

平成27年度博士学位論文

ゲノム・タンパク質情報解析技術の適用による
貯蔵mRNAに結合するmRNA前駆体3'末端開裂複合体の
立体構造と機能部位の推定

お茶の水女子大学大学院
人間文化創成科学研究科 ライフサイエンス専攻
青砥 早希

平成28年 3月

目次

論文要旨	3
序 章	
脊椎動物の初期発生における遺伝子発現制御	8
本研究の目的と構成	13
本研究で使用したゲノム・タンパク質情報	14
第一章 CPSFサブユニットの立体構造予測	
背景	18
方法	22
1. CPSFサブユニットのアミノ酸配列の入手	22
2. CPSFサブユニットの天然変性領域の検索	22
4. コンパラティブモデリングによるサブユニットの立体構造予測	25
5. CPSF100における亜鉛イオン配位部位探索	26
結果	28
1. CPSF構成因子の天然変性部位	28
2. CPSF構成因子のテンプレート配列の検索	28
3. 各サブユニットの予測立体構造	32
4. CPSF100における活性部位探索	34
考察	37
第二章 複合体立体構造予測ツールの開発	
背景	41
方法	46
1. ヘテロ三量体の既知立体構造の取得	46
2. 既存二量体予測ツールの検討	46
3. デコイ構造のクラスタリング	47
4. クラスタリングスコアを用いた既存の複合体予測手法の改良	49
5. 候補構造数の決定	51
6. 結合パターンによるクラスタリングを取り入れた構造予測手法の精度	52
結果および考察	53
1. 既存二量体予測ツールの検討	53
2. デコイ構造のクラスタリング	56
3. 式(1)によるクラスタースコア(Gscore)の算出	58
4. 候補構造数の決定	62
5. 式(1)を用いた構造予測手法の精度	62
6. 小括	65

第三章	複合体におけるサブユニット界面の特徴抽出	
背景		68
方法		70
	1. パラログ間で共有しない結合残基の特徴抽出	70
	2. CPSFサブユニットへの適用	75
結果		77
	1. サブユニット構成を変化させた複合体およびそのパラログセット	77
	2. パラログセットから得られた結合部位の分類	77
	3. CPSFサブユニットにおけるパラログ間で異なる結合残基の予測	95
考察		97
第四章	CPSFサブユニットの複合体予測	
背景		101
方法		105
	1. 複合体立体構造予測に用いたサブユニット構造	105
	2. 分子系統解析によるCPSFサブユニットの結合ペア予測	105
	3. CPSF複合体立体構造の予測	107
	4. 候補構造の選出	108
	5. 候補構造の評価	108
	6. CPSF複合体と結合する分子の結合部位予測	110
	7. mRNAの結合部位予測	113
	8. CPSF160における一塩基変異(SNV)検索	114
	9. CPSF100の天然変性領域の機能予測	115
結果		116
	1. CPSFサブユニットの類似配列検索	116
	2. CPSFサブユニットのアラインメントおよび分子系統樹の作成	118
	3. CPSFサブユニットの分子系統解析	120
	4. 候補構造の選出	124
	5. 候補構造の評価	124
	6. CPSF複合体とCPEBおよびGld2の結合部位予測	132
	7. CPSF160上のSNV検索	141
	8. CPSF100の天然変性領域の機能予測	145
考察		147
終章		
	初期発生機構研究における本研究の位置づけ	160
	本研究によって得られた成果	161
	初期発生機構研究における本研究の展望	169
	複合体機能解析分野における本研究の位置づけと展望	169
用語説明		171
参考文献		179

論文要旨

ゲノム・タンパク質情報解析技術の適用による 貯蔵mRNAに結合するmRNA前駆体3'末端開裂複合体の立体構造と機能部位の推定

青砥 早希

真核生物の初期発生では、個体の生涯において、この時期にしか発現しない遺伝子が多く存在する。これらの遺伝子の発現場所と発現時期は、いくつかの機構によって精密に調節されていることが明らかになっている。特に、DNAからmRNAへの転写過程の制御機構が詳細に調べられており、近年ではDNAのメチル化など、エピゲノミックな調節機能に着目した研究が展開されてきた。一方、これらの遺伝子がmRNAからタンパク質へ翻訳される過程でも、多くの制御がはたらいっていることが近年明らかになってきている。アフリカツメガエルの卵細胞では、減数分裂を担うタンパク質の合成は、一次卵母細胞形成時に転写され蓄積された母系mRNAの翻訳抑制により制御されている。母系mRNAの翻訳抑制については、様々な生物で確認されており、少なくとも動物界でこの機構は保存されていることが明らかになっている。

DNAから転写されたmRNAは、キャッピング、ポリA鎖伸長、スプライシングなどのプロセッシングを受けた後に翻訳される。通常、ポリA鎖の修飾を受けるmRNAは、核内でポリA鎖が伸長し、細胞質に輸送されるとすぐにリボソームによって翻訳される。しかし母系mRNAは、核内でポリA鎖が十分に伸張されることなく細胞質に輸送され、細胞質で長期間蓄積される。翻訳活性時には、細胞質でポリA鎖の伸長が行われる。細胞質に局在する母系mRNAは、ポリAポリメラーゼ(Gld2)やmRNA3'末端切断およびポリアデニル化特異因子(CPSF)、細胞質ポリアデニレーション因子結合タンパク質(CPEB)などのタンパク質群と結合し、mRNA-タンパク質複合体(mRNPs)を形成する。mRNPsの主なはたらきは、母系mRNAの翻訳抑制(貯蔵型mRNPs)と、翻訳開始時の母系mRNAのプロセッシングの完了(翻訳活性型mRNPs)であると考えられている。母系mRNPsの構成因子は、前述したタンパク質以外にも数多く同定されており、これらの分子のはたらきは、その詳細が明らかになってきた。しかし、母系mRNPsがどのようにその機能を発揮するのかを理解するためには、母系mRNAの周りでmRNPs構成因子がどのように相互作用するかを原子分解能で明らかにする必要がある。そのためには、母系mRNPsの複合体構造を決定することが必須だが、実験的な手法として一般的であるNMR測定やX線結晶構造解析を行うには、mRNPsは分子量があまりにも大きく複雑な分子であるために、データ測定と処理が困難である。そこで本研究では、計算科学的な手法を用いて、母系mRNPsの構成因子の1つであるCPSFの三次元構造を予測し、その構造を基に母系mRNAの翻訳制御に関与するCPSF、Gld2、CPEBおよび母系mRNAの相互作用について考察した。

母系mRNPsの構成因子の1つであるCPSFは、それ自体もタンパク質の複合体であり、CPSF160、CPSF100、CPSF73、CPSF30の4つのサブユニットを有する。そこで、CPSF複合体立体構造を予測するにあたって、CPSFの各サブユニットの立体構造を予測し、サブユニットを組み合わせた。さらに得られた複合体構造をもとに、RNA結合残基予測および、CPSF

複合体と直接相互作用することがわかっているGld2とCPEBの結合残基を予測した。

CPSF73とCPSF100は、共通祖先配列由来のタンパク質であった。共通祖先由来の2つのタンパク質が相互作用する場合、それらのタンパク質はそのアミノ酸組成や立体構造が類似であるため、その相互作用面は相同なドメインで形成される傾向にある。CPSF73とCPSF100は、この2つのタンパク質が共通してもつドメインで相互作用しており、その予測二量体構造には空間対象性が見られた。予測構造全体はU字に似た構造をとっており、CPSF100が、mRNAの認識を行うCPSF1つ60とエンドヌクレアーゼ活性をもつCPSF73をつなぎとめていた。またCPSF160のCPSF100との結合残基は、多くがCPSF160パラログにはないアミノ酸残基で構成されていた。CPSF160パラログはCPSF73パラログやCPSF100パラログとは結合していないため、CPSF160は、アミノ酸残基の挿入によってCPSF160固有の結合部位を獲得したと考えられる。予測されたCPSF複合体構造を用いて、CPEB、Gld2およびmRNAの結合部位を予測した。CPSF160は3つのドメインから構成され、CPEBとGld2はCPSF160の同じドメインと相互作用することが予測された。CPSF160上の両予測相互作用部位が近接していたことは、CPEBとGld2とがCPSFとの相互作用時にお互いも相互作用するという実験結果と符合する。予測RNA結合部位の1つは、CPSF-CPEBの予測結合部位に近接した位置であった。このRNA予測結合部位は、CPSF73の活性部位と約80Å離れていた。このRNA結合部位とCPSF73の活性部位間の距離は、mRNA上のCPSF160認識部位とCPSF73のmRNA開裂部位間の距離と合致した。Gld2はCPSF160とCPSF73に挟まれるように結合し、mRNAとの結合部位および活性部位を外側に向けて、CPSF複合体に埋まり込むように結合していると予想された。これらの予測結果を総合すると、貯蔵mRNPs中ではGld2はmRNAのポリA鎖付与部位より4~8塩基上流の位置に結合し、CPSF73とCPSF160、CPEBの少なくとも3分子と直接相互作用することでmRNAのポリA鎖付与部位から離れた場所に存在していると予測された。これらの相互作用は、Gld2によるmRNAのポリA鎖伸長を阻害する構造であり、結果的に母系mRNAの翻訳抑制機構の維持に寄与していると考えられる。母系mRNPsの翻訳抑制解除の引き金と考えられているCPEBのリン酸化と、母系mRNAの翻訳活性に必須とされるポリA鎖の伸長の2つの現象をつなぐ分子機構は未知である。本研究で行なった貯蔵mRNPsの相互作用の予測結果から、翻訳抑制機構のメカニズムはCPSFの構造にGld2が埋まるように結合し、さらにCPEBがストッパーとなるようにGld2と結合することで、Gld2がmRNAのポリA鎖付与部位に結合できないような立体構造をとっていることだと予想される。このことはCPEBのリン酸化によりGld2がCPEB、CPSFとの結合を維持できなくなると、Gld2はCPSF複合体から離れmRNAのポリA鎖付与部位を認識できるようになり、翻訳抑制が解除されることを意味する。*Xenopus leavis*の未成熟卵のGld2を抗体により阻害すると、核膜崩壊が促進されること、また核膜崩壊に必須な母系mRNAであるMos mRNAのポリA鎖伸長が阻害されることが示されている。この2つの現象は相反しているようであるが、今回の予測構造に基づけば、Gld2抗体がGld2のCPEB、CPSFとの相互作用による翻訳抑制維持を阻害したことでMos mRNPsの立体構造が変化し、その翻訳が行われたと解釈することができる。

本研究により初期発生において複雑な遺伝子発現を担う母系mRNAの翻訳抑制維持を実現するメカニズムの新たなモデルを示唆することができた。本研究で示した三次元的なサブユニットの機能は、個々の分子機能の解明の進んだ母系mRNPsの翻訳抑制機構研究において、機構の詳細を解明するための新たな指針となる。

外国語要旨

Structural bioinformatics analyses for quaternary structure and functional sites prediction of mRNA cleavage and polyadenylation specificity factor (CPSF)

Saki Aoto

During the oocyte maturation and early embryonic development in higher eukaryotes, gene expressions are controlled not only by transcription, but also by translation. In immature oocytes, transcription activity of gene is low, but transcripts stored in the cytosol are utilized for high protein expression. The stored mRNAs named maternal mRNAs are transcribed before meiosis and wait for the time that they should be translated in the oocyte cytoplasm. In the oocyte maturation process, new proteins are translated solely from the maternal mRNAs. This phenomenon is also observed in neurons.

Messenger RNAs generally undergo a number of maturation processes in nucleus, such as splicing, 5'-capping, and poly(A)-tail elongation. The matured mRNAs are then translated immediately after they are transported to cytoplasm. Maternal mRNAs are, however, transported to cytoplasm only with a short poly(A) tail and elongation of the poly(A) tail is suppressed until the timing of translation. The maternal mRNAs are known to be associated with several different proteins and form mRNPs. The complex molecule protects themselves from degradation and regulates the timing of poly(A)-tail elongation.

All maternal mRNAs have cytoplasmic polyadenylation element (CPE) and a highly conserved AAUAAA hexanucleotide sequence (HEX) in their 3'-UTR as the targets for proteins forming the mRNP. CPE is the binding site of CPE-binding protein (CPEB), one of the core components of the maternal mRNP that recruits poly(A) deadenylase and is involved in translation repression. CPEB also binds cleavage and polyadenylation specificity factor complex (CPSF). CPSF binds to HEX and plays a key role in translation activation. CPSF cleaves immature mRNA 3'-UTR prior to polyadenylation by Gld2, one of the poly(A) polymerases (PAPs). Phosphorylation of CPEB is considered to trigger a structural change in the maternal mRNP, which in turn changes the level of translation in maternal mRNAs. The mechanisms of translation regulation by maternal mRNPs have been studied focusing on the function of each component. However, integration of these functions for the understanding of the overall mechanisms of translation repression in maternal mRNPs has not been tried. For the integration, three-dimensional structure of maternal mRNPs complex is required. An effort has been made for determining the structure of maternal mRNPs, but it remains to be elucidated. To understand the mechanisms of translational regulation of the maternal mRNPs, I predicted a three-dimensional structure of CPSF, the core component of maternal mRNPs, that consists of CPSF160, CPSF100,

CPSF73 and CPSF30. Based on the predicted structure, the interactions of core components to maternal mRNPs and mRNA were predicted as well.

