ライフサイエンスデータベース統合推進事業
統合データ解析トライアル
研究開発課題
「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」

研究開発終了報告書

研究開発期間： 平成25年9月～平成26年1月

研究代表者　:藤井　聡
（九州工業大学大学院　情報工学研究院
生命情報工学研究系、助教）

§1 研究開発のねらい

現在、様々なゲノム・プロテオームに関する情報、例えば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（PPI）など非常に多くのデータベースが公開されている。しかし、それらの情報単独ではその価値を見出すことが難しいことがある。特に、多くのタンパク質多くの部位、つまりどこにでも存在するような情報としての価値の低いものに関しては単独では意味を見出すことは非常に難しい。しかし、そのような価値の低い情報であっても、ある情報は必ず別のある情報と一緒に存在しているような共起関係が分かれば、それらの情報の価値を高めることが可能となる。また、タンパク質をはじめとする生体分子は３次元の立体構造を取ることで機能している。ゆえに、タンパク質に対して付加されている情報もタンパク質の３次元構造として表されたときに有機的に意味のあるものとなる。例えば、タンパク質の配列として見ると離れた部分にあった情報が３次元で表した時に近傍になることもある。したがって、タンパク質中において情報同士が３次元的に近傍に存在するもしくはある一定の距離関係で存在するということが、情報が共起しているということを現すことになる。本研究の目的は、タンパク質にアノテーションされている様々な情報同士の関係性について、情報の集積だけでなく３次元的な距離関係も加味して共起関係を調べ、重要な関係性を検出する手法を開発することである。

タンパク質構造の立体構造は PDBj（http://pdbj.org/）から得ることができる。今回はすべてのタンパク質を対象として解析をしていくため PDB 構造全データを対象として取り扱う。また、今回共起関係を検出するターゲットは PROSITE（http://prosite.expasy.org/）から得ることのできるすべての機能ドメインと機能モチーフを使用する。PROSITE にはタンパク質のドメインやファミリーから小さな機能サイトまで含まれている。つまり元々情報としての価値の高いモチーフから、どこにでも存在するような価値の低いモチーフまで含まれている。そのため、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 を最初にターゲットとした。また、PROSITE は CATH や SCOP のような構造ベースではなく配列ベースで定義されているモチーフである。そのため、配列という一次元データとして意味を持つ情報を 3 次元の構造中に表すことでより有機的な結び付きを得ることができるのかという検証も想定している。今回共起関係については１対１の関係のみに絞って、まずは解析手法を確立することを目標とした。共起関係の検出方法としては、①タンパク質構造中で近傍に存在するもの、②タンパク質全体で高頻度に見られるもの、③その両方に当てはまるなどの条件に合致するものという 3 種類の方法での共起関係の検出を目指した。また発展的な展開として単一のタンパク質内だけから得られる共起関係だけでなく、2 つタンパク質が相互作用することによって現れる共起関係も調べることも考えている。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、統計的な優位性を含めてデータベースとして公開するまでを第一目標とする。データベースは MySQL で構築し、検索サイトは PHP5 により作成する。さらに、研究室で開発している PDBnet（http://dna00.bio.kyutech.ac.jp/pdbnet/）に統合する形にし、より利便性を高めることも考えている。その後、疾病関連遺伝子変異を OMIM や Human Validation DB から得てそれらとの共起関係への対応や、どのような情報に対しても対応できるように汎用化したプログラムにすることも視野に入れている。

§2 研究成果

- タンパク質の 3 次元構造データは PDBj より全 PDB 構造を取得した。共起を検出するターゲットのドメイン・モチーフの情報は PROSITE から得た。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た。タンパク質の冗長化について、PDB 中の残基番号と Uniprot のアミノ

酸番号を結びつけているEMBL-EBIのSIFTS（http://www.ebi.ac.uk/pdbe/docs/sifts/）を使い行った。

- 共起関係の検出は当初予定していた①タンパク質構造中で近傍に存在するもの、②タンパク質全体で高頻度に見られるもの、③その両方に当てはまるなどの条件に合致するものという3種類の方法を一通り行い、共起関係を検出できた。また、2つのタンパク質が相互作用することによって現れる共起関係についても応用できた。Tryptophan hydroxylaseに存在するLeucine zipper patternのように、結晶構造が出てなかった時代の論文に述べられた知見が発掘できていたりしたので、あながち間違いでは結果が得られているのかもしれない。更に検証をしていく必要があるだろう。
- 今回、PROSITEにはどこにでも現れるような情報としての価値の低いモチーフも含まれているが、そのような価値の低かったモチーフが、他のモチーフと共起することによって価値が高くなって現れてきているのも確認できた。
- 最終的にその得られた共起関係のリストはデータベースにして、PDBnet内に検索ツールとして公開している。

