2013.11.29
統合データ解析トライアル・中間激励会

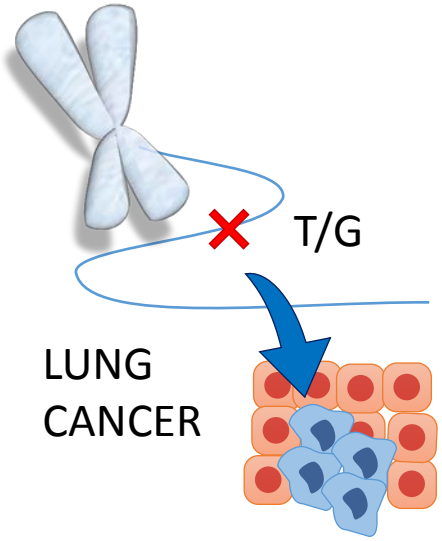
共起関係解析によるタンパク質の機能モジュール探索法の開発

九州工業大学・情報工・生命情報

藤井　聡

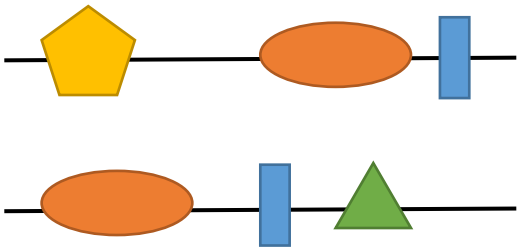
背景

疾病関連遺伝子



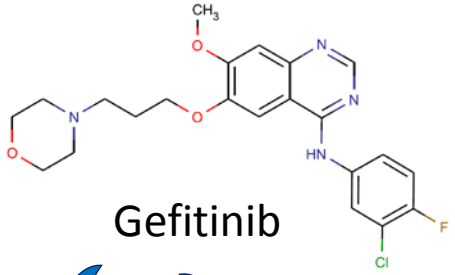
OMIM
NHGRI GWAS Catalog
Human Variation DB
etc...

ドメイン・モチーフ



PROSITE
Pfam
InterPro
CATH
SCOP
etc..

ドラッグターゲット

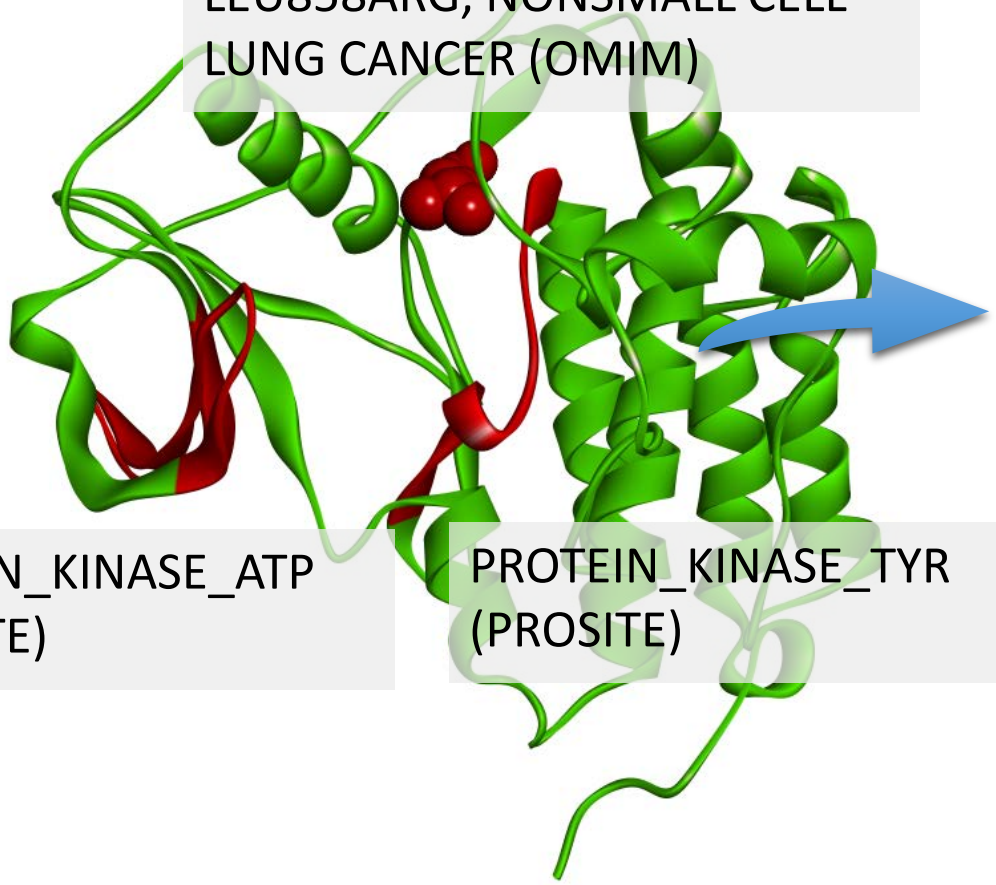


EGFR　CYP2D6　ABCG2

DrugBank
PubChem
KEGG DRUG
etc..

