたタンパク質数}} \geq 0.8$$

PDBnet -Co-occurrence Search Tool-

http://dna00.bio.kyutech.ac.jp/pdbnet/co-search.php



Search Cooccurrence of PROSIE domain

Please fill or choose necessary entries below.

Fold Enrichment >	2.0
Enrichiment FDR <	0.05
Minimun number >	2
Ca Distance:	3.0 to 6.0
Dist/Enrich >	0.8
Target:	Inter-molecular

Search Clear

- 共起に関する条件を入力すると、その条件で絞ったPROSITEモチーフの共起リストを表示する。
- 共起のリストから、各々の共起を持つタンパク質、さらにPDB構造までリンクで追うことができる。
- Jmolにより構造上の共起関係を確認することも可能。
- 現状は、上記のような一方向の検索しかできない。タンパク質名やモチーフの名前等から検索はできない。
- 将来的にはPDBnetからこれらの情報へアクセスできるようにする予定である。

Webツール デモ

実例:
PS00107_PS01351の共起関係

MotifA:
PS00107:
Protein kinases ATP-binding region signature.
MotifB:
PS01351:
MAP kinase signature.

Show 10 entries Search:

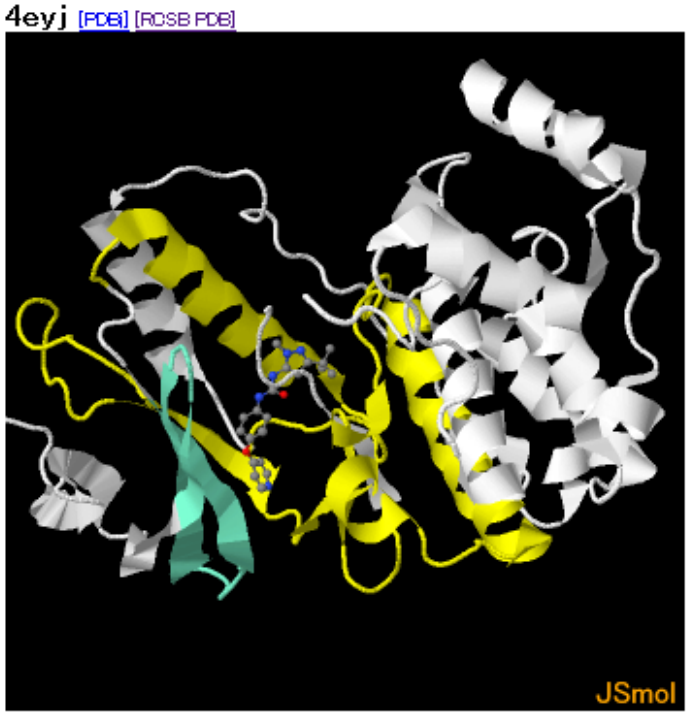
ProteinA	StartA	EndA	ProteinB	StartB	EndB	N_{pdb}	$Dist_{min}$	$Dist_{ave}$	$Dist_{sd}$
A3FQ79	19	43	A3FQ79	48	148	2	4.60	4.62	0.01
A9UJZ9	31	55	A9UJZ9	60	163	2	4.58	4.60	0.03
B6KP12	151	175	B6KP12	180	283	1	4.45	4.45	0.00
O15264	31	55	O15264	60	162	4	4.34	4.51	0.12
P16892	19	43	P16892	47	149	9	4.50	4.55	0.05
P27361	48	72	P27361	76	178	2	4.33	4.36	0.04
P28482	31	55	P28482	59	161	20	4.01	4.41	0.16
P47811	30	54	P47811	59	162	21	3.97	4.31	0.15
P49137	30	54	P49137	59	162	1	4.34	4.34	0.00
P53778	33	57	P53778	62	165	2	4.27	4.27	0.00

Showing 1 to 10 of 17 entries Previous Next

実例:
PS00107_PS01351の共起関係

MotifA:
PS00107:31 – 55
Protein kinases ATP-binding region signature.
O15264
Name:Mitogen-activated protein kinase 13(MAPK13)
Organism:Homosapiens