The 3D structures of the CPSF subunits were built by comparative modeling method. To assemble CPSF subunits, the in-house protein-interface prediction tool was improved and utilized. For predicting an assembly of three or more complex, the biggest problem was to figure out direct interactions (partners) among the subunits. To address this problem in CPSF, information from the molecular phylogenetic analyses of each CPSF subunit were used. IntS9, a close homolog of CPSF100, and IntS11, a close homolog of CPSF73 are known to make a direct interaction, hence CPSF100 and CPSF73 likely interact with each other directly. From this observation, a docking calculation between CPSF100 and CPSF73 was performed. The 3D structure of CPSF160-CPSF100-CPSF73 complex was then built by docking CPSF160 and CPSF100-CPSF73 heterodimer. CPSF160-CPSF100-CPSF73 complex turned out to have a U-shape and CPSF160 bound CPSF73 through CPSF100. The molecular function of CPSF100 has been unclear, but the association with CPSF100 was known to be required for efficient endonuclease activity of CPSF73. Based on the model, CPSF100 seems important for bringing CPSF73 close to CPSF160 that binds mRNAs. The prediction of the binding site of CPEB and Gld2 on CPSF resulted in finding both binding sites on the same domain of CPSF160, and this enabled direct interactions with each other. This modeling structure is consistent with the previous experiment showing that CPEB on CPSF bound to Gld2. Gld2 also binds to CPSF73, bridging the opening of the U-shape with its catalytic site facing away from CPSF. Several RNA-binding sites were predicted on CPSF including a one on CPSF160 located next to the binding site of CPEB. The RNA-binding site was about 80 Å apart from the catalytic site of CPSF73, and this distance matched the distance between HEX and the site cleaved by CPSF73. As a result, the catalytic site of Gld2 was located far from poly(A)-binding site on mRNA and this configuration was seemingly sustained by the interaction of CPEB. The binding pose of Gld2 prevents itself from cleaving the maternal mRNA, hence the translation of the maternal mRNAs should be suppressed. At the initiation of the translation of maternal mRNA, CPEB is reported to be phosphorylated. Once CPEB is phosphorylated, it is speculated that CPEB could not sustain all the interactions and Gld2 should be released from both CPEB and CPSF, then Gld2 recognized the 3'-terminus of the mRNAs and Gld2 should initiate polyadenylation.

In this study, the 3D structure of the core components of maternal mRNPs in translational repression state was built and predicted the bindings sites of Gld2 and CPEB. The structure of CPSF helped to build a hypothesis for the mechanism of mRNPs translational repression, and could be used to obtain the model that the binding pose of Gld2 is altered after phosphorylation of CPEB. The predicted CPSF complex structure will pave the way for the understanding of the mechanism of the translational regulation on maternal mRNPs.

序 章

脊椎動物の初期発生における遺伝子発現制御

1つの細胞である卵は、受精が成立すると十兆個以上の細胞に分裂し(Ben-Tabou de-Leon et al. 2007)、それぞれの細胞が数十もしくは数百の機能を獲得して個体を形成する。この初期発生の過程では、個体の生涯において、この時期にしか発現しない遺伝子発現が多くみられ、これらの遺伝子の発現場所と発現時期はいくつかの機構によって精密に調節されている。初期発生におけるタンパク質合成は主に転写調節で行われ、その転写調節は転写因子や転写調節因子、シグナリング分子の網目のように複雑な遺伝子発現制御機構により行われていることが、ウニやショウジョウバエの胚を用いた研究により詳細に分かっている(Ettensohn 2013, Li and Davidson 2009)。

一方、未受精卵ではDNAからmRNAへの転写が抑制されていることが分かっている(Han et al. 2003)。この時、卵はタンパク質を発現してないのではなく、減数分裂開始以前に転写し蓄積しておいたmRNAを翻訳して減数分裂や初期発生をすすめる(Dworkin et al. 1985, Lin 2010)。このmRNAは母系mRNA(maternal mRNA)と呼ばれ、母系mRNAによる遺伝子発現は転写反応を経ないため、卵細胞質内にタンパク質を速やかに供給することができる。この現象は受精直後に見られる卵内の新規タンパク質合成量が数十倍に跳ね上がる現象の原因と考えられている(Lin 2010)。

母系mRNAによる遺伝子発現

一般にmRNAは核内でスプライシングやポリA鎖の伸長などのプロセッシングを完了したのちに核外に移行する(Proudfoot et al. 2002, Piñol-Roma and Dreyfuss 1992)。しかし、母系mRNAはポリA鎖が短い状態で細胞質で貯蔵されていることが確認されている。また母系mRNAの転写時期だと考えられている減数分裂前期の初期の卵では、ポリA鎖が伸長した

mRNAが観察されている(Huarte et al. 1987, Huarte et al. 1992)。このことから母系mRNAのポリA鎖は転写時には伸長されており、細胞質に移行された後、脱アデニル化され、翻訳活性化の際に改めて細胞質内でポリA鎖伸長などのプロセッシングが行われる(Mendez et al. 2000a)。この現象はcytoplasmic polyadenylationと呼ばれる(Wilt 1973, Richter et al. 2008)。cytoplasmic polyadenylationは生殖細胞の胚発生の初期および、ニューロンでの炎症反応時にしか観察されていない(Weill et al. 2012)。

細胞質に移行した母系mRNAはmaternal messenger ribonucleoproteinと呼ばれるタンパク質との複合体を形成している。maternal messenger ribonucleoproteinの機能は、母系mRNAのリボヌクレアーゼからの保護や翻訳抑制、また翻訳開始時の母系mRNAのプロセッシングの完了など、多岐にわたる(Dominski et al. 2013, Kim and Richter 2006)。母系mRNAとmaternal messenger ribonucleoproteinの複合体は一般的には総称して母系mRNPsと呼ばれるが、本研究では翻訳抑制をうけ、細胞質で貯蔵されている時の母系mRNPsを貯蔵型mRNPs、プロセッシングを完了し成熟した母系mRNAと内包している母系mRNPsを翻訳活性型mRNPsと定義し、またmaternal messenger ribonucleoproteinをmRNPsと区別するため、mRNPsサブユニットと記載することにした。

貯蔵型mRNPsを構成する母系mRNAは、3'末端側配列に以下に、ポリA鎖伸長を受けるmRNAに特異的なHex配列、母系mRNAに特徴的なCPE配列などの配列をもち(詳しくは巻末の用語説明を参照)、これらの配列の位置関係は、母系mRNAがコードする遺伝子の翻訳時期と密接に関わっていることが分かっている(Belloc et al. 2008b, Piqué et al. 2008)。

母系mRNAの発現時期は、未成熟卵として体内で細胞周期を停止していた卵が成熟卵となるため、細胞周期を開始してから受精後に細胞周期が数サイクル廻るまでの短い期間と考えられている(Nothias et al. 1995)。しかし母系mRNAがコードしているタンパク質の中に

は発現時期がもっと短いものも存在する。*Xenopus leavis*では母系mRNAの伸長の行われる時期によって母系mRNPsの構成タンパク質が異なっていることが分かっており(Radford et al. 1979)、少なくとも3つのタイプの母系mRNPsがあると考えられている(Belloc et al. 2008)。3タイプの母系mRNPsが主に翻訳を行う時期は、第一減数分裂前期、第一減数分裂中期、および分裂間期から第二減数分裂中期の間であり、翻訳時期が同じmRNA配列にはCPE配列の個数や位置に共通性があることが示唆されている(Belloc et al. 2008b, Piqué et al. 2008)。母系mRNA上の配列を変化させることで、この配列を認識するmRNPsサブユニットの構成が変化し、結果として母系mRNAがコードする遺伝子の翻訳時期が制御されていると考えられている(Piqué et al. 2008)。

貯蔵mRNPsを構成するmRNPsサブユニット

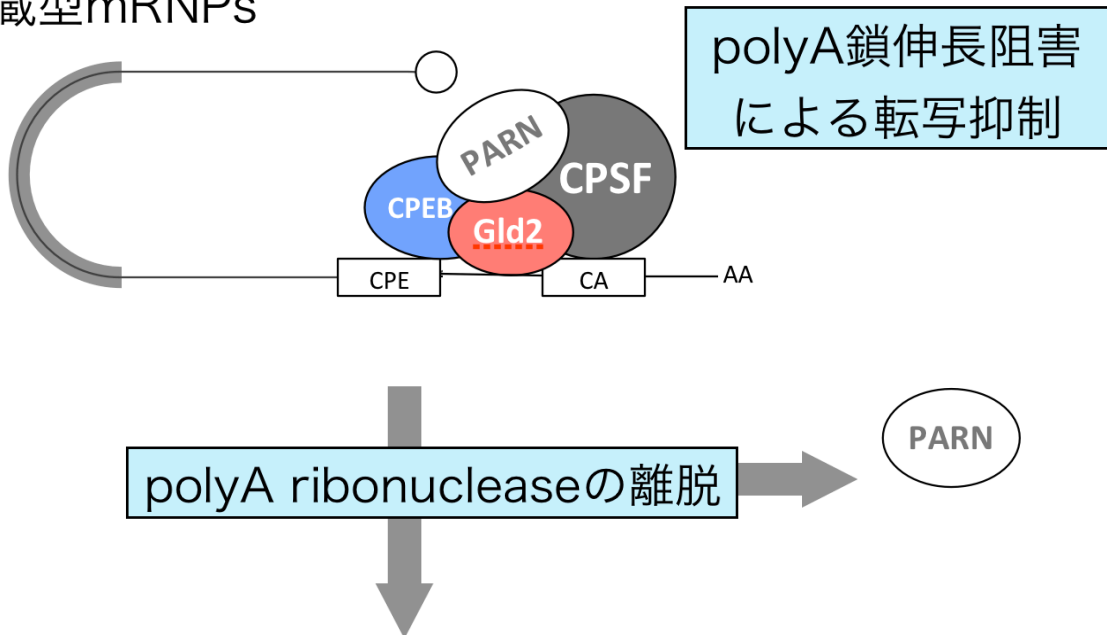
貯蔵mRNPsを構成する分子は、母系mRNPsの翻訳制御に必須な機能を有し、基本的な構造を構成している分子と、ある翻訳時期の母系mRNAとしか貯蔵mRNPsを構成しない分子とが存在する(Mendez et al. 2002)。前者は翻訳抑制に直接関与する分子であり、貯蔵mRNPsの状態を変化させることで母系mRNPsを翻訳可能な状態に変化させる分子である(Kronja and Orr-Weaver 2011)。後者の分子の主なはたらきは翻訳時期の調節であり、例としてMaskinやpumilioなどのタンパク質分子が挙げられる(Stebbins-Boaz et al. 1999, Nakahata et al. 2003)。これらの分子のバリエーションによって、貯蔵mRNPは減数分裂期および胚発生の初期の細胞周期調節や胚の各器官の構築に必須な分子を、卵成熟から初期胚発生の間の遺伝子発現に最適な期間まで翻訳抑制している(Mendez et al. 2002)。これらすべての母系mRNPs構成分子を含めると、少なくとも27種の分子が母系mRNAと直接または間接的に相互作用することが報告されている(Radford et al. 2008)。

現在わかっている、すべての翻訳時期に共通して散見されるタンパク質は、細胞質で貯蔵されるmRNAに特異的に結合し、翻訳抑制にはたらくとされるCPEB、mRNA前駆体3'末端開裂複合体として知られ、CPEBやGld2を母系mRNAにリクルートするはたらきをもつCPSF、細胞質特異的にはたらくポリAポリメラーゼであるGld2、貯蔵mRNPsでは、Gld2と拮抗することでポリA鎖伸長を抑制し、結果として翻訳抑制を担うポリAエンドヌクレアーゼであるPARN、翻訳活性化後、伸長した母系mRNAのポリA鎖を保護するePABの5つのタンパク質または複合体である(詳しくは巻末の用語説明を参照)。

翻訳抑制が行われている貯蔵mRNPsには、CPEB、CPSF、Gld2、PARNが存在する(図序-2. a)。ホルモンなどによる翻訳活性化刺激が伝わると、CPEBのリン酸化やPARNの離脱が起こり、母系mRNAのポリA鎖伸長が起こり、翻訳が行われる(図序-2. b)。このようにCPEB、Gld2は、翻訳活性化の際に大きく複合体構造、またはタンパク質表面の化学的性質を変化させると考えられる。このとき、貯蔵mRNPsのような複合体構造では、その状態変化によって各構成因子に何らかの相互作用の変化がおこる。貯蔵mRNPsにおいて、CPEB、およびGld2は、mRNA以外にCPSFとの結合が示唆されており、また、CPSFは、CPSF160、CPSF100、CPSF73、およびCPSF30と呼ばれるサブユニットをもつ複合体であることから(詳しくは巻末の用語説明を参照)、CPSFの周囲に大きな相互作用の変化が起きると考えられた。そこで、本研究では、CPSF複合体を、貯蔵mRNPsから翻訳活性型mRNPsへ状態転移する際の相互作用変化の中心に位置する分子であるとした。