  PDBnet - Cooccurrence Search Tool -

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

  現状ではPROSITEにしか対応していないので汎用性は低い。将来的に統合データベースにある様々な情報について増やしていったり、今回これらを算出するために使ったプログラムを汎用化して様々なデータに適用できるようにしたりする必要があるだろう。

§3 研究開発計画および計画に対する達成状況

（1）達成状況

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することはできた。それ以外のP-famドメイン情報やOMIMやHuman Variation DBにある疾病関連遺伝子変異との共起関係への発展、またプログラム・ツールの汎用化までは対応することができなかった。

(2)ツールの将来性への展望

今回の研究開発の結果について、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、同じモチーフの共起関係を持つタンパク質には同じファミリーに属するタンパク質が集まっていたり、酵素の活性部位付近に共起関係が存在していたりと、結果に抽出されてくるべき結果が抽出できていた。今回の結果は、もっともらしい堅実な成果だと考えている。

今回の共起関係を検出する方法として、タンパク質の立体構造を使った3次元的な距離関係により共起関係を調べる試みをしたが、今回の手法ではその3次元的な情報を生かし切れた方法とは言い難い。今後、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もっと生かした結果を導き出せると考えている。

今回の共起関係を見るという、情報を結びつけるという非常に単純な発想であっても、情報の価値を高めることができることができた。今後情報を統合していくにあたって、このように情報同士を結び付けることが重要だと考えている。これまで世界中でライフサイエンスにおける様々な情報が集められてきた。しかし、情報があまりにも溢れかえっていて、利用者はそれらの情報の中でどれが重要なのかなど判断し難い状況であった。しかも情報それぞれがオリジナルな規格を持ち、それぞれ独立した情報としてしか扱えなかった。それが、近年の統合化プロジェクトをはじめとしたライフサイエンス情報の統合化がすすむことにより、様々な情報を俯瞰してみることができるようになってきた。いわば、完全にバラバラであった情報が、まとめられ部分的に重なりも見える和集合を見ている状況になっ

たと言えるだろう。そして、その次の段階として必要なことは、その中で特に重要な部分である様々な情報同士が関係し合う部分言わば積集合の部分を見出すことだと考えている。何をもって繋がりがあるという積集合とみなすかが重要なポイントとなってくるだろう。今回の私の開発研究では、タンパク質構造という３次元空間上での情報の集積によって関係性を調べようとした。しかし、タンパク質構造を使って結び付けて意味のある情報は、原子単位やタンパク質のアミノ酸配列単位でアノテーションされている情報、例えば今回扱ったドメイン・モチーフの情報や疾病関連 SNP などまでであろう。タンパク質単位、遺伝子単位でアノテーションされている GO や疾病関連遺伝子などの情報における関係性を調べるためには、タンパク質同士の関係性や遺伝子制御による関係性を現したインタラクトーム（プロテオーム）やトランスクリプトームから抽出する必要があると考えらえる。更にその上の階層にはメタボローム、フェノームも存在し EC 番号のようにこの階層でないと意味を見出しにくい情報も当然存在する。このように各階層において、様々な情報が存在している。まずは、それぞれの階層で情報の特徴であったり、情報同士の関係性であったりを抽出する方法を考案するのが重要ではないかと考えている。その次のステップとして、階層を縦断できる情報もあるので、それらを使って全体の関係性を調べていくと良いのではないかと考えている。

私の研究室ではこれまで分子構造の階層であるストラクチュロームと、タンパク質-タンパク質相互作用からのインタラクトームの階層の統合を行っており PDBnet という Web ツールとして公開している。ストラクチュロームにおける情報の集積及び関係性を調べるという観点から３次元構造を利用した今回の研究開発を計画した。今後はインタクトームにおいて情報の集積や情報同士の関係性を調べ、ストラクチュロームからインタラクトーム間を縦断して調べることができるようにすることも視野に入れている。更にはトランスクリプトームからメタボロームまでまとめて俯瞰できるようにできたらとも考えている。