etc..

EGFR tyrosine kinase domainの結晶構造



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。

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、タンパク質−タンパク質相互作用などが挙げられる。
- 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する。

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。

研究開発の主なスケジュール

研究開発項目	平成25年 10月	平成25年 11月	平成25年 12月	平成26年 1月
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開発	←	—	—	—
3. データベース作成ならび検索サイトの作成			←	→

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構造を取得する。

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から得る。
- ドメイン・モチーフの位置についての情報が存在しないので、タンパク質配列に対してPROSITEのps_scanにより配列に対して予測計算を行い求める。
- Pfamのドメイン情報も進行状況次第で取り入れる。

タンパク質構造の冗長化

- 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。
- すでに前研究でUniprotを利用しタンパク質構造情報の冗長化は行っているので、それをドメイン・モチーフにも適応させる。

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構造を取得する。

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から得る。
- ドメイン・モチーフの位置についての情報が存在しないので、タンパク質配列に対してPROSITEのps_scanにより配列に対して予測計算を行い求める。
- Pfamのドメイン情報も進行状況次第で取り入れる。

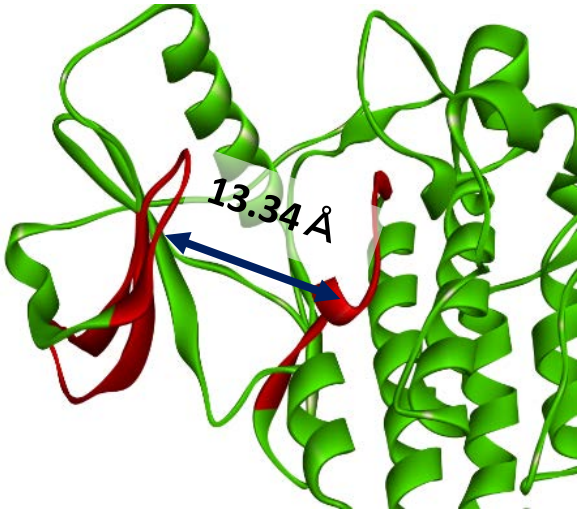
タンパク質構造の冗長化

- タンパク質の構造
- すでに前研究でUr

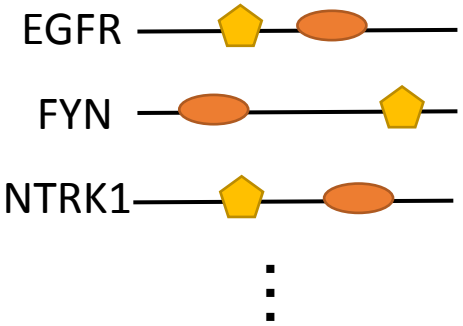
問題点：モチーフの位置が微妙にずれることがある。
⇒PDB chainをタンパク質配列にBlast等でアライメントして位置を定める必要あり。

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	PDB.chain.ID_A	Prosite_A	Start_A	End_B	Uniprot. ID_B	PDB.chain.ID_B	Prosite_B	Start_B	End_B	Cα.dist (min)	Cα.dist (ave.)
P09326	2edoA	PS00001	25	28	P09326	2edoA	PS00006	29	32	3.804	10.017
P15424	3i61A	PS00008	560	565	P15424	3i61A	PS51194	355	512	9.029	27.395
P63577	3p32A	PS00005	35	37	P63577	3p32A	PS00006	31	34	3.838	7.817
P38501	3h4fA	PS00006	5	8	P38501	3h4fB	PS00008	225	230	35.963	39.736
Q9X273	3azrB	PS00008	252	257	Q9X273	3azrB	PS00008	293	298	8.404	16.302