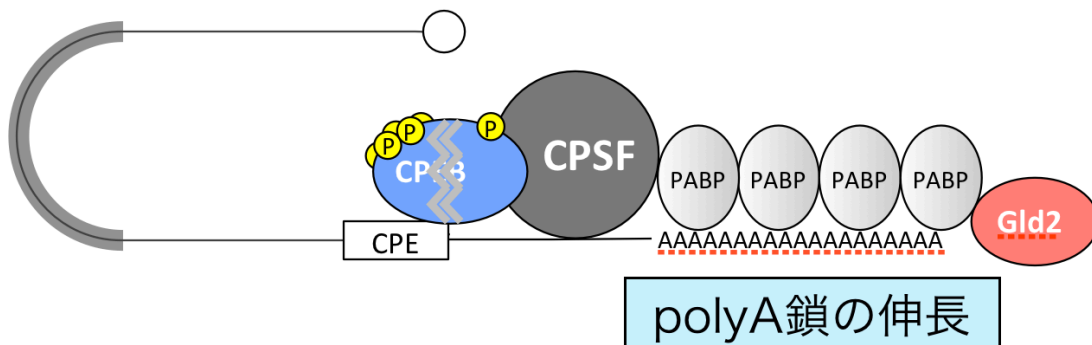
a

貯蔵型mRNPs



b

翻訳活性型mRNPs



図序-1. 貯蔵型mRNPsから翻訳活性型mRNPsへの移行

翻訳抑制が行われている貯蔵型mRNPsでは、翻訳抑制因子であるCPEBとポリA鎖付与に先立ってmRNAの3'末端を壊裂するとされるCPSFが母系mRNAに結合している。CPSFはpolyA polymeraseであるGld2をリクルートするが、貯蔵型mRNPs中では、polyA ribonucleaseであるPARNがGld2が拮抗することでpolyA鎖の伸長を阻害し、翻訳抑制を行っていると考えられている(a)。ホルモンなどの刺激により、CPEBのリン酸化、PARNの離脱が行われると、polyA鎖の伸長が行われ、翻訳活性型mRNPsとなる(b)。

本研究の目的と構成

分子生物学的実験によって母系mRNAに結合する分子や母系mRNPsを構成するサブユニットの正体は明らかになってきているが、それらのサブユニットがどのように相互作用しているかは明らかになっていない。mRNPsのような比較的大きな複合体構造の相互作用を明らかにすることは、各サブユニットの分子機能のつながりや個々の分子機能の重要性を明らかにすることに繋がる。母系mRNPsのサブユニットについては、主に翻訳抑制が解除される際に活性を発揮する酵素などのはたらきや重要性が研究されてきており、母系mRNPsの転写抑制がどのように維持されているかといったmRNPsの機能そのものについては未だ不明な点が多い。酵素活性にはその立体構造が重要であるため、これらの酵素が生体内でどのような環境におかれているかを理解することは、分子機能にとどまらず、生体反応として理解するためにも重要である。複合体を構成している分子は、他の分子との相互作用によって、その活性が阻害されたり、反応速度が変化することが知られている (Hyde et al. 1988)。

そのため、母系mRNPsの翻訳抑制機能を詳細に理解するには、母系mRNPsの相互作用を明らかにする必要がある。多くのサブユニットを持つmRNPsの相互作用を明らかにするためには、立体構造を明らかにすることがもっとも直接的な方法であり、特にサブユニットどうしの相互作用と各サブユニットの分子機能の関連性を調べるには、原子レベルの分解能をもつ立体構造を得ることが肝要である。

そこで、本研究では、母系mRNPsによる翻訳抑制機構の解明を目的とし、その相互作用の変化の中心と考えられるCPSF複合体の立体構造を立体構造を生命情報学的手法をもちいて作成し、その立体構造をもとに、CPSF複合体を中心とした三次元相互作用、および貯蔵mRNPs複合体中におけるCPSF複合体の機能解明を行った。

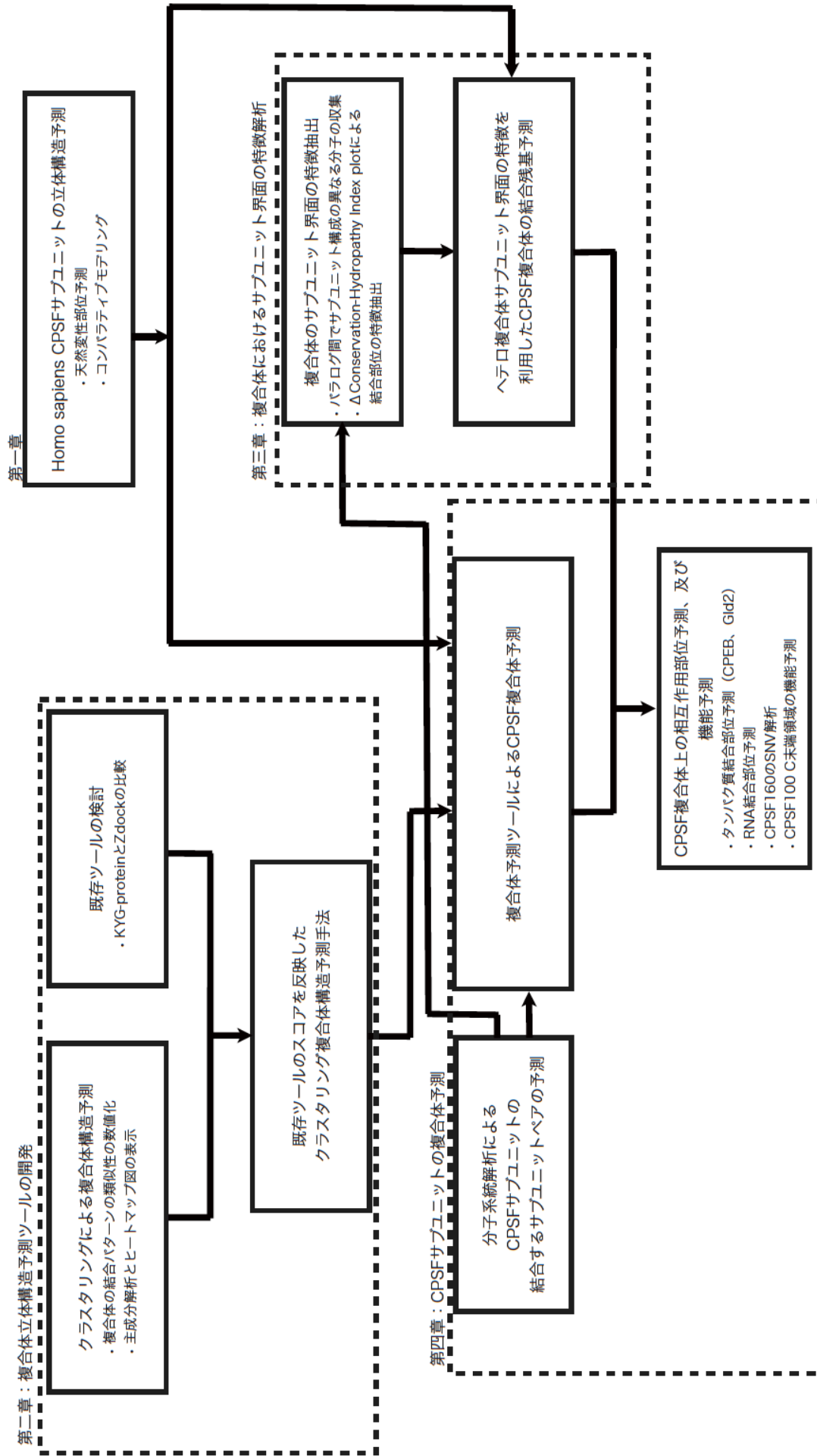
第一章ではCPSFサブユニットの予測立体構造を作成し、第二章ではCPSFサブユニットの相互作用を予測するために、3量体以上の複合体を想定した複合体構造予測ソフトウェアの開発を行った。また第三章では、今までの予測ソフトウェアでは予測しづらいサブユニット界面の特徴を調べ、CPSFサブユニットがそのようなサブユニット界面を有するか調べた。さらに第四章では分子系統解析を行うことで、CPSFのどのサブユニットどうしが結合しているか推定し、CPSF複合体の組み上げを行った。そしてCPSF複合体の予測立体構造をもとに、CPSFサブユニットの機能とGld2およびCPEBとの相互作用部位から予測される貯蔵mRNPsの転写抑制の維持を実現するメカニズムについて考察を行った(図序-2)。

本研究で使用したゲノム・タンパク質情報

mRNPsの翻訳制御に関する先行研究では*Xenopus leavis*の卵細胞を用いて実験を行った文献が最も多かった。そこで本研究でも*Xenopus leavis*のCPSF複合体の立体構造予測を試みたが、*Xenopus leavis*では、CPSF複合体のサブユニットであるCPSF160のアミノ酸配列全長、および全ゲノム配列が決定されていなかった。そのため、本研究では、全ゲノム配列が明らかであり、もっともアノテーションが豊富な*Homo sapiens*由来のタンパク質分子の情報を用いて行った。

さらにCPSF複合体のサブユニット構成は、動物界では広く保存されているが、Yeastや*Arabidopsis thaliana*では異なることが明らかになっている(Millevoi and Vagner 2010)。このことから、サブユニットの組み合わせは流動的であり、現時点で定まった定義がない。本研究では比較的広い生物種で複数の手法を用いて複合体の要素として確認されている、CPSF160、CPSF100、CPSF73、CPSF30をCPSF複合体のサブユニットと定義した。これら

のCPSFサブユニットの遺伝子情報およびタンパク質情報はアメリカ合衆国国立生物工学情報センター(NCBI: <http://www.ncbi.nlm.nih.gov/>)などのデータベースから収集した。



図序-2. 本研究の流れ

第一章

CPSFサブユニットの立体構造予測

背景

複合体立体構造予測は、各サブユニットのタンパク質表面の物理化学的特性を利用して、結合構造の予測を行う。サブユニット表面の特徴を捉えるには立体構造を取得するのが近道である。しかし、CPSFサブユニットの立体構造は、*Homo sapiens* CPSF73の立体構造の一部がX線結晶解析によって明らかにされているが、そのほかのサブユニットの立体構造は決定されていない。そこで本章では生命情報学的手法を用い、各サブユニットの立体構造を予測した。

立体構造予測の手法は、実験的に決定されたタンパク質の立体構造を何らかの形で利用している。しかし現段階ではすべてのタンパク質の立体構造のパターンが明らかにされているかどうかは分からない(Yura and Hayward 2006)。またタンパク質には特定の立体構造をとる部分と取らない部分がある。ここで立体構造をとる部分とは、単量体でとる部分とする。タンパク質は疎水性のアミノ酸残基をもつため、そのようなアミノ酸残基が多い部分がタンパク質内部に位置することで、一般的には単量体のタンパク質は球に近い構造をとると考えられている。しかし、実際にはNMRで立体構造の決定できない、可変的な領域が多くあることが分かってきており、このような領域を天然変性領域と呼ぶ。

タンパク質の天然変性領域

これまでもドメイン間をつなぐループ部分やアミノ酸配列の末端などはふらふらとした構造であり、立体構造を特定しにくい構造であることは知られていたが、そのような領域は比較的小さいものだと考えられていた。しかし、100残基以上の広い範囲に天然変性領域をもつ分子は珍しいものではないということが分かってきている。特に*Homo sapiens*においては、転写因子の約49%に天然変性領域を持つことが推定されている (Nishizawa 2009)。

この領域を広範囲にもつタンパク質は天然変性タンパク質と呼ばれ、この領域はそのタンパク質分子単体(または基質などが存在せず機能を発揮できない状態)では立体構造を形成しない(Zhang et al. 2007)。天然変性領域は、原核生物に比べ真核生物で、また転写やシグナル伝達などの他分子の認識や結合を必要とする分子群で、多く見つかっている(Uversky 2002)。また他のタンパク質と結合することによって定まった形をとるタンパク質も存在する(Wright and Dyson 1999, Tompa 2005)。このことから天然変性領域は、複数の分子との入り組んだ結合を可能にするため、単体のときは柔軟な状態を維持しているのではないかと考えられている (Nishizawa 2009)。

複合体を形成し翻訳制御というはたらきをもつ、CPSFサブユニットも天然変性領域をもつ可能性が高いと考えられた。そこで本研究では、各CPSFサブユニットが天然変性領域を有するか予測を行い、天然変性領域ではないと予測される領域について立体構造の予測を行った。またタンパク質立体構造では、最も一般的に行われ、かつ既知立体構造を利用した構造予測のうち、最も基本的な手法であるコンパラティブモデリングを用いた(図1-1)。

コンパラティブモデリング (ホモロジーモデリング)

コンパラティブモデリングとは、実験的に立体構造が決定されているタンパク質のアミノ酸配列と立体構造が未知のアミノ酸配列のアライメントを元に、立体構造未知のアミノ酸配列の立体構造を予測する手法である(Martí-Renom et al. 2000)。この時立体構造が決定されているタンパク質をテンプレートと呼び、このテンプレートの立体構造をもとに立体構造未知のタンパク質の構造を予測する。テンプレートとするタンパク質は、立体構造未知のタンパク質と相同なタンパク質、もしくは配列一致度の高いタンパク質を選択する。

これは与えられた環境におけるタンパク質の立体構造はそのアミノ酸配列によって決まり、そのフォールディングはその環境においてエネルギーが最小になるような構造をとる(アミノ酸配列が似ていれば、似た構造をとる)と考えるのが妥当であるという仮説(Anfinsen et al. 1961)、および何らかの進化的な関係性があるタンパク質どうしは似ている傾向があるという経験則(Chothia and Lesk 1986)に基づく。