§4 研究参加者

氏名	所属	役職	研究開発項目	参加時期
○藤井 聡	九州工業大学大学院 情報工学研究院生命情報工学研究系	助教		H25.10-H26.1

§5 成果発表等

(1)原著論文発表 （国内(和文)誌 0 件、国際(欧文)誌 0 件）

(2)その他の著作物(総説、書籍など)

なし

(3)国際学会発表及び主要な国内学会発表

① 招待講演 （国内会議 0 件、国際会議 0 件）

② 口頭発表 （国内会議 0 件、国際会議 0 件）

③ ポスター発表 （国内会議 1 件、国際会議 0 件）

1. 藤井 聡(九工大・情報工)、PDBnet: タンパク質相互作用ネットワークにおける構造情報と生化学実験情報の統合、第 36 回日本分子生物学会、神戸ポートアイランド、12/3-12/6

(4)知財出願

①国内出願 (0 件)

②海外出願 (0 件)

③その他の知的財産権

なし

(5)受賞・報道等

なし

§6 自己評価

研究開発のねらいとして考えていたことは、研究開発期間でほぼ達成することができたと考えている。元々4 か月という研究期間を考えてかなり現実的な申請をしていたので、大きな路線変更や想定外な事態もなく、堅実な結果を出すことができたと考えている。また、今回の自分が取った手法の問題点も分かったので、今後研究の展開で改善していきたい。今回この統合推進事業に参加することで、ライフサイエンスデータベースに関しての現状や方針、また最新の技術の一端にも触れることができ、非常に刺激になった。そのような今回得た知識を即座に今回の研究に反映することができればよかったが、研究期間の短さからそれが叶わなかった。それらも今後自身の研究へ反映していきたい。このように今後の課題・展望が明確になったというのが一番大きな成果だったかもしれない。