Prosite モチーフ数: 2,006

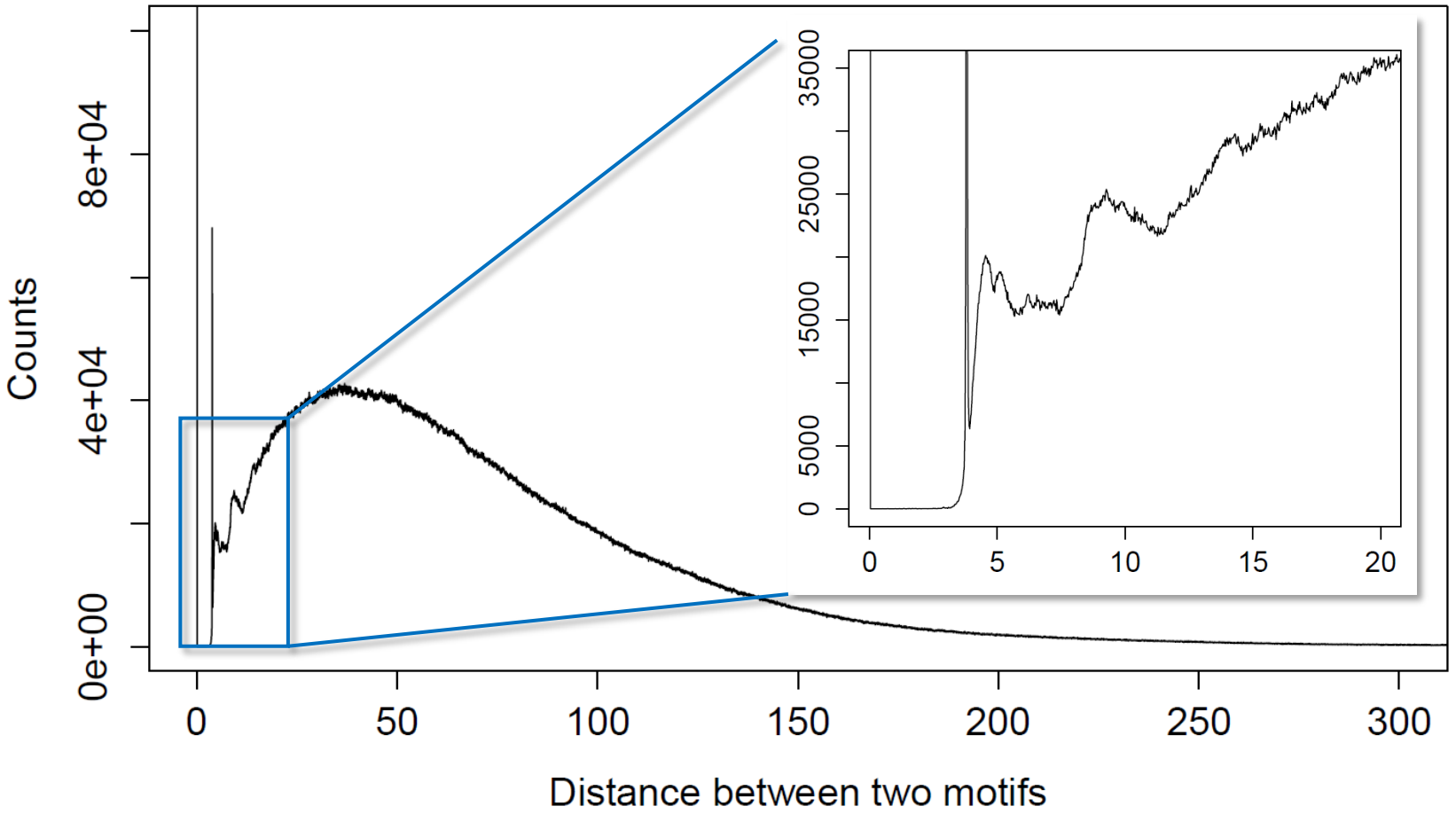
総PDBchain数: 221,581

総タンパク質数: 28,865

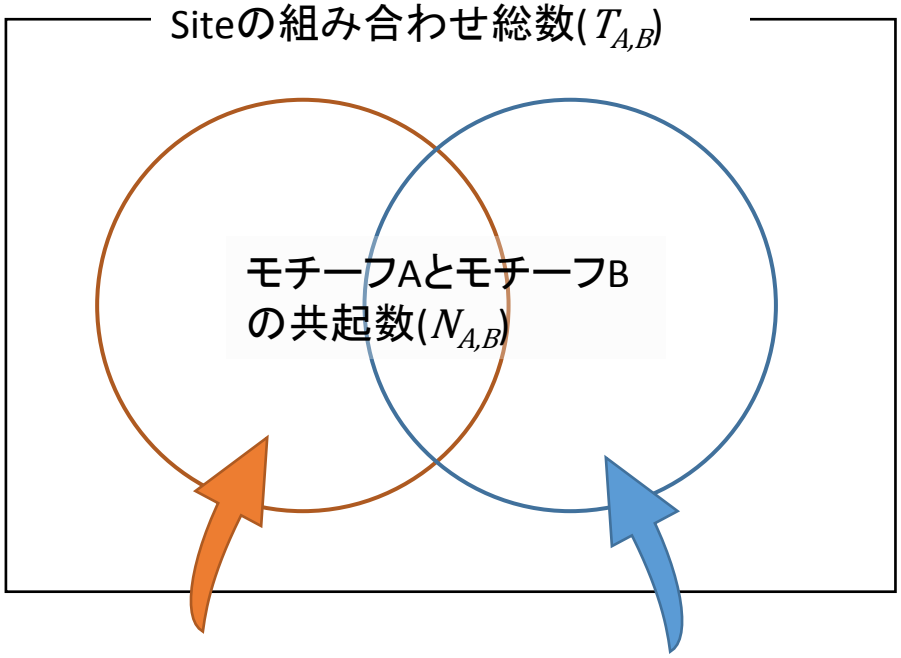
モチーフHit数: 3,163,170

モチーフ組み合わせ数: 146,673,695

モチーフ間のCα最短距離の分布



共起関係の検出法



Hitしたモチーフ AとモチーフBの潜在Siteの組み合わせ数(N_A)

Hitしたモチーフ BとモチーフAの潜在Siteの組み合わせ数(N_B)

[i]というPDB構造のchain [j]におけるモチーフAの潜在Site数($S_{A,i,j}$)

$$S_{A,i,j} = \begin{cases} L_{i,j} - m_A + 1, & L_{i,j} \geq m_A \\ 0, & L_{i,j} < m_A \end{cases}$$

$L_{i,j}$: [i]というPDB構造のchain [j]の配列長

m_A:モチーフAの長さ

$$T_{A,B} = \sum_{i}^{All\ PDBs} \left(\sum_{j}^{chains} S_{A,i,j} \times \sum_{j}^{chains} S_{B,i,j} \right)$$

$$N_A = \sum_{i}^{motif A\ hitted\ PDBs} \sum_{j}^{chains} S_{B,i,j}$$

$$N_B = \sum_{i}^{motif B\ hitted\ PDBs} \sum_{j}^{chains} S_{A,i,j}$$

結果：
共起関係のenrichment

Motif_combination	$N_{A,B}$	N_A	N_B	$T_{A,B}$	p-value	FDR
PS00115_PS51133	2362	8917230	473465	91929148880	0	0
PS00783_PS01106	77	965236	329600	94485660482	0	0
PS00006_PS51388	396	1.27E+09	17486	85477197004	9.99E-16	3.55E-15
PS00163_PS50310	16	395938	1511892	1.00611E+11	0.000159	0.000356
PS00008_PS01194	5827	1.54E+09	364364	97893945372	0.073852	0.130755
PS00585_PS01073	1	611674	310966	91321151148	0.615961	0.922663
PS00008_PS50031	21	1.25E+09	2252	79521254315	0.993606	1
PS00008_PS50810	48	1.14E+09	5461	73018045354	0.999994	1