コンパラティブモデリングでは、テンプレートの配列と立体構造が未知の配列との配列一致度は理想的には40%以上、少なくとも35%以上は必要と考えられている(Contreras-Moreira et al. 2003, Sadowski and Jones 2007)。テンプレート配列を一本に絞り込むことができない場合は、複数の類似タンパク質のアミノ酸配列をマルチプルアライメントしたものをテンプレートとして用いて、コンパラティブモデリングを行う。特にテンプレートの立体構造の一部が欠けていたり、構造をつくりたいタンパク質のアミノ酸配列よりテンプレート配列が短い場合、欠損部分を補う立体構造をもつ類似配列を組み合わせアライメントを作成することで、立体構造予測の精度をあげることができる(Kosinski et al. 2003)。

コンパラティブモデリングでは、作成したアライメントに基づいて、テンプレートの立体構造のアミノ酸残基の側鎖を、対応する立体構造未知のアミノ酸残基の側鎖に置き換える。こうして作成した予測立体構造は、側鎖の形状によっては不自然な構造となっている場合があるため、エネルギーが低くなるように構造を再計算することで安定な立体構造を得る(Summa et al. 2007)。

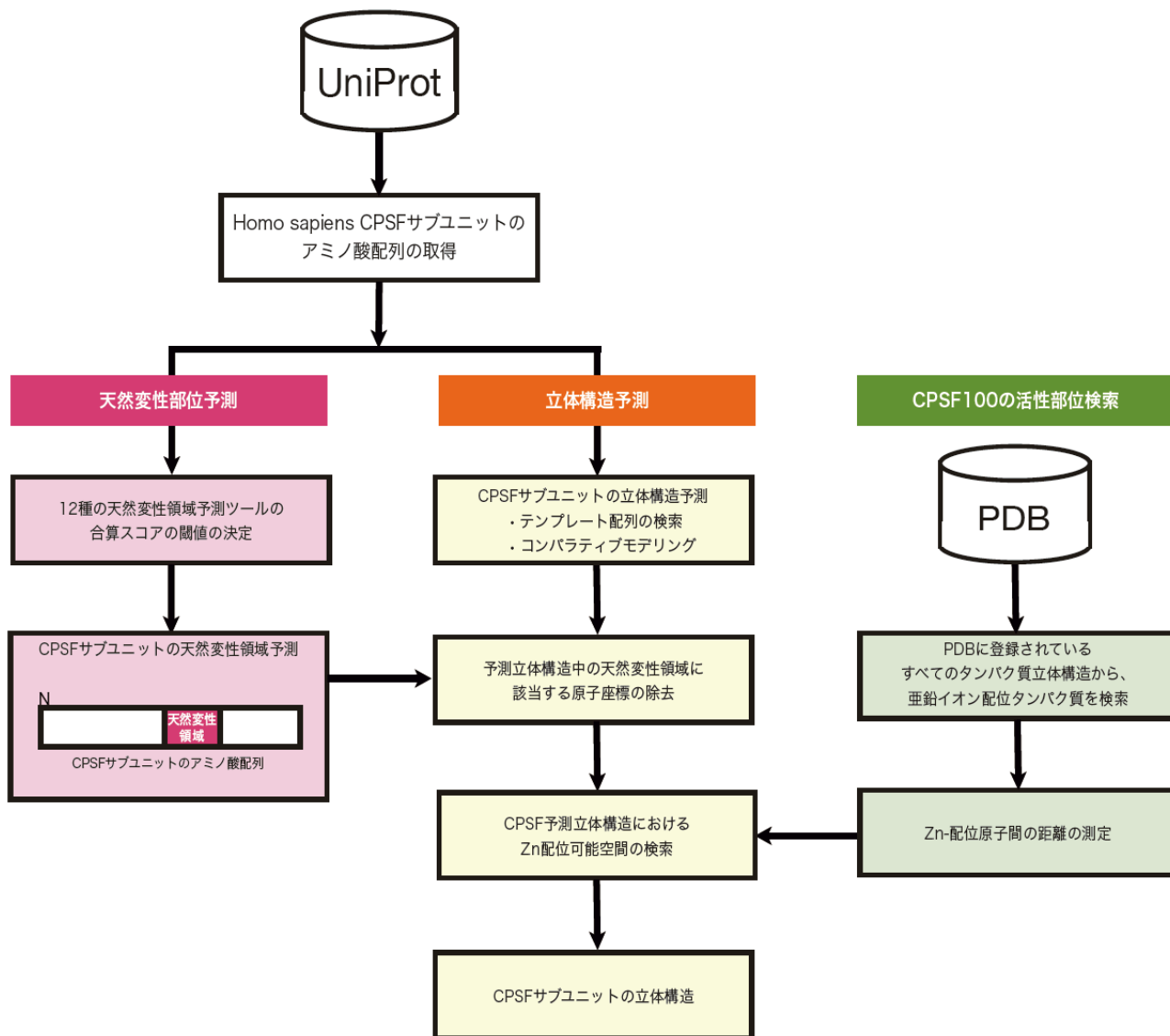


図1-1. 本章の流れ

本章では、*Homo sapiens* CPSFサブユニットのアミノ酸配列を用いて、天然変性領域予測（マゼンダ色で示す手順）、および立体構造予測（オレンジ）を行った。また、*Homo sapiens* CPSF100は、エンドヌクレアーゼ活性を示す活性部位がアミノ酸配列上では保存されていないことが明らかであったので、エンドヌクレアーゼ活性に必要な亜鉛イオンを配位するアミノ酸残基、もしくは原子が立体構造中に存在するか、調べた（緑）。

方法

1. CPSFサブユニットのアミノ酸配列の入手

本研究で用いた*Homo sapiens*のCPSFサブユニットのアミノ酸残基配列は、文献をもとにしたアノテーション情報のあるUniProt (The UniProt Consortium 2011)のデータから収集した。部分配列などを解析に使用してしまうことを防ぐため、CPSFサブユニットの分子名が、UniProtのデータにおいて、実験的に確認された分子名としてアノテーションされているものに限った。

2. CPSFサブユニットの天然変性領域の検索

天然変性領域は単体では定まった構造を取らないため、単体の立体構造を予測することはできない。そこでCPSFサブユニットの立体構造予測に先立って、予測不可能な領域がどの程度あるか、調べることにした。

2-1. 天然変性領域予測プログラム、およびその利用方法

アミノ酸配列から天然変性領域を予測するソフトウェアは既に多く存在するため(Ferron et al. 2006)、CPSFサブユニットの天然変性領域の予測は既存のソフトウェアを使用した。バイオインフォマティクスにおいて、特にアミノ酸配列からなんらかの予測を行なう場合、多数の異なる予測プログラムの結果を統合することで予測精度が向上することが示されている(Ginalski et al. 2003)。

本研究でも、様々な計算手法を使用した天然変性領域予測ツール10個を用い、12種の天然変性領域の予測を行った。これらのツールは、それぞれスコアの表示の仕方が異なるため、全てのツールにおいて、構造をとると予測される領域のスコアを0.0、天然変性領域と

予測される領域のスコアを1.0になるように変換し、各ツールのスコアの重みを統一した。DISOPRED2、FoldIndex、GlobPlot2、DisEMBLはアミノ酸残基一つ一つがDisorder(天然変性領域)かOrder(構造をとる領域)で結果が表示されるため、結果がDisorderとなっているアミノ酸残基のスコアを1.0とし、Orderとなったアミノ酸残基のスコアは0.0とした。POODLE、DISPROT、IUpred、RONN、DISpro、SPRITZは、スコアが0.00から1.00の間の数値として算出されるのでスコアをそのままプロットした。これらの予測プログラムでは、スコアが0.50以上のアミノ酸残基を天然変性領域となる残基としている(Ferron et al. 2006)。そこで、本研究でも12個の予測結果の平均スコアが0.5以上であったことを示す、スコア合計値が6.0以上の領域を天然変性領域とした。

2-2. CPSF160、CPSF100、CPSF30の天然変性領域の推定

*Homo sapiens*のCPSF73は一部の立体構造が決定されている(Mandel et al. 2006)。非構造予測領域は基本的には実験的に立体構造が決定できない領域であるので(Wright and Dyson. 1999)、12種の天然変性領域予測プログラムのスコア(図1-2. a)からその合算値(図1-2. b)を算出し、*Homo sapiens* CPSF73の立体構造情報のあるアミノ酸配列部分(図1-2. c)と比較することで、スコア合計値6.0以上が天然変性領域の予測結果として妥当であるかを検証した。また、この検証結果から得られた合計スコアの値をもとにCPSF160、CPSF100、CPSF30の天然変性領域の推定を行った。

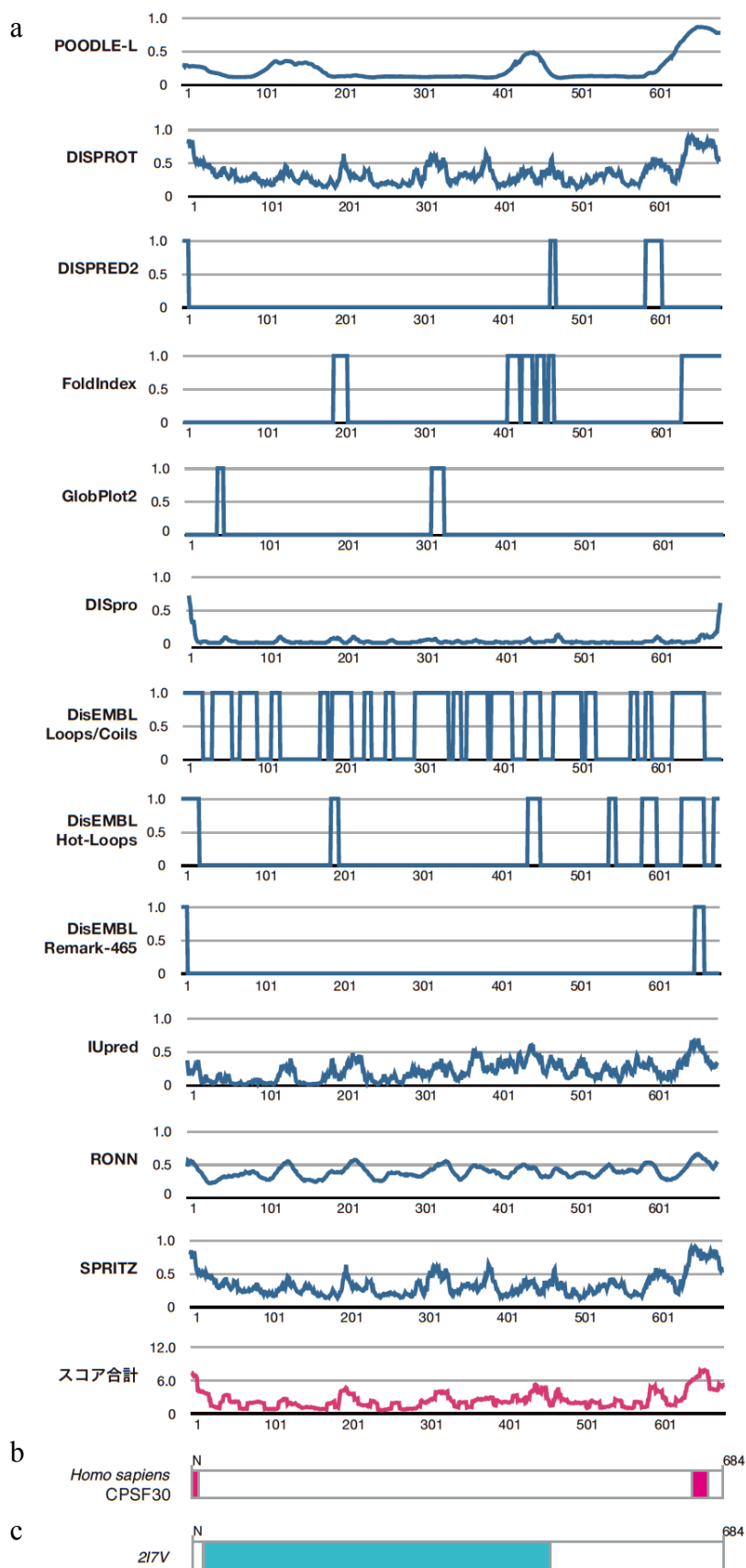


図1-2. CPSF73の天然変性領域検索結果

a. 横軸はアミノ酸残基番号を、縦軸はスコアを示している。青い折れ線グラフでは表1に示した各天然変性領域予測ツールによるスコアを、マゼンダの折れ線グラフではスコア合計値を示した。b. 12種の予測ツールのスコアの合計値が6.00以上のアミノ酸残基位置をマゼンダで示した。c. *Homo sapiens* CPSF73のアミノ酸配列上で立体構造 (PDB ID: 2I7VA) が得られている領域を青緑色で示した。

3. CPSFサブユニットのテンプレート配列の検索

コンパラティブモデリングのために、実験的に決定された立体構造のデータベースである、PDB(Berman et al. 2000)に登録されているすべての立体構造情報に対し、類似性検索(blastp: Altschul et al.1990)を実施することにより、テンプレート配列を検索した。このとき、E-valueが 10^{-4} 以下のアミノ酸配列のみを収集し、このうち、配列一致度が高く、各CPSFサブユニットのアミノ酸配列の全長わたって、類似なアミノ酸配列をもつ立体構造を選択した。テンプレート配列の取得のために使用したアミノ酸配列は、UniProtに登録されていたアミノ酸配列を用いた。各CPSFサブユニットのUniProt IDはQ10570.2(*Homo sapiens* CPSF160)、Q9P2I0.2(*Homo sapiens* CPSF100)、Q9UKF6.1(*Homo sapiens* CPSF73)、O95639.1(*Homo sapiens* CPSF30)である。