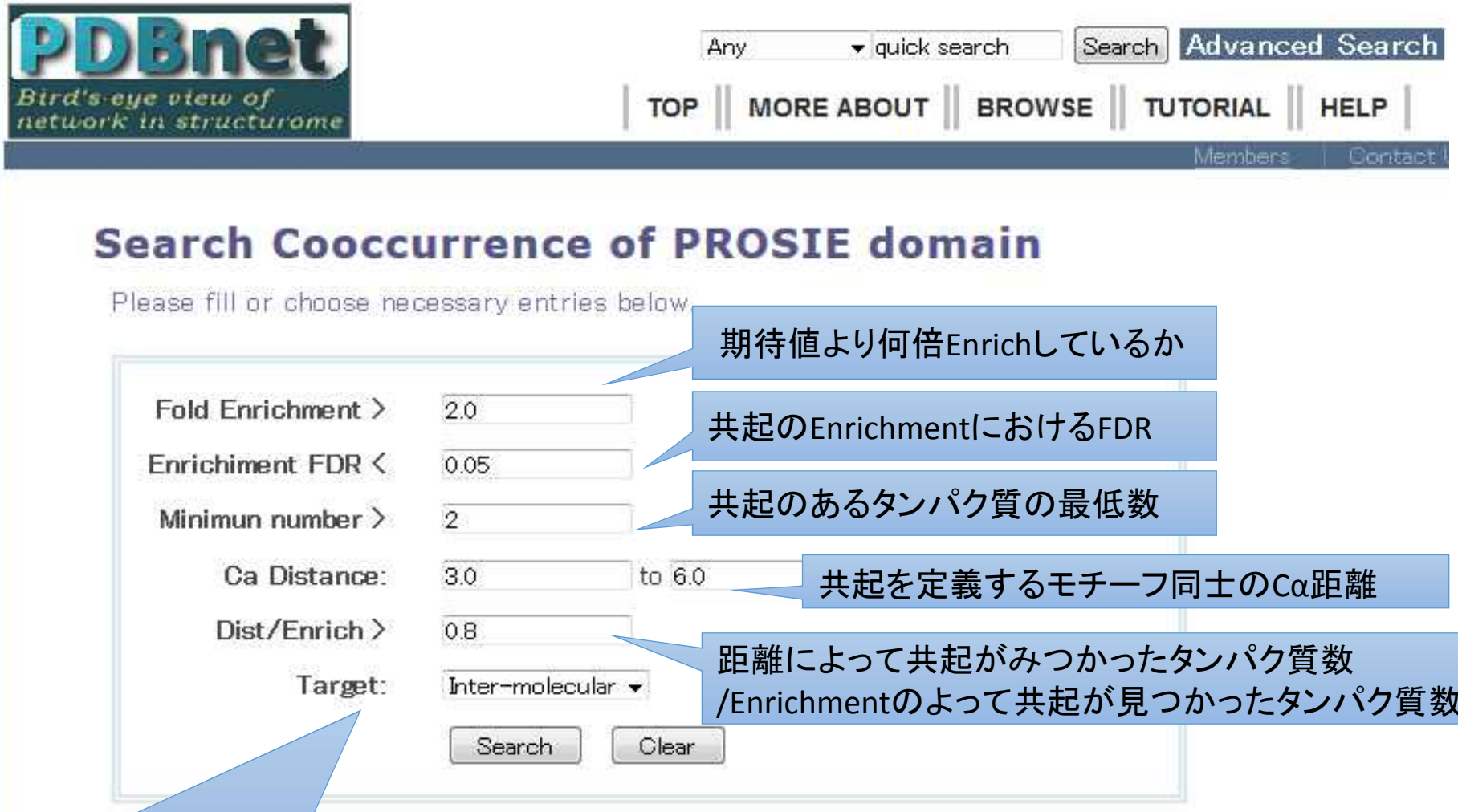
以上

PDBnet
- Co-occurrence Srearch Tool -

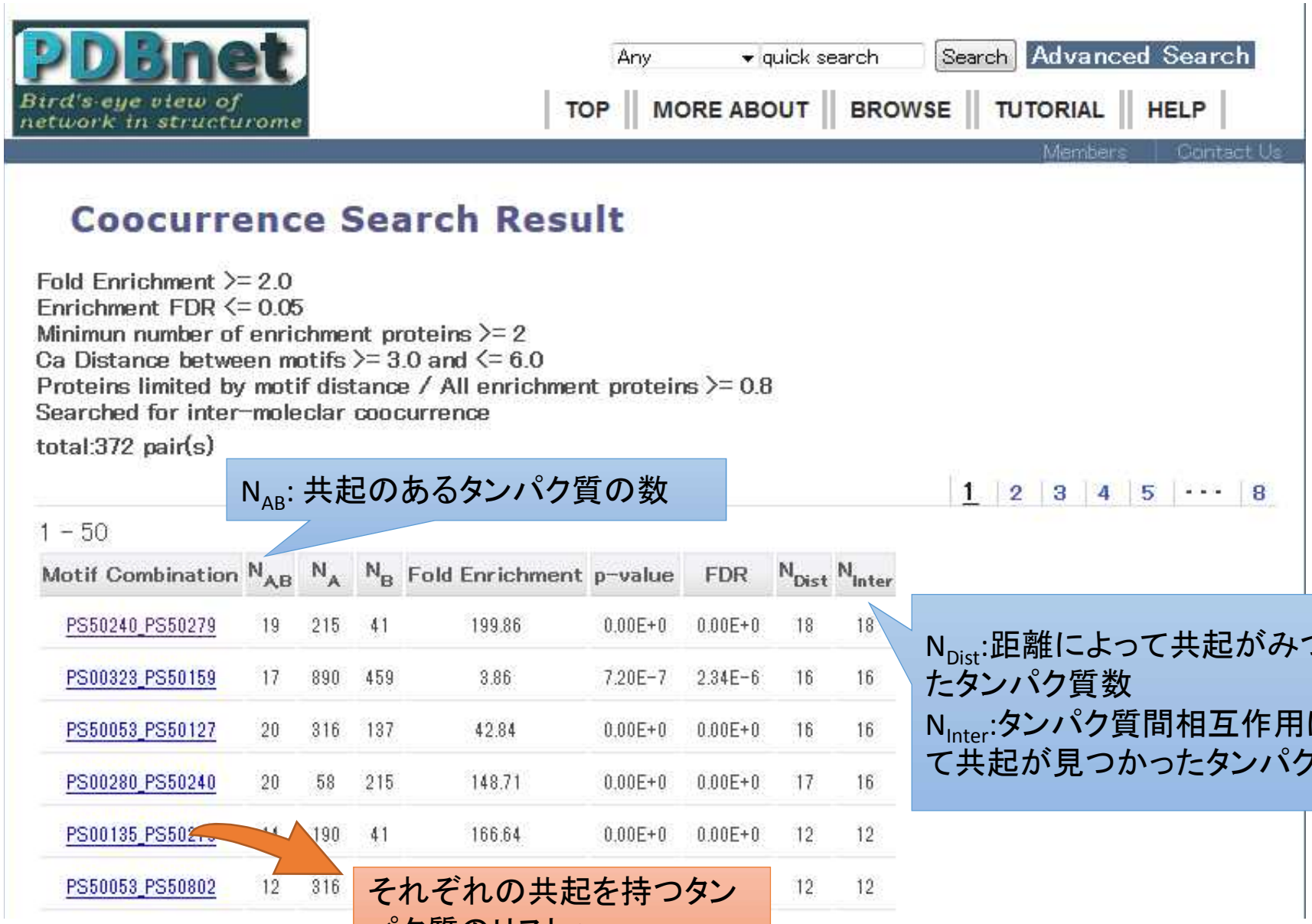
Co-occurrence Searchの機能

- 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。

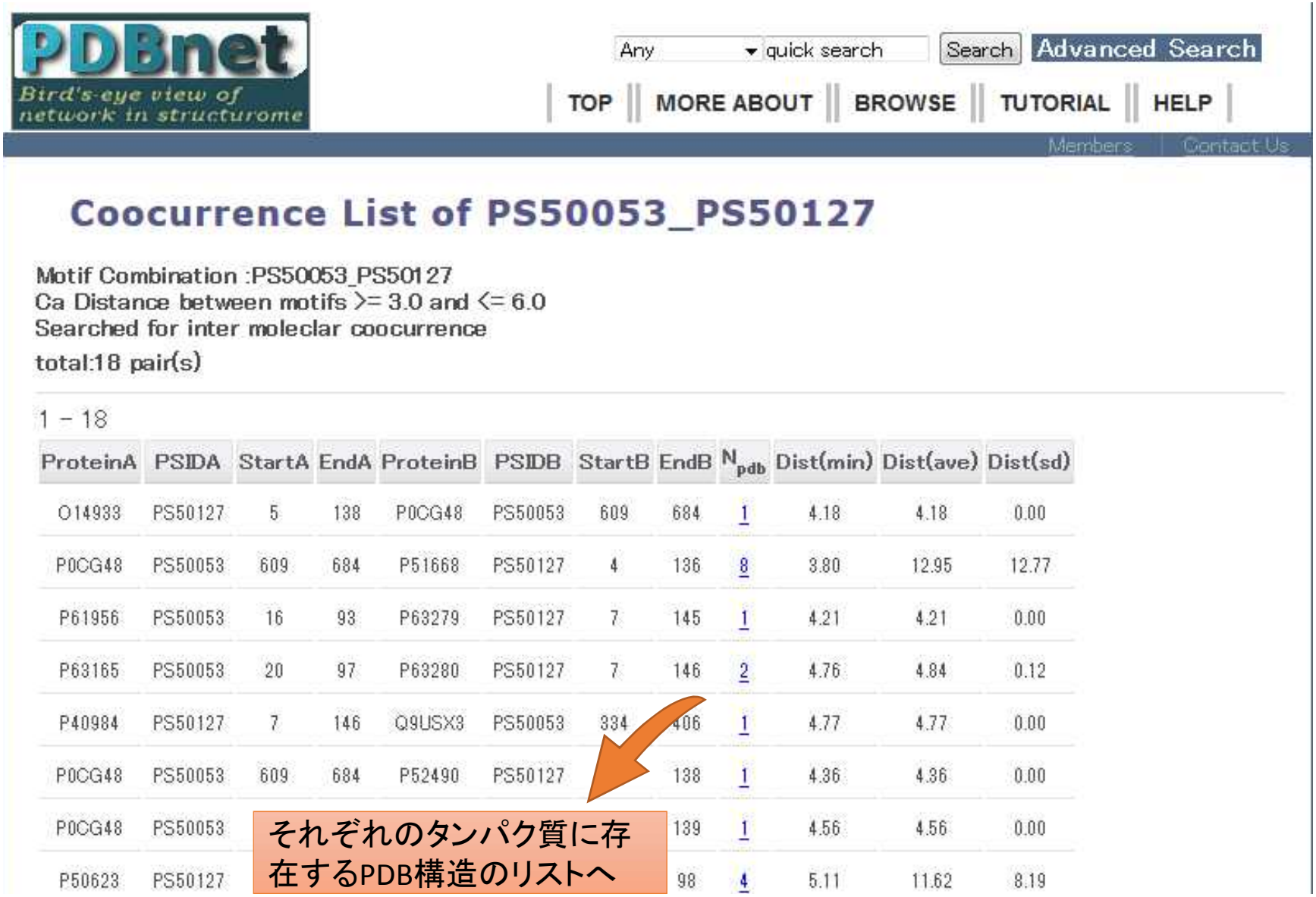
共起関係のデータベースにアクセスする検索ページ



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



Coocurrence List of PS50053_PS50127

Motif Combination :PS50053_PS50127