*p-valueは超幾何分布により算出

FDR < 0.05: 12,867entries

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

FDRはBH法により計算

距離と共起のEnrichmentの両方での

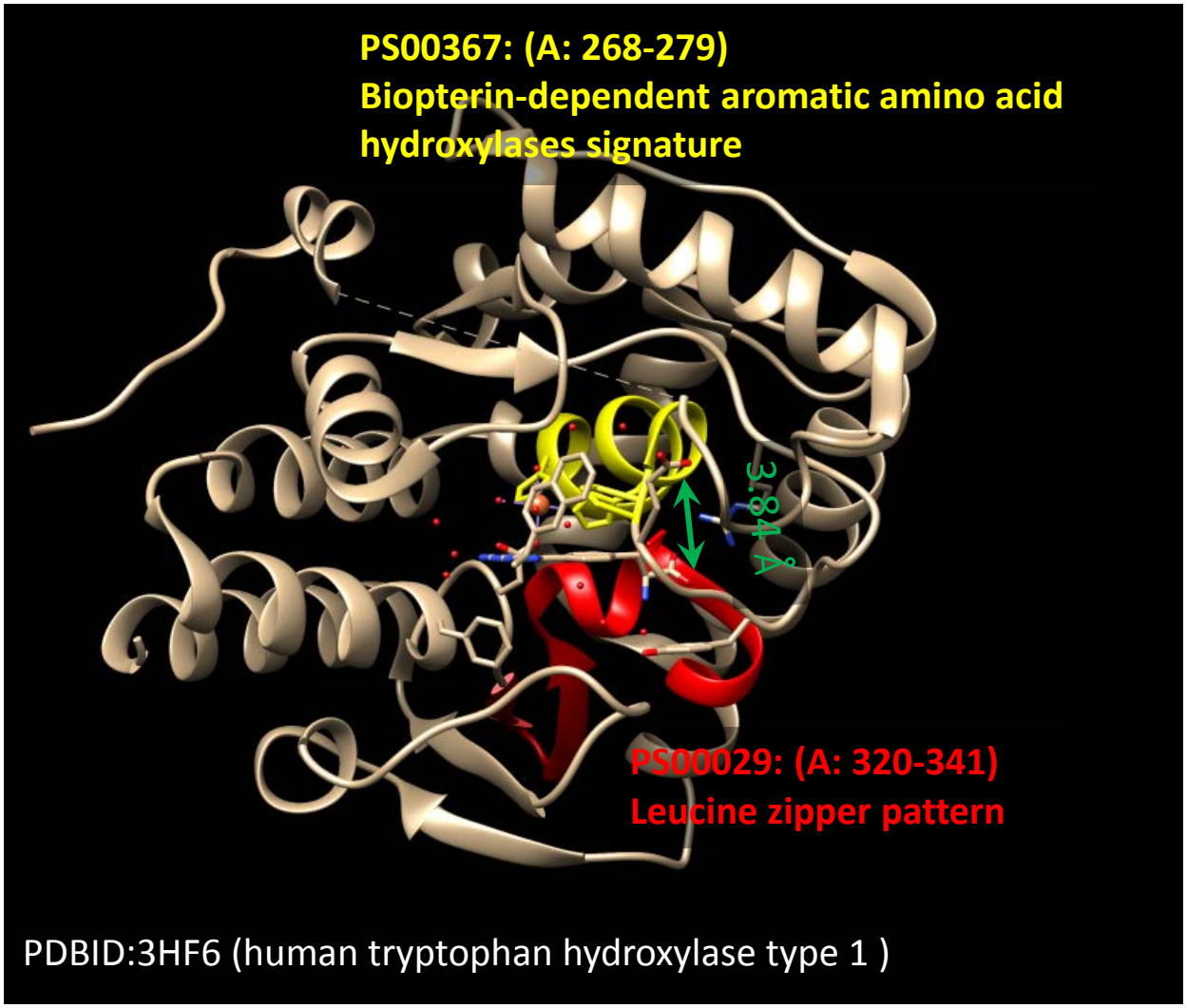
- 共起のEnrichmentにおける有意性が FDR< 0.05である。
- 含まれているモチーフ同士のCα最短距離がすべて3.5 - 6.0 Å内にある。

上の2つの条件に合うモチーフの共起のみにしぼった。

Motif combination	$N_{A,B}$	N_A	N_B	$T_{A,B}$	p-value	FDR	N(3.5 - 6.0 Å)
PS00621_PS50240	31	19265	978872	8.82E+10	0	0	31
PS00135_PS00286	**7**	**860346**	**3720**	**9.73E+10**	**0**	**0**	**7**
PS50883_PS50925	20	23706	31138	6.01E+10	0	0	20
PS51096_PS51480	4	60	4242	3.91E+10	0	0	4
PS00623_PS00626	5	103158	141064	9.63E+10	1.45E-08	4.01E-08	5
PS00135_PS51390	2	752375	1426	8.92E+10	2.87E-07	7.50E-07	2
PS00029_PS00367	**7**	**8426690**	**17903**	**9.67E+10**	**0.00022**	**0.000489**	**7**
PS00107_PS00221	5	3359325	51156	9.8E+10	0.009238	0.018139	5
PS00299_PS50002	2	300387	160106	8.62E+10	0.019142	0.036359	2
PS50240_PS51004	1	55998	73610	1.91E+10	0.020202	0.0383	1