4. コンパラティブモデリングによるサブユニットの立体構造予測

Homo sapiens CPSFサブユニットのアミノ酸配列と、類似性検索により取得したテンプレート配列のアラインメントには、ALAdGAPsを用いた(Hijikata et al. 2011)。このアラインメントをもとに、コンパラティブモデリングを行った。

コンパラティブモデリングにはMODELLER(Webb et al. 2014)を使用した。MODELLERでは、エネルギーの異なるモデルを複数作成することができる。本研究では各サブユニットごとに5つのモデルを作成し、構造全体のエネルギーの最も低い構造を予測構造として選出した。さらに10残基、または40残基ごとにタンパク質の構造のエネルギーを計算するProSA(Wiederstein and Sippl 2007)を用いて、得られた予測構造に部分的にエネルギーが高くなるような不安定な構造がないか確認した。

5. CPSF100における亜鉛イオン配位部位探索

CPSF73とCPSF100はパラログな分子であるため、その立体構造には類似性があると推察された。*Homo sapiens* CPSF73の立体構造はその一部がX線結晶解析により明らかになっており(PDBID: 2I7V)、CPSF73はタンパク質内部に2つの亜鉛イオンを配位していることが明らかになっている (Mandel et al. 2006)。これら配位子は、アミノ酸配列上でHXHXDH (Xは任意のアミノ酸残基)と呼ばれる、CPSF73において高度に保存されたモチーフとして確認できる(Dominski et al. 2010)。このことから、CPSF100がCPSF73と類似の構造をとっているならば、CPSF100がその立体構造維持のためにCPSF73同様に亜鉛イオンを配位している可能性があると考えられた。そこでCPSF100の予測立体構造が亜鉛イオンを配位できるか調べた。

5-1. 亜鉛イオン配位タンパク質における亜鉛イオン-配位子間の原子間距離の測定

CPSF73の立体構造では1つの亜鉛イオンが配位するために4つの配位子が関与しており、2つの亜鉛イオンが1つの配位子を共有する形で7つの原子と配位している。ここで配位に使われている原子は窒素、酸素、硫黄原子であった。そこで、これらの原子が亜鉛イオンの配位子となるためにはどの程度の原子間距離が必要であるか、PDBに登録してあるすべての亜鉛イオンを配位するタンパク質を対象に、亜鉛を配位する際の亜鉛イオン-配位子間間の距離を測定することとした。立体構造中の亜鉛原子の探索、および亜鉛イオン-配位子間間の距離の測定はインハウスプログラムにより行い、得られた原子間距離を原子種ごとに棒グラフにし、その分布を調べた。

5-2. CPSF100予測立体構造における亜鉛イオン配位子の検索

タンパク質が亜鉛イオンを配位する際の亜鉛イオンと配位子間の平均距離とその分布が求められたので、この距離を満たすように4つの原子が位置しているような空間があるか、CPSF100の予測立体構造および CPSF73の予測立体構造を対象に調べた。

配位子となる4つ原子の探索は、次のように行った。まずCPSF100の予測立体構造および CPSF73の予測立体構造の座標情報から窒素原子、酸素原子、硫黄原子を探索し、見つかった原子の近傍にさらに3つ、窒素原子、酸素原子、硫黄原子のいずれかの原子が存在するか検索する。窒素原子、酸素原子、硫黄原子のいずれかの原子が4つ集まっている箇所があれば、その4原子の重心を計算し、重心からその4原子の距離が亜鉛イオンと配位子間の平均距離およびその分布を満たすか計算する。重心と4つの原子の距離が亜鉛イオンと配位子間の平均距離およびその分布を満たしている場合、それらの4分子をもつアミノ酸残基を配位子候補として選出した。

結果

1. CPSF構成因子の天然変性部位

CPSF73の天然変性領域の予測結果と立体構造の比較から、12種の天然変性領域のスコア合計を6.0とすることは妥当と考えられたので、このスコアを閾値とし、CPSF160、CPSF100、CPSF30の天然変性領域の予測を行った。CPSF160、CPSF100、CPSF30の天然変性領域結果をそれぞれ図1-3、図1-4、図1-5に示す。この結果、CPSF160およびCPSF100では、複数の天然変性領域が予測されたが、いずれも50残基程度の短い天然変性領域が予測された(図1-3. b、図1-4. b)。このことからCPSF160およびCPSF100は配列全長の立体構造を予測することが可能であることがわかった。また、CPSF30は全長270残基のうち70残基(残基番号172~272)が天然変性領域として推定された(図1-5. b)。CPSF30はアミノ酸配列全長の約4分の1が天然変性領域であることがわかり、CPSF30の多くの部分が単体の立体構造を予測できないことが分かった。

2. CPSF構成因子のテンプレート配列の検索

PDBを対象とした類似配列検索の結果、CPSF160は*Thermus thermophilus*のリボヌクレアーゼ TTHA0252の立体構造であるPDB ID: 3E11、CPSF73は*Homo sapiens* CPSF73の部分立体構造であるPDB ID: 2I7V、CPSF30は*Homo sapiens* CPSF30のPDB ID: 2D9Nの既知構造を取得した。CPSF100は400~500残基の間に複数の天然変性領域を持っていたため(図1-4. b)、その部分のテンプレート配列は得られず、CPSF73のテンプレートと同一のPDB ID: 2I7Vを375残基で二つに分割した前後の領域がそれぞれ、CPSF2_HUMANの1残基目から481残基目以前の配列と481残基目以降から782残基まで配列と類似であった。そこで

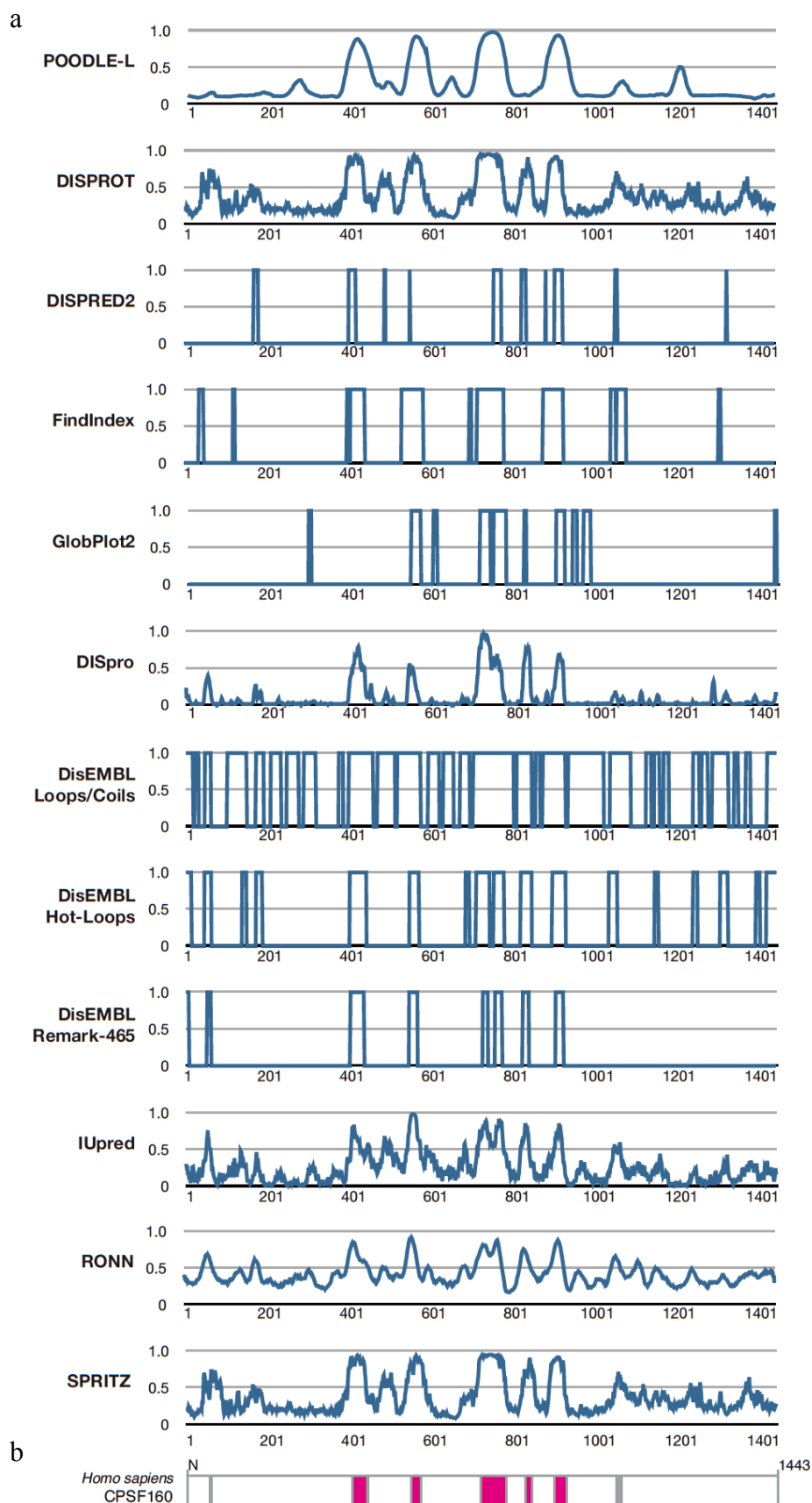


図1-3. CPSF160の天然変性領域の推定結果

a. 横軸はアミノ酸残基番号を、縦軸はスコアを示している。青い折れ線グラフでは各天然変性領域予測ツールによるスコアを示した。b. 12種の予測ツールの各スコアの合計値が6.00以上のアミノ酸残基位置をマゼンダで示した。

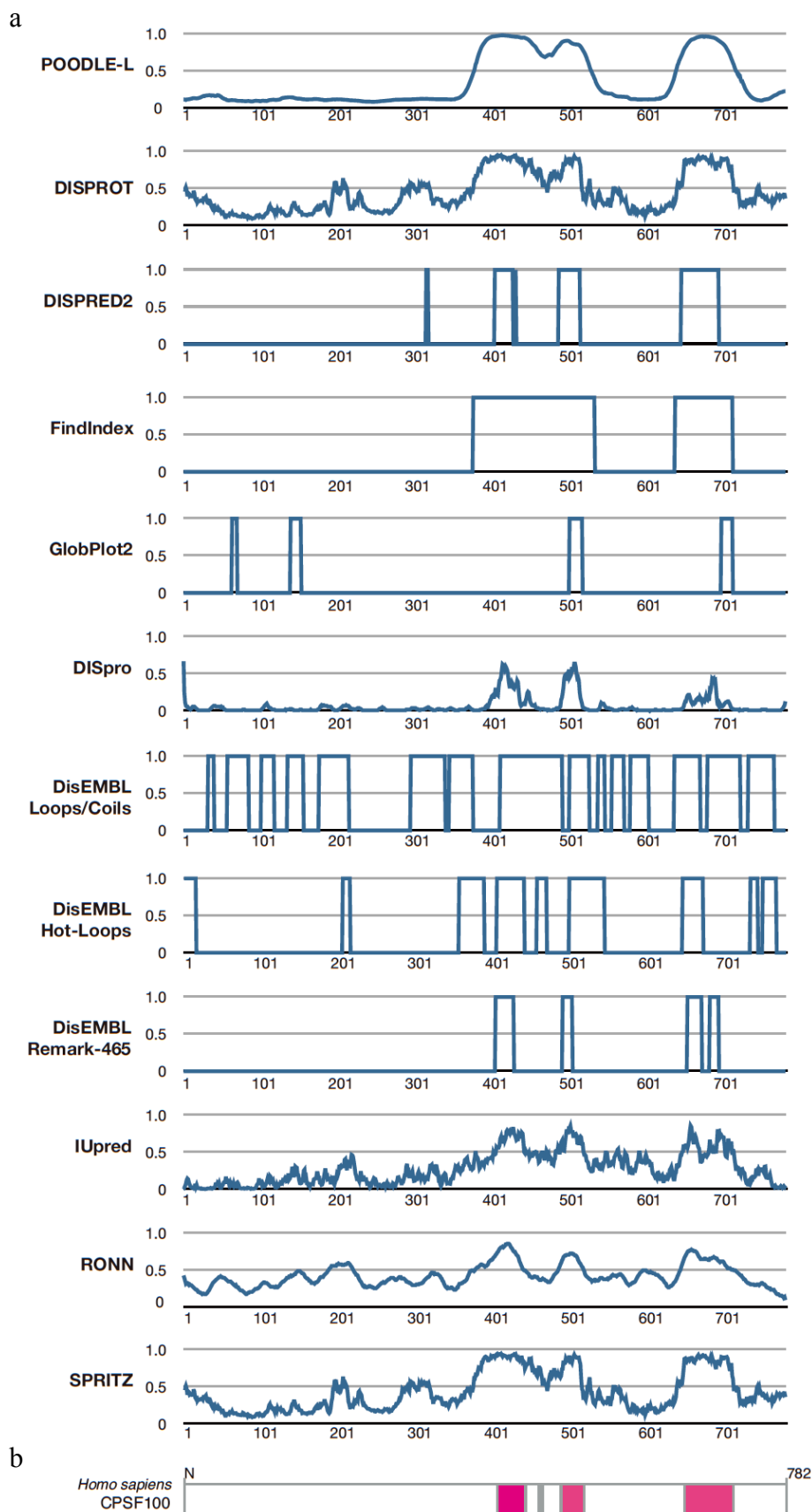


図1-4. CPSF100の天然変性領域の推定結果

a. X軸はアミノ酸残基番号を、Y軸はスコアを示している。青い折れ線グラフでは各天然変性領域予測ツールによるスコアを示した。b. 12種の予測ツールの各スコアの合計値が6.00以上のアミノ酸残基位置をマゼンダで示した。