MotifB:
PS01351:60 – 162
MAP kinase signature.
O15264
Name:Mitogen-activated protein kinase 13(MAPK13)
Organism:Homosapiens



Show 10 entries Search:

ChainA	StartA	EndA	ChainB	StartB	EndB	$Dist_{min}$	Structure
3coiA	31	55	3coiA	60	162	4.55	Jmol
4exuA	31	55	4exuA	60	162	4.55	Jmol
4eyjA	31	55	4eyjA	60	162	4.34	Jmol
4eymA	31	55	4eymA	60	162	4.61	Jmol

Showing 1 to 4 of 4 entries Previous Next

実例:
PS00017_PS00152の共起関係

MotifA:
PS00017:
ATP/GTP-binding site motif A (P-loop).
MotifB:
PS00152:
ATP synthase alpha and beta subunits signature.

Show 10 entries Search:

ProteinA	StartA	EndA	ProteinB	StartB	EndB	N_{pdb}	$Dist_{min}$	$Dist_{ave}$	$Dist_{sd}$
B7UMA6	177	184	B7UMA6	356	365	2	6.50	6.51	0.02
O57728	234	617	O57728	804	813	11	5.98	6.76	0.51
P00825	172	179	P00825	363	372	2	6.56	6.66	0.13
P00825	172	179	P06450	356	365	2	9.45	17.76	11.75
P00829	206	213	P00829	396	405	357	5.93	44.09	36.50
P00829	206	213	P00829	395	404	3	6.54	29.19	20.44
P00829	206	213	P19483	406	415	360	6.52	43.25	36.32
P00830	190	197	P07251	400	409	558	6.86	73.82	42.14
P00830	190	197	P00830	379	388	558	5.73	74.29	41.98
P06450	170	177	P00825	363	372	2	7.02	15.92	12.58

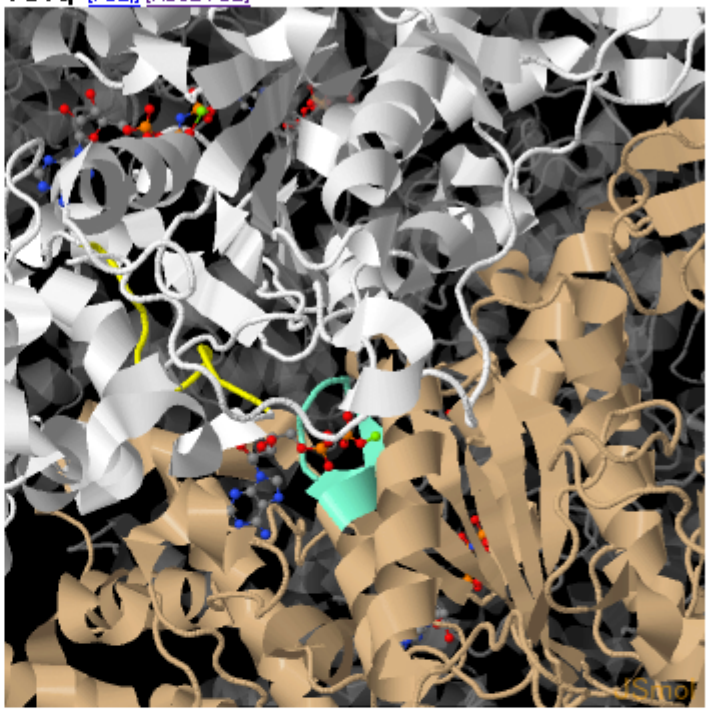
Showing 1 to 10 of 33 entries ◀ Previous Next ▶

実例:
PS00017_PS00152の共起関係

MotifA:
PS00017:206 – 213
ATP/GTP-binding site motif A (P-loop).
P00829
Name:ATP synthase subunit beta, mitochondrial(ATP5B)
Organism:BosTauurus

MotifB:
PS00152:406 – 415
ATP synthase alpha and beta subunits signature.
P19483
Name:ATP synthase subunit alpha, mitochondrial(ATP5A1)
Organism:BosTauurus

1e1q [PDB] [RCSB PDB]



Show 10 entries Search:

ChainA	StartA	EndA	ChainB	StartB	EndB	$Dist_{min}$	Structure
1e1qD	156	163	1e1qC	363	372	7.79	Jmol
1e1qE	156	163	1e1qA	363	372	10.03	Jmol
1e1qF	156	163	1e1qB	363	372	9.45	Jmol
1efrD	156	163	1efrC	363	372	7.95	Jmol
1efrE	156	163	1efrA	363	372	9.87	Jmol

デモ終了

本ツールの汎用性について

- 現状ではPROSITEにしか対応していないのでまだ汎用性は低い。
- 共起を検出する手法としては単純なので、将来的には、原子単位、アミノ酸配列単位で付加されている情報に対しても適用できるだろう。

これまで作られてきた統合データベースの新たな活用法について

- 今回はタンパク質構造（PDB）を情報を結びつけるための媒体として活用した。それにより、情報を3次元の構造中に表すことにより有機的に結び付けることができた。

本ツールを活用した有用な知識の発見について

- 客観的な評価指標を示せていないので評価するのは難しいが、いくつか具体的な例を観察すると、活性部位付近に共起関係が存在するような、抽出されてくるべき結果は抽出できていた。
- 今後、機能未知なタンパク質、領域等を実際に検証することで有用な知見の発見につながる可能性もあるだろう。