Ca Distance between motifs >= 3.0 and <= 6.0
Searched for inter moleclar coocurrence
total:18 pair(s)

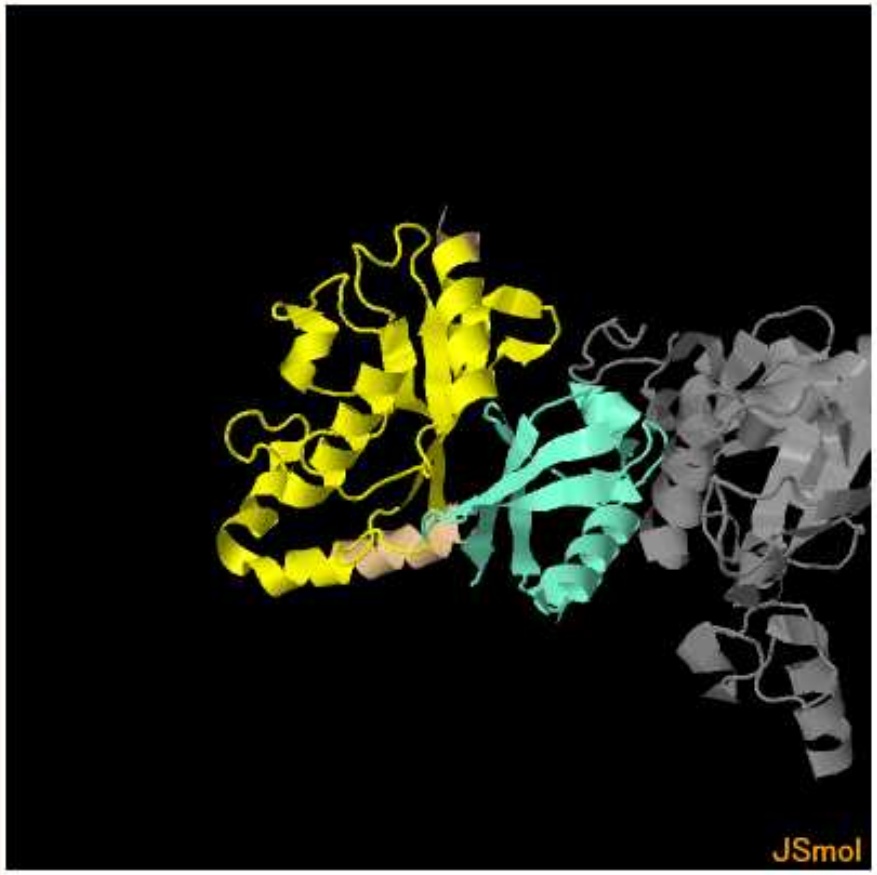
1 - 18

ProteinA	PSIDA	StartA	EndA	ProteinB	PSIDB	StartB	EndB	N_{pdb}	Dist(min)	Dist(ave)	Dist(sd)
O14933	PS50127	5	138	P0CG48	PS50053	609	684	1	4.18	4.18	0.00
P0CG48	PS50053	609	684	P51668	PS50127	4	136	8	3.80	12.95	12.77
P61956	PS50053	16	93	P63279	PS50127	7	145	1	4.21	4.21	0.00
P63165	PS50053	20	97	P63280	PS50127	7	146	2	4.76	4.84	0.12
P40984	PS50127	7	146	Q9USX3	PS50053	334	406	1	4.77	4.77	0.00
P0CG48	PS50053	609	684	P52490	PS50127		138	1	4.36	4.36	0.00
P0CG48	PS50053						139	1	4.56	4.56	0.00
P50623	PS50127						98	4	5.11	11.62	8.19

それぞれのタンパク質に存在するPDB構造のリストへ

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

Jmolによる共起の状態の確認



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。