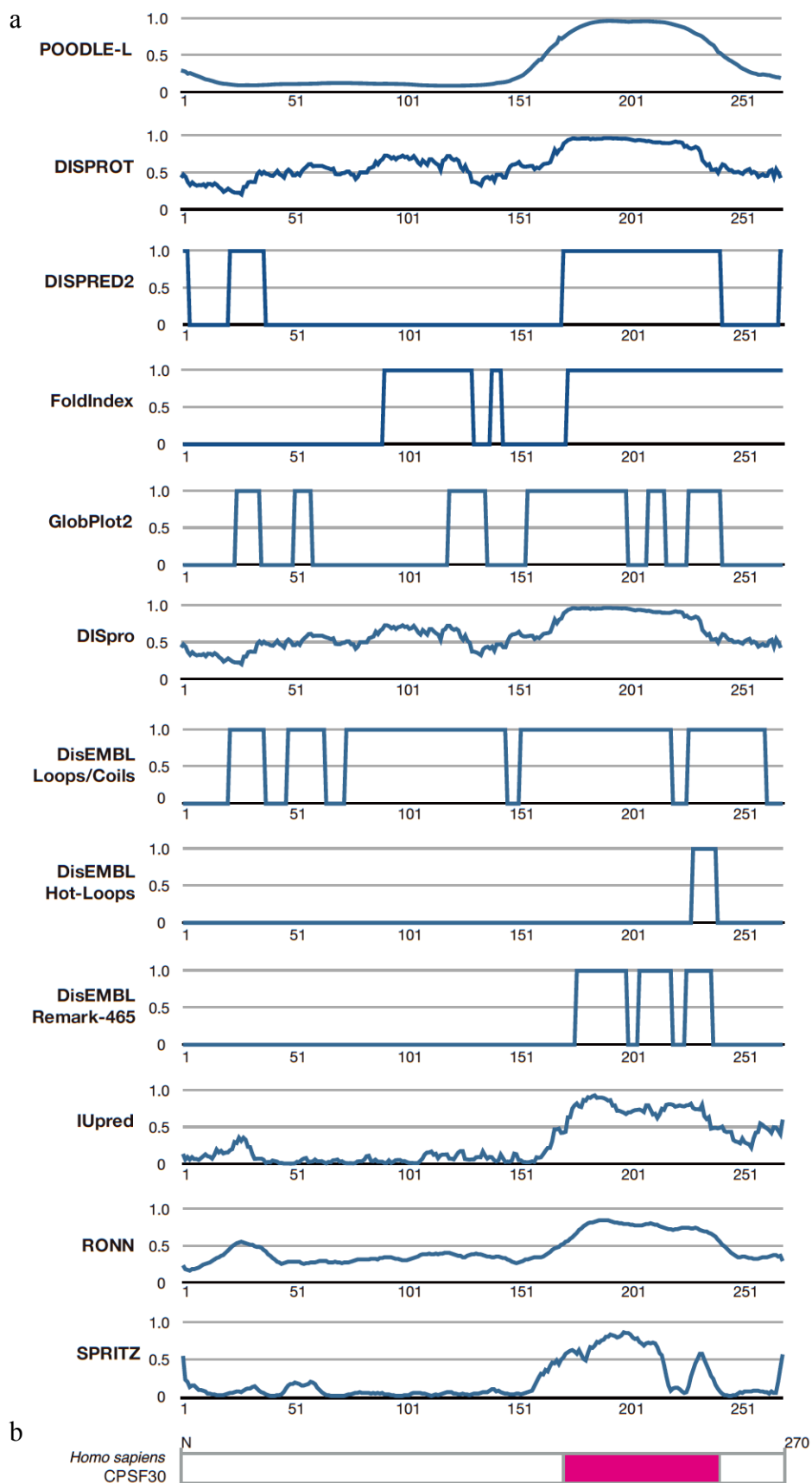


図1-5. CPSF30の天然変性領域の推定結果

a. 横軸はアミノ酸残基番号を、縦軸はスコアを示している。青い折れ線グラフでは各天然変性領域予測ツールによるスコアを示した。b. 12種の予測ツールの各スコアの合計値が6.00以上のアミノ酸残基位置をマゼンダで示した。

CPSF160、CPSF73、CPSF30は得られた構造をそのままテンプレート構造とし、CPSF100は375残基前後で2IV7を分割したものの2つをテンプレート構造とした。

CPSF30の立体構造予測に用いたPDB ID: 2D9Nの既知構造のアミノ酸配列は配列長が77残基しかなく、CPSF30が単体でも立体構造をとると考えられる領域の半分以上の立体構造を予測することができなかった。また2PDB ID: 2D9Nの既知構造以外に、CPSF30とアミノ酸配列の類似性が比較的高く、かつアミノ酸配列の配列長が長いものがなかったため、他の既知構造を2D9Nで予測ができなかった部分のテンプレートとすることもできなかった。結果としてCPSF30は、PDB ID: 2D9Nの既知構造をテンプレートとして予測した約80残基の領域しか立体構造を作成することができず、アミノ酸配列全長の半分以上の構造が不明であることから、CPS30のタンパク質表面が十分に予測出来ない可能性が高いと考えられた。複合体構造予測にはタンパク質間結合部位となるタンパク質分子表面が得られていることが重要であるため、CPSF30の立体構造は複合体構造予測に使用しないこととした。

3. 各サブユニットの予測立体構造

予測したCPSFサブユニットの立体構造を図1-6に示す。今回ホモロジーモデリングに用いたMODELLERでは、天然変性領域は分子の外側に突き出たループとして構造が予測される。このような構造は結合部位予測の妨げになるので、予測立体構造において天然変性領域であるためにループと予測された構造は、削除しなければならない。本研究では、CPSF100のC末端側の天然変性領域(図1-4. b)が明らかに飛び出たループとして予測されたため、CPS100の予測立体構造ではこの領域の座標を削除した。

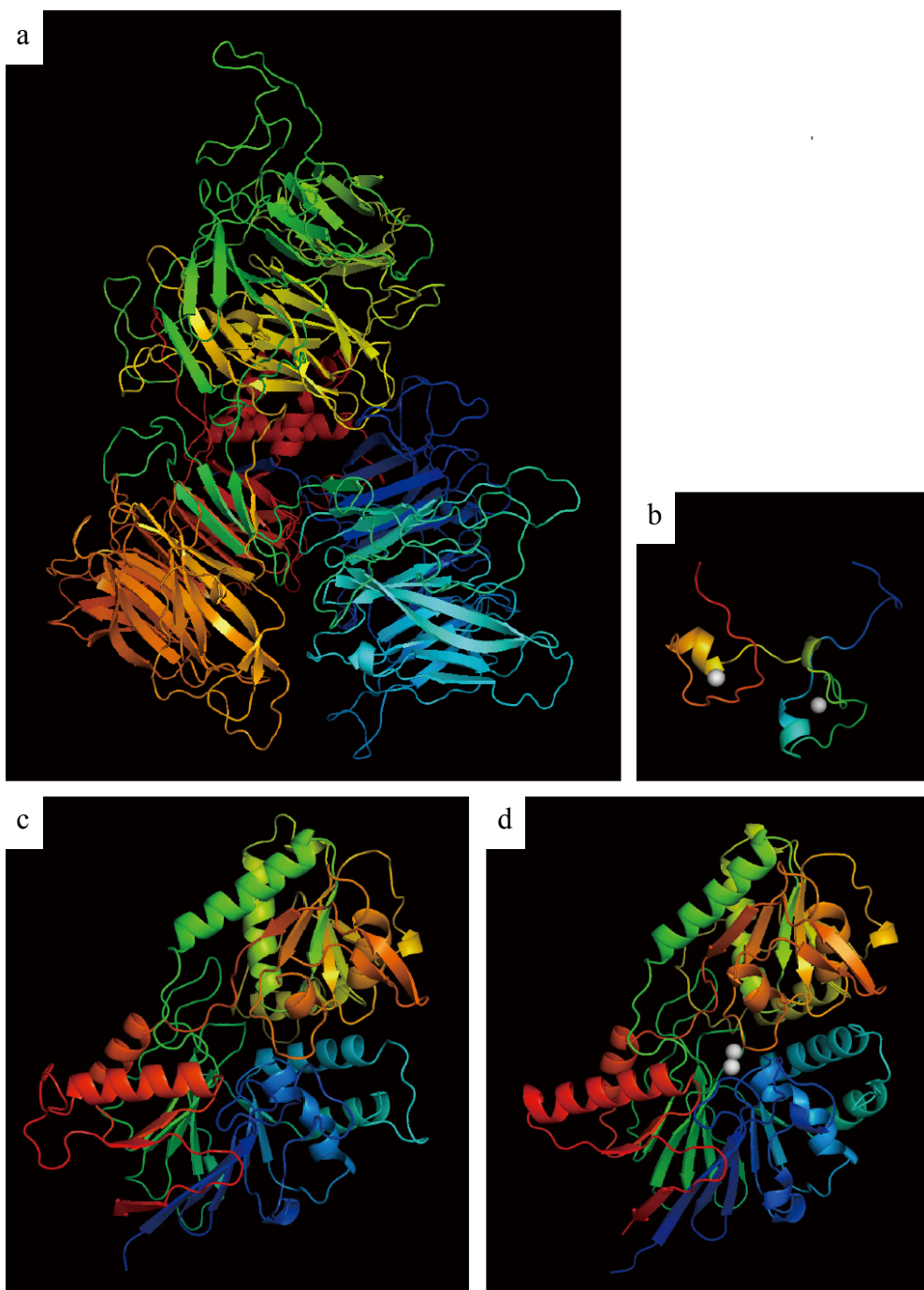


図1-6. MODELLERで得たCPSFサブユニットの予測立体構造

Homo sapiens CPSF160 (a)、*Homo sapiens* CPSF30 (b)、*Homo sapiens* CPSF100 (c)、*Homo sapiens* CPSF73(d)の構造を示す。CPSF30およびCPSF73は亜鉛イオンを配位することが分かっていたので、テンプレート構造とアミノ酸配列の主軸を重ね合わせることで、予測立体構造上の亜鉛イオンの配位位置を予測した。

4. CPSF100における活性部位探索

Homo sapiens CPSF73でも、亜鉛イオン配位部位となるHXHXDHモチーフがアミノ酸配列上に存在していたが、*Homo sapiens* CPSF100のアミノ酸配列にはこのモチーフはみられなかった。しかしモチーフとしては保存されていなくとも、立体構造上で探索すると配位子となるアミノ酸残基、または亜鉛イオンが配位できる原子が空間的に保存されている可能性がある。そこでCPSF73およびCPSF100の予測立体構造を対象に、亜鉛イオンの配位子が保存されているか調べた。

亜鉛イオンの配位に使われている原子は窒素、酸素、硫黄原子であった。これらの原子が亜鉛イオンを配位する際の亜鉛原子-配位子間の適切な距離を測定したところ、最も存在頻度の高い配位距離は、亜鉛イオン-窒素原子間では2.1Å、亜鉛イオン-酸素原子間では2.1Å、亜鉛イオン-硫黄原子間では2.3Åであった (図1-7)。

この結果から、CPSF73の予測立体構造およびCPSF100の予測立体構造に対して、窒素、酸素、硫黄のいずれかの原子4つが互いに隣接して存在するかを探索し、これらの原子が4つ集まっている場所があった場合、4つの原子からその重心までの距離が2.3Å以内にあるか調べた。CPSF73の予測立体構造における配位子探索の結果を図1-8. a、CPSF100における配位子探索の結果を図1-8. bに示す。*Homo sapiens*のPDBデータを鋳型に作成した*Xenopus leavis* CPSF73では、template PDB(図1-8. a)で配位子として働くアミノ酸残基が亜鉛イオンの配位子として選択されたが(図1-8. b)、同じ*Homo sapiens*のPDBデータを鋳型に作成した*Xenopus leavis* CPSF100では、CPSF73における亜鉛イオンの配位部位付近には配位子は予測されなかった(図1-8. c)。このことからCPSF100は亜鉛イオンを配位しておらず、CPSF100はエンドヌクレアーゼ活性をもたないと考えられる。