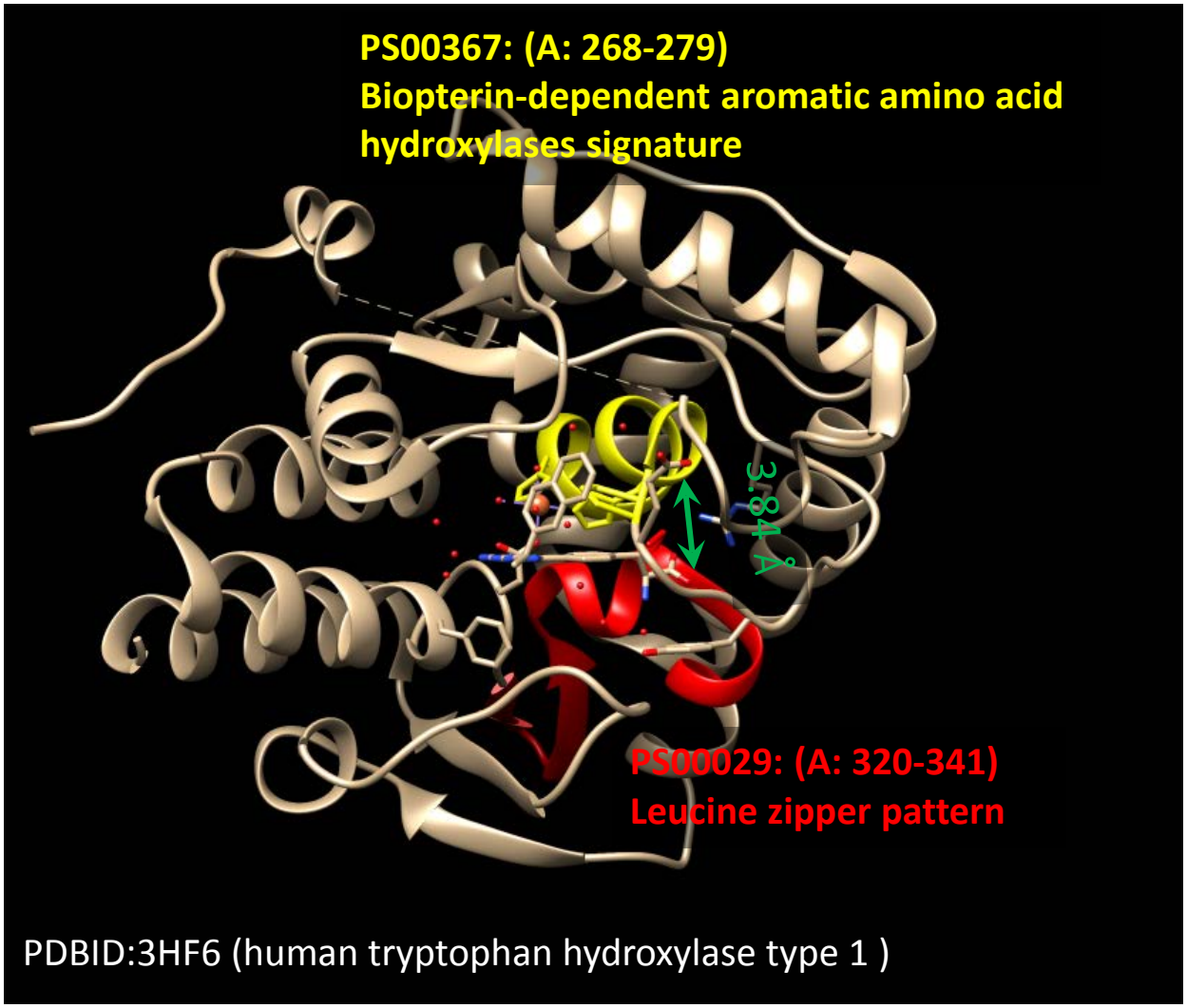
今後の本研究の将来性

- 空間における集積性を調べる空間統計学により評価したり、物理化学で扱われるPMF(Potential Mean Force)により評価したりすれば、タンパク質の立体構造を更に生かした結果を導き出せるだろう。
- タンパク質構造を使って任意の情報を結びつける手法の1つのスタンダードとしたい。

まとめ

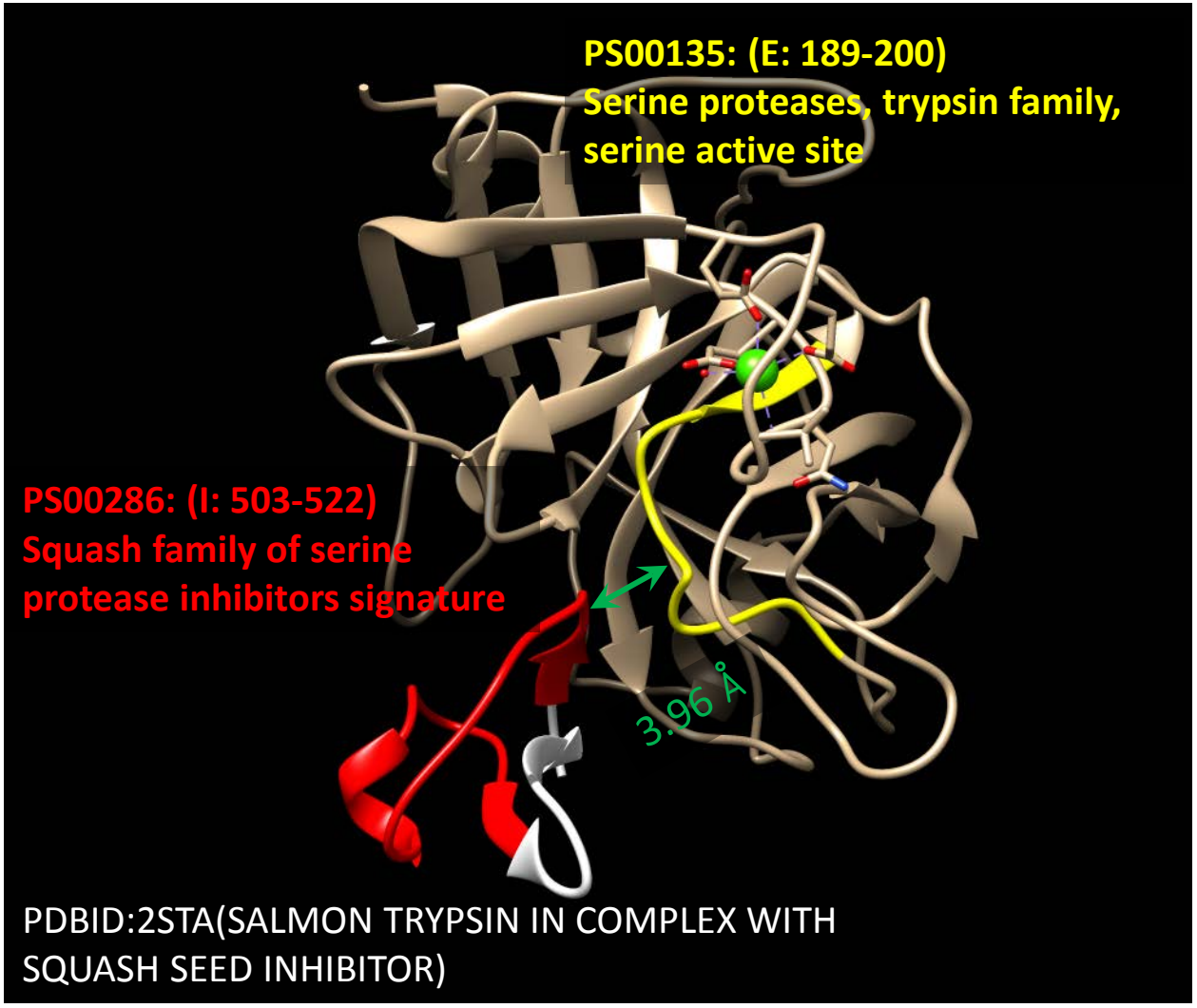
- ほぼ当初の研究開発計画通りに進行した。PROSITEに対する共起関係の検出を行い、得られた共起関係のリストをデータベースとして公開した。
- 発展的な展開としては、2つのタンパク質の相互作用による共起関係の検出には対応した。
- 今回の結果についての客観的な評価を示せていないが、主観的な観察によっては尤もらしい結果は得られていた。
- 今後汎用化、共起検出の統計手法の改良によって、更に有用な手法となるだろう。