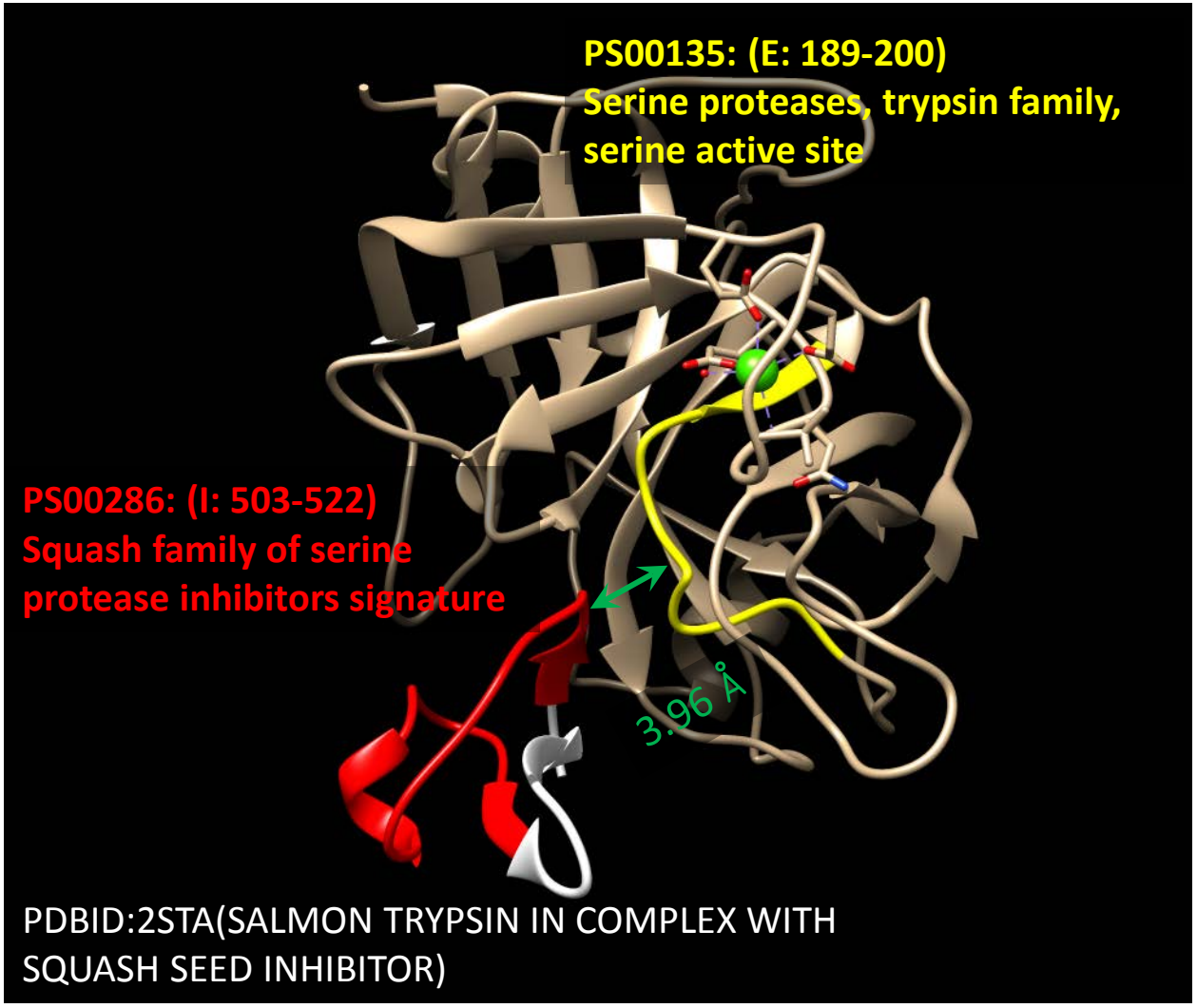
181 entries

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

まとめ

- 現状はスケジュールどおりに進行している。課題は残っているが、共起関係の検出を一通り行った。
- タンパク質の冗長化が現状ではうまく行えていない。Blastを行い冗長化を行う必要がある。
- モチーフ同士の距離の関係では、相互作用しているかのようなピークが見られた。
- 共起関係のenrichmentによっても絞ることができた。
- 距離の関係と、Enrichmentの両方を使って絞り込むと、重要そうな共起関係が検出できていた。

- 今後は、残っている課題を解決して、出力をデータベースにまとめ検索サイトを作成する。