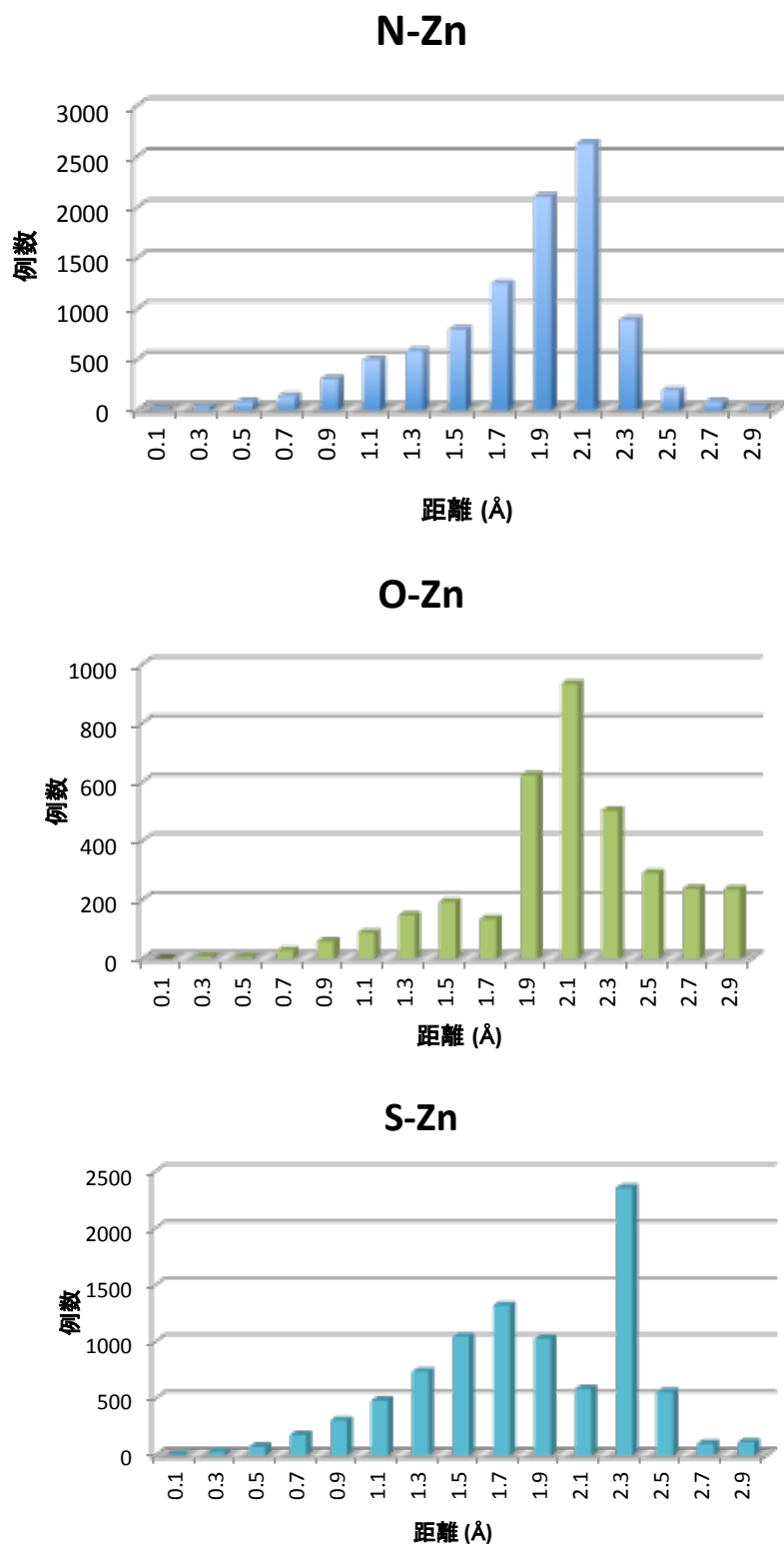


図1-7. 亜鉛イオンを配位するタンパク質中の亜鉛イオンと各配位子の距離
 PDBから、亜鉛イオンを配位するタンパク質すべてを選出し、タンパク質中で、亜鉛イオンが配位する原子と、その距離を調べた。亜鉛イオン(Zn)を配位する原子として、窒素原子(N)、酸素原子(O)、硫黄原子(S)が得られた。各グラフは3種の配位子における、亜鉛イオンとその配位子となる原子との原子間距離、およびその原子間距離を有する配位原子数(例数)を示す。

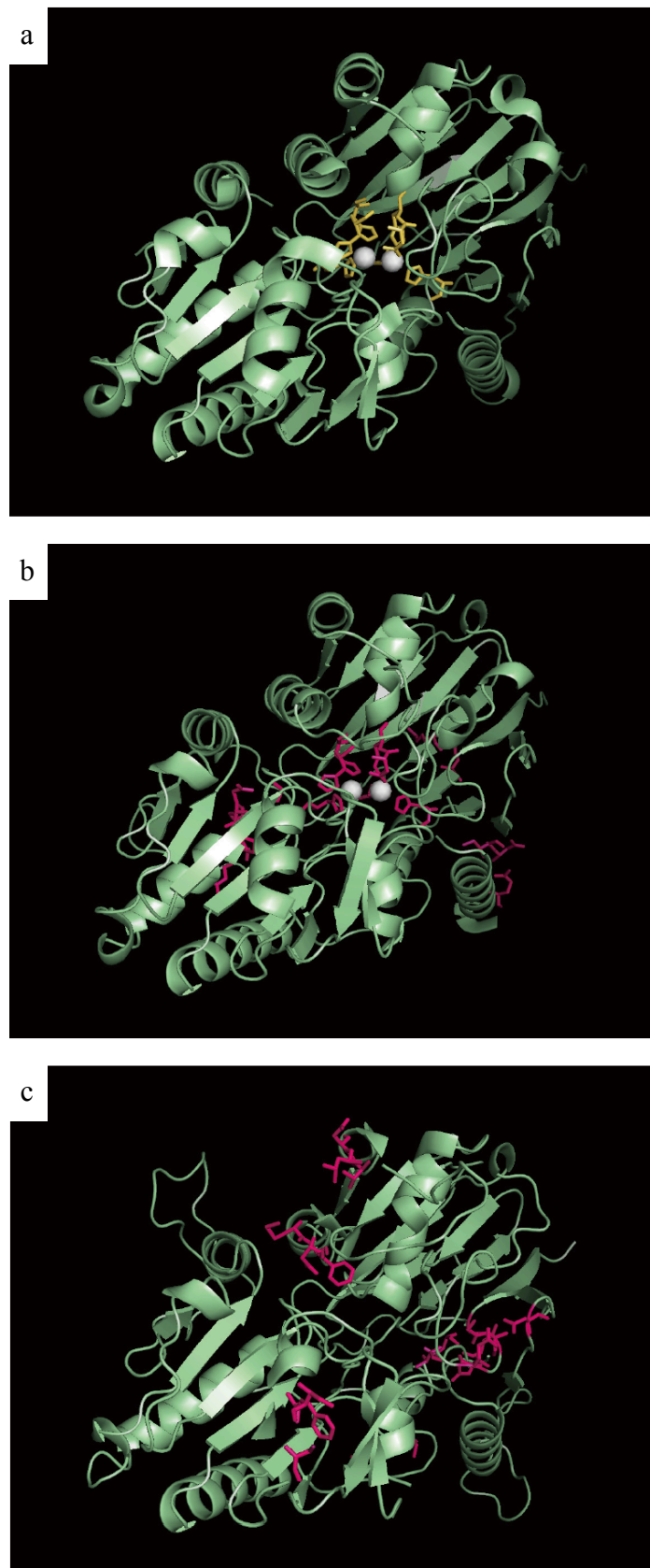


図1-8. 予測した亜鉛イオンを配位しているアミノ酸残基

Xenopus leavis CPSF73 の予測立体構造(a) 及び *Xenopus leavis* CPSF100の予測立体構造 (c) 上に予予測された亜鉛イオン配位子となりうるアミノ酸残基を赤で示した。また 既知構造である *Homo sapiens* CPSF73 において実際に亜鉛イオンを配位しているアミノ酸残基を黄色で示した。

考察

高次複合体を構成するタンパク質分子は、特定の構造を持たない天然変性領域と呼ばれる領域を多く持つことが知られている(Ward et al. 2004.)。天然変性領域をもつ分子は天然変性領域をもたない分子に比べ、柔軟な構造となるため、特定の分子との結合に有効だと考えられている(Ward et al. 2004)。しかし特定の条件下でないと構造を取らない配列のため、天然変性領域の配列が取りうる構造を予測することは非常に困難である。そこで天然変性領域を特定し、それ以外の部分の立体構造を予測した。

天然変性領域予測の結果、*Homo sapiens* CPSF73を除いた*Homo sapiens* CPSF160、*Homo sapiens* CPSF100、*Homo sapiens* CPSF30の3つのサブユニットは、アミノ酸配列の中央部付近、もしくは中央部からややN末端側に100残基程度の天然変性領域があることが分かった(図1-3. b, 1-4. b, 1-5. b)。配列全長の中央部に大きな非構造領域が存在すると立体構造の予測が困難となるが、CPSF160、CPSF100、CPSF73の天然変性領域は配列全長の20%以下であり、これら3つのサブユニットについては、全体構造を充分反映した立体構造を予測できた。

一方、CPSF30では天然変性領域が全長の約3分の1近くあり、天然変性領域の特性はCPSF30の単体としての機能に大きく影響を与えていると考えられる。天然変性領域は転写やシグナル伝達などの複数の分子の認識や結合を必要とする分子群でよく見られる構造であり(Nishizawa 2009)、CPSF30はCPSF複合体中で複数のサブユニットと結合、または結合を補助する様な分子機能があると考えられる(Kaufman et al. 2004)。

CPSF30も他のCPSFサブユニット同様CPSF複合体中でどのサブユニットと結合するか、またサブユニット単体としての機能も明らかになっていない。分子生物学実験により、

CPSF30は様々な特性をもつことが明らかになっている。CPSF30はmRNAとの結合を担っていると考えられるZinc fingerドメインを複数もっており(Barabino et al. 1997, Xiang et al. 2014)、特にCPSF30は、Hex配列近傍に存在するU-rich領域に高い親和性を示すことから、CPSF30-CPSF160間の結合が示唆されている(Kaufmann et al. 2004)。

そのため、CPSF30がどのようにCPSF複合体に結合しているかはその機能を検証する上で重要であると考えられるが、CPSF30は低分子量でありながら立体構造の予測できた領域がアミノ酸配列全長の4分の1しかなかった。複合体構造予測にはタンパク質表面の立体構造が重要であるため、これ以後のCPSF30は複合体構造予測には用いないこととした。

CPSF73とCPSF100はMBLスーパーファミリーに属する分子であり、互いにとても良く似たドメイン構成の分子である。しかしCPSF73ではN末端に10残基、C末端に20残基程度の天然変性領域が予測されたのに対し、CPSF100では400残基から500残基の範囲に30~60残基の複数、600残基目から700残基目にかけて大きな天然変性領域が予測された。この部分はCPSF73およびCPSF100に特異的なC末端のドメイン(図1-9)にあたり、CPSF73とCPSF100のC末端ドメインは、配列類似性は低くないにもかかわらず、CPSF100のC末端領域のみに大きな非構造領域が予測されたことになる。このことからCPSF100のC末端側ドメインはCPSF73のC末端側ドメインと比較し柔軟な構造をしており、この領域はCPSF73とは異なる性質をCPSF100の機能に与えていると考えられた。

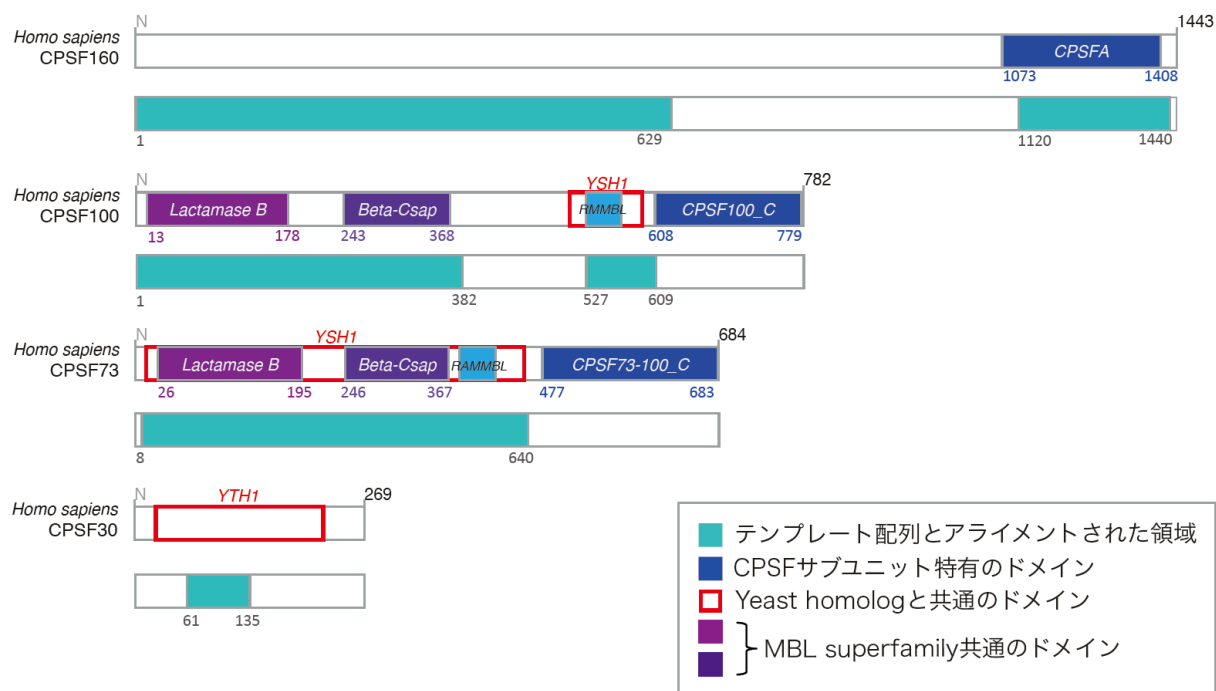


図1-9. CPSFサブユニットのドメイン構成とテンプレート配列の見つかった領域
 各CPSFサブユニットのドメイン構成を上側にボックスで表示し、ドメインごとに色分けし、ドメイン名をボックス内に記載している。各サブユニットのアミノ酸配列とテンプレート配列がアラインメントできた領域を下側に並べて緑色で表示した。またその既知構造のPDB IDをボックス内に記載した。全てのCPSFサブユニットで、テンプレートとして使用できる既知構造が見つかった。しかし、CPSF73およびCPSF100特有の配列であるC末端部分と類似性の高い配列構造は得られなかった。