実例：
PS00029_PS00367の共起関係



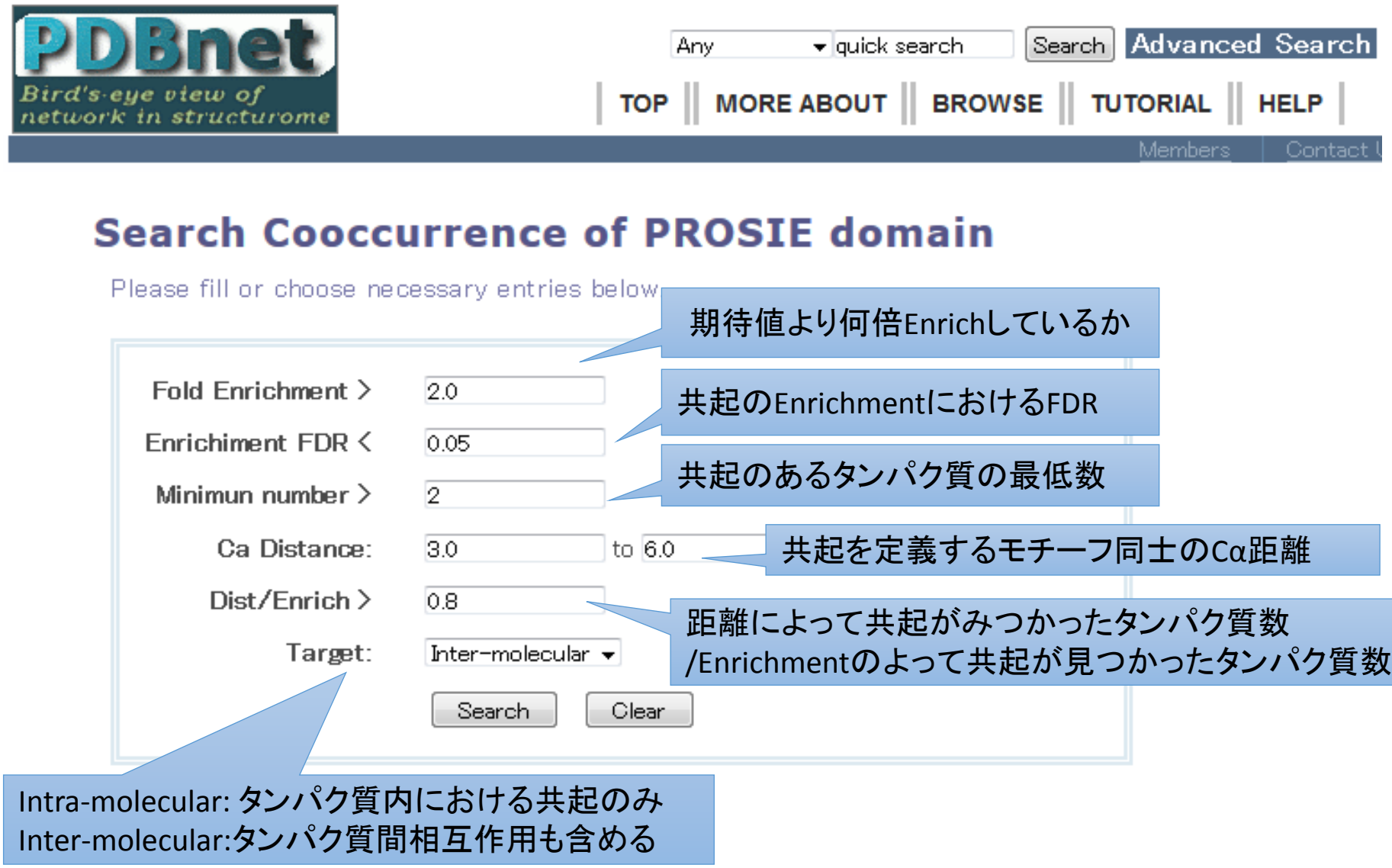
P17752	Tryptophan 5-hydroxylase 1 (Human)
	3hf6A
	1mlwA
	3hf8A
	3hfbA
P04177	Tyrosine 3-monooxygenase (Rat)
	1tohA
	2tohA
P70080	Tryptophan 5-hydroxylase 1 (Chicken)
	3e2tA

実例：
PS00286_PS00135の共起関係

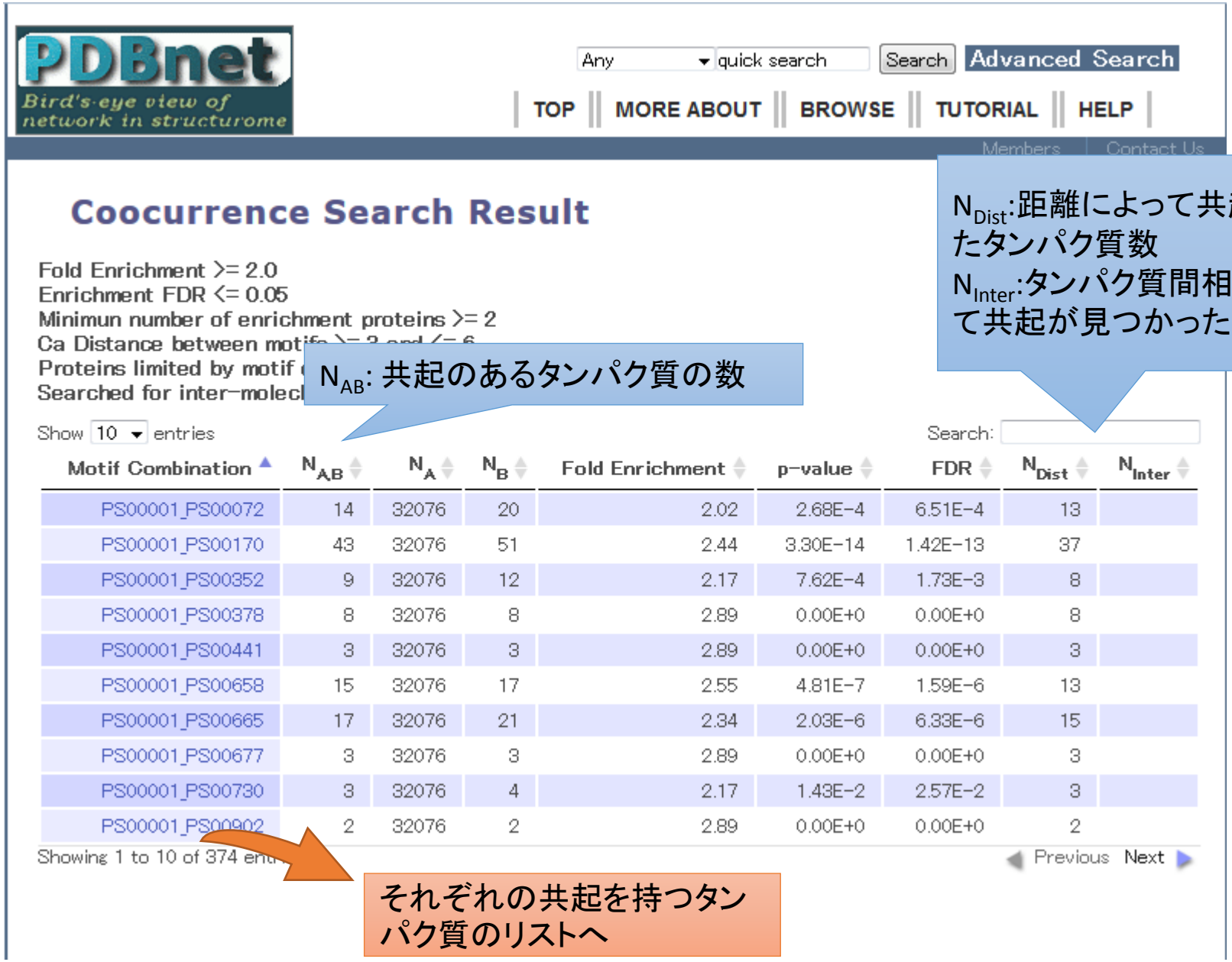


P01074	2staI	P35031	2staE
P01074	1ppeI	P00760	1ppeE
P10293	2btcI	P00760	2btcE
P12071	1h9iI	P00761	1h9iE
P10295	1f2sI	P00760	1f2sE
P10293	2stbI	P35031	2stbE
P30709	1mctI	P00761	1mctA

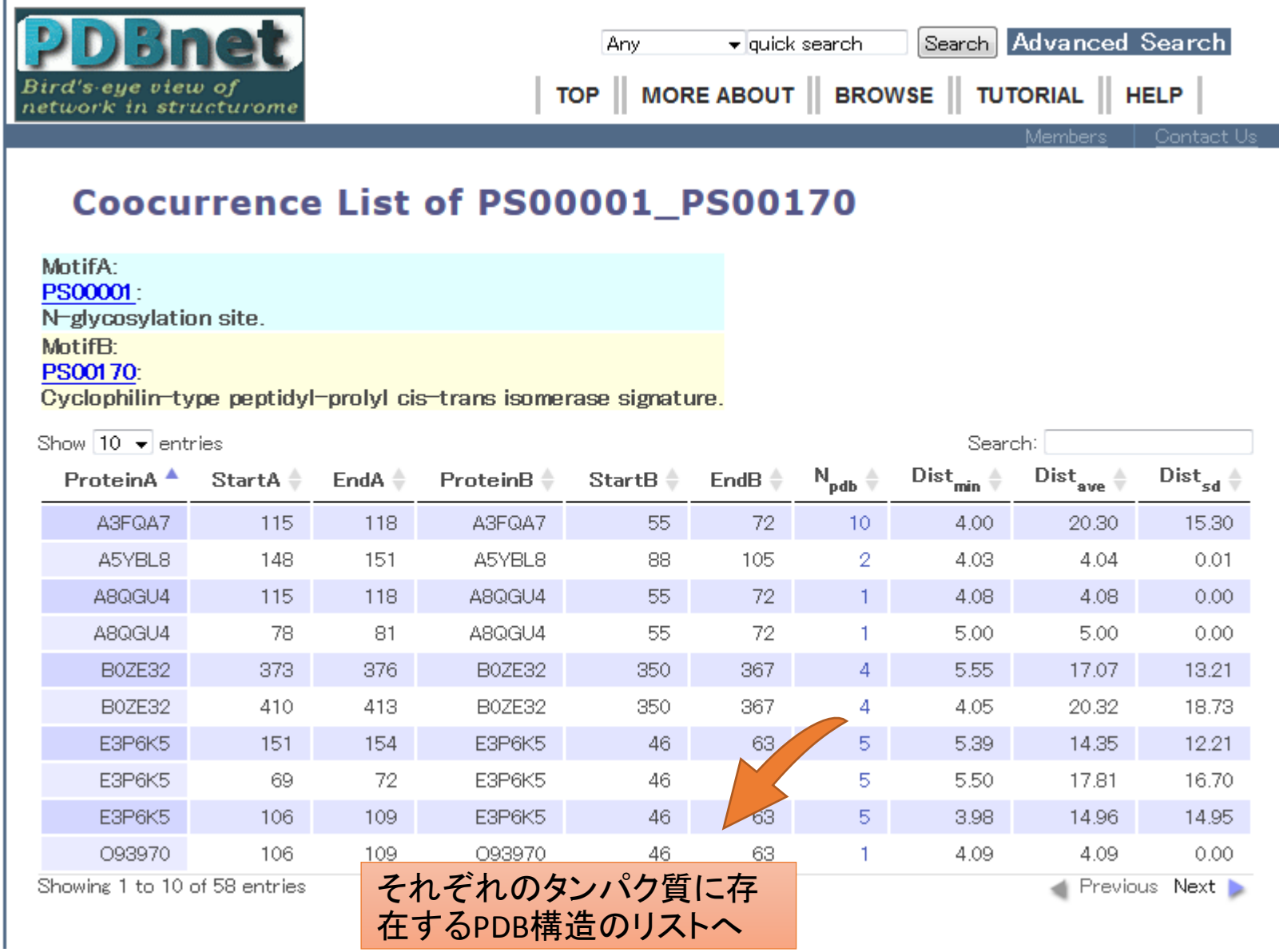
共起関係のデータベースにアクセスする検索ページ



共起関係の検索結果表示のページ



あるモチーフの共起関係を持つタンパク質のリスト



あるモチーフの共起関係を持つある1つのタンパク質についてのPDB構造のリスト

PDBnet
Bird's-eye view of network in structurome

Any quick search Search Advanced Search

TOP | MORE ABOUT | BROWSE | TUTORIAL | HELP

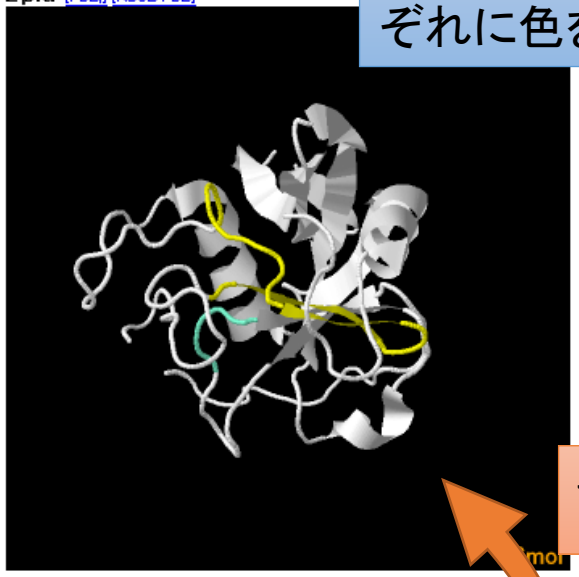
Members | Contact Us

Coocurrence Motifs in Structures

MotifA:
PS00001:115 – 118
N-glycosylation site.
A3FQA7
Name:Peptidyl-prolyl cis-trans isomerase(cgd2_4120)
Organism:CryptosporidiumparvumIowaII

MotifB:
PS00170:55 – 72
Cyclophilin-type peptidyl-prolyl cis-trans isomerase signature.
A3FQA7
Name:Peptidyl-prolyl cis-trans isomerase(cgd2_4120)
Organism:CryptosporidiumparvumIowaII

2plu [PDB] [RCSB PDB]



共起しているそれぞれのmotifそれぞれに色を付けて表現している。

Jmolを使って分子構造を見る。

Show 10 entries Search:

ChainA	StartA	EndA	ChainB	StartB	EndB	$Dist_{min}$	Structure
2pluA	131	134	2pluA	71	88	4.00	Jmol
2poyA	131	134	2poyA	71	88	4.09	Jmol
2poyB	131	134	2poyB	71	88	4.08	Jmol
2poyC	131	134	2poyC	71	88	4.01	Jmol

Showing 1 to 4 of 4 entries Previous Next