第二章

複合体立体構造予測ツールの開発

背景

タンパク質分子の立体構造の決定方法としては、X線結晶解析やNMRによる解析が最も一般的な手段である。しかし、NMRでは分子量の大きいものでも50kDa程度を対象としており(Markley et al. 2003)、X線結晶解析では一般的には数十kDa、比較的大きな分子量のタンパク質を扱う場合でも数百kDaを対象としている(Berman et al. 2013)。そのため、これらの解析手法のみでは、数千kDa以上になるタンパク質複合体の立体構造の決定を行なうことはできない。これらの解析手法にかわって、近年、タンパク質複合体の立体構造解析は、Cryo-EM(非晶質凍結切片クライオ電子顕微鏡)が広く用いられるようになった(Bai et al. 2015)。しかし、Cryo-EMはX線結晶解析などと比較すると解像度が低く、Cryo-EMだけでは原子レベルの立体構造を明らかにすることはできない。そのため、Cryo-EMによるタンパク質複合体の立体構造解析は、X線結晶解析、NMR、コンパラティブモデリングやドッキング手法など様々な手法を組み合わせで行われる(Lindert et al. 2009)。この様な現状から、タンパク質複合体の立体構造決定には、生命情報学的手法を用いた複合体立体構造の予測が期待されている(Shi et al. 2014)。そこで本研究では、CPSFのような多数のサブユニットが結合する複合体を想定した複合体構造予測手法の開発を行った。

二量体の立体構造予測と三量体以上の複合体の立体構造予測

どの分子とどの分子が結合しているかという問題は、それらの分子の機能や複雑な機能調節への理解に直結するため(von Mering et al. 2002)、分子生物学研究を進める重要な要素と言える。そのため、タンパク質間相互作用の検出には様々な手法が用いられており(Ito et al. 2001, Ho et al. 2002, Cho et al. 1998, Marcotte et al. 1999)、近年では大規模なタンパク質間相互作用の検出も行われるようになってきた(Rual et al. 2005)。

三量体以上の複合体の構造予測においても、複合体のどのサブユニットどうしが直接結合しているのかは非常に重要な手がかりである。しかし、実験手法によるタンパク質間相互作用の検出には人手がかかり、また手法によっては検出限界もある(von Mering et al. 2002)。そのためどのサブユニットどうしが直接結合しているのか明確ではない複合体を対象とする可能性が高い点において、二量体の複合体構造予測にはない困難さが生ずる場合がある。

二量体のタンパク質複合体の立体構造は、ZDOCK (Chen and Weng. 2002)、PatchDock(Schneidman-Duhovny et al. 2005)、ClusPro(Comeau et al. 2004)、FireDock(Andrusier et al. 2007)、GRAMM(Tovchigrechko and Vakser. 2006)、RosettaDOCK(Lyskov and Gray. 2008)など多くのドッキングツールが存在しているため、比較的簡単に予測を行なうことができる。これらのドッキングツールは、結合されるタンパク質（レセプター）と結合するタンパク質（リガンド）を想定した1対1のドッキングを行うプログラムである。ドッキングプログラムは始めのステップで分子ペアが取りうるすべての結合構造（候補構造）を人工的に作り出すが、この行程は片方の分子の立体構造を固定してもう一方の構造を回転・移動させることで行われる(Li et al. 2003, Chen et al. 2003)。二量体ではすべてのパターンをある程度効率的に求めることができる。

一方、三量体以上の分子の結合の全パターンを求める場合では、ひとつの分子を固定してもX、Y、Z軸の移動とX、Y、Z軸方向の回転の全パターンを計算すべき分子が複数になるため二量体の場合に比べ格段に計算に要する時間が長くなる。そのため三量体以上の複合体において、一度にすべてのサブユニットの結合のパターンを計算によって求める場合、想定されるパターン数が増大するため計算量が非常に多くなってしまう(Alber et al. 2007)。

本研究では1サブユニットずつ組み上げるようにドッキングを行なうことで、結合される構造(単量体とは限らない)と結合する構造の1構造対1構造のドッキングとして扱い、CPSF複合体構造の作成を行った。このような手続きを採用することで、既存のドッキングツールを利用して三量体以上の構造を作成することができる。

ドッキングによる複合体構造予測は、大まかに2つのステップで行われる(図2-1)。1つ目のステップはドッキングと呼ばれ、二量体のタンパク質が取りうる結合構造(候補構造)を複数個作り出す。ドッキングツールは前述したZDOCK (Chen and Weng 2002)、PatchDock(Schneidman-Duhovny et al. 2005)、ClusPro(Comeau et al. 2004)、FireDock(Andrusier et al. 2007)、GRAMM(Tovchigrechko and Vakser 2006)、RosettaDOCK(Lyskov and Gray 2008)などが挙げられる。この行程では候補構造の生成以外に、コンパラティブ・モデリングの時と同様、原子間やアミノ酸残基間のエネルギーを算出し、エネルギーの高いものは不安定な構造として排除する。ZDOCKでは、実際のタンパク質構造に近い構造を生成するためには候補構造が2000体あれば十分としている(Chen et al. 2003)。これら多数の候補構造をデコイと呼ぶ。一般にこの一つ目のステップである候補構造を多数作成する過程は、上記のどのプログラムもほぼ同じである。2つ目のステップは複数ある候補構造から尤もらしい立体構造を絞り込む行程である。ドッキングツールでは複数個(数千個以上)の複合体候補構造を作成するため、その中から尤もらしい構造を様々な方法で選出することで、予測構造を得る。このときの選出方法は、アミノ酸組成に基づいて結合残基を予測するもの、各サブユニットの立体構造の相補性に基づいて結合部位を予測するものなど様々な手法を用いたソフトウェアがあるが、上記のソフトウェアでは、共通してそれぞれの予測手法に基づいたスコアを各候補構造に振り、ランキングすることで最も高い候補構造を選出するという手続きを取っている。スコアリングの方法は、利用している既知構造のデータベー

スや重要視する結合部位の特性によって各プログラムごとに異なり、本研究ではタンパク質表面のアミノ酸組成や相補性によるスコアリングを行った。スコアリングに並行して用いられるデコイの絞り込み方法として、予測構造時のエネルギーが低く、物理化学的に安定な構造を選出する手法がある(Chen and Weng 2002)。しかし、実際の複合体構造に最も近い構造をもつ予測複合体立体構造の候補構造が、ツールにより発生させた数千個のデコイのうち、必ずしもエネルギーが最小になるわけではない(Lorenzen and Zhang 2007, Callebaut et al. 1999)。同様にZDOCKで作成した候補構造からエネルギー値の低い順に限られた数の候補構造を選出しても、様々な結合様式の複合体が見られる(Lorenzen and Zhang 2007)。このような予測結果から尤もらしい1つの構造を選出するのは非常に困難である。

そこで本研究では、タンパク質表面のアミノ酸組成や相補性によるスコアリングに加え、候補構造のクラスタリングを行い、どのような結合様式の構造の候補構造が尤もらしい形状のデコイとして予測されやすいかを明らかにすることで、実際の構造に近い候補構造を選出する方法を作成した。

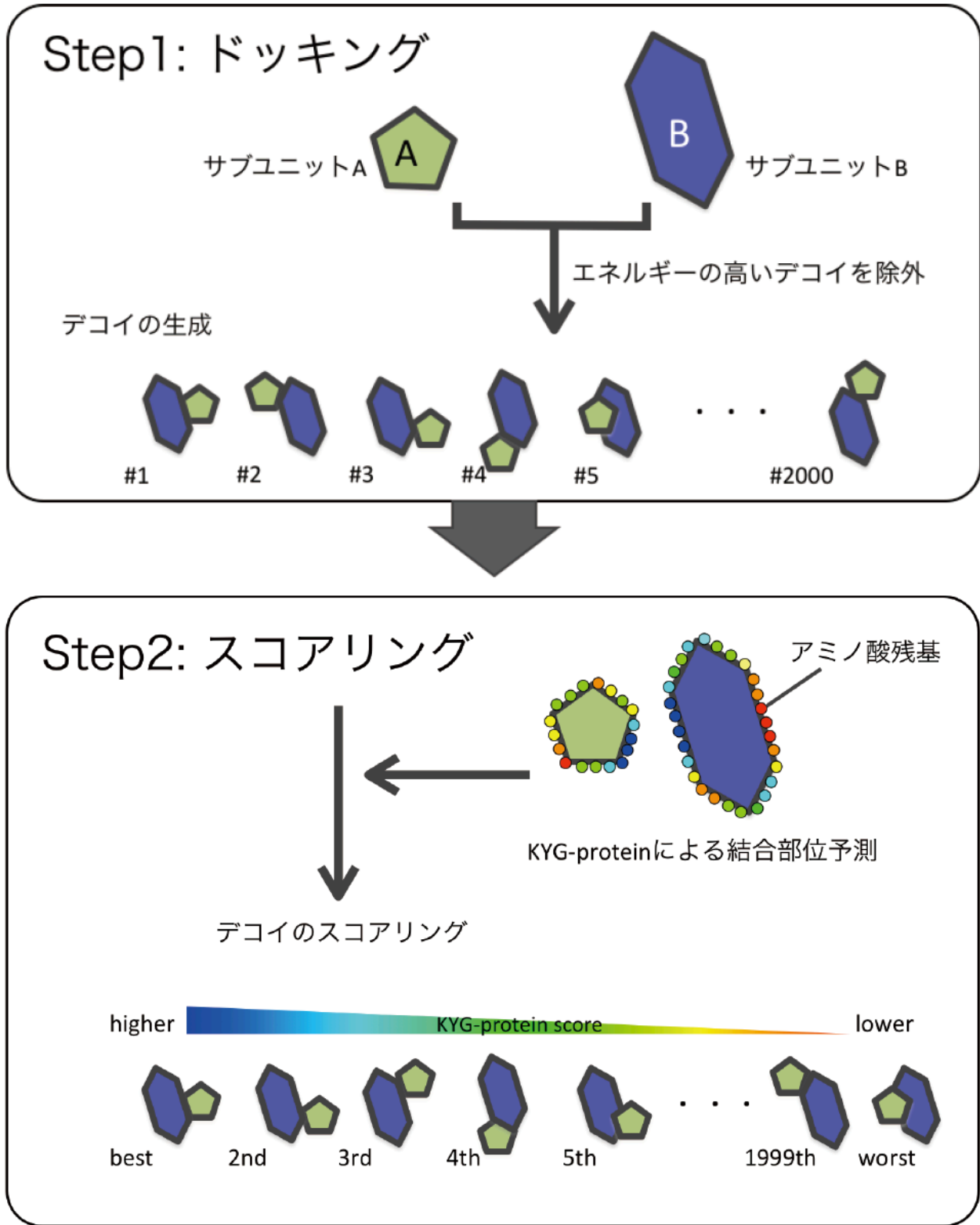


図2-1. 一般的な複合体構造予測のステップ

ステップ1の剛体ドッキングは多くの複合体構造予測ツールに共通する行程である。ステップ2は複数ある候補構造から尤もらしい立体構造を絞り込む行程であり、複合体構造予測ツールごとに特色が異なる。ここでは本研究で用いたKYG-proteinによるスコアリング方法を挙げている。この行程はスコアリングの元となる結合部位の特徴やその重み付けが各プログラムで異なるが、一般的に各候補構造(デコイ)にスコアを振った後、スコアに基づいてデコイを順位付けする。

方法

1. ヘテロ三量体の既知立体構造の取得

2つ以上のタンパク質と同時に結合するタンパク質分子の結合部位の特性について調べるため、ヘテロ三量体の既知構造の検索を行なった。立体構造データベースであるProtein Data Bank(PDB)では機能的な複合体構成に基づいて構造を検索することができない。そこでYura *et al.* (2009) の機能的な複合体構成を持つ既知複合体構造のリスト (<http://cib.cf.ocha.ac.jp/DACISIS/>)より、三量体構造をとる複合体のPDBのIDを取得した。またこのリストにはヘテロ複合体かホモ複合体かの記載がなかったため、全ての三量体構造についてPDBに記載されたサブユニットのアノテーションを確認した。

2. 既存二量体予測ツールの検討

PDBにより得た既知ヘテロ三量体構造を分割し、一量体の立体構造データを作成した。一量体化した三つの立体構造データのうち実際に結合している2つを選択し、既存ツールであるZDOCKを用いて候補構造(デコイ)を2000体作成した。

こうして作成した2000体のデコイは、KYG-protein(Kato *et al.* 2014)およびZDOCKを用いて、各ツールの算出するスコアより順位付けした。ZDOCKはデコイを大量生成すると同時に、2つの分子どうしのタンパク質表面の凹凸や複合体構造の安定性を考慮して2量体構造を評価する。KYG-proteinは各アミノ酸残基がどれくらいタンパク質間結合部位に現れやすいかを数値化した値を用いて、結合部位におけるアミノ酸残基の構成からその結合部位を評価するツールである。ZDOCKとKYG-proteinの予測結果を比較するため、まず全てのデコイについて、デコイ上で結合残基になっているアミノ酸残基が「デコイを作成したサブユニットどうしの実際の結合残基」、「デコイを作成したサブユニットどうしの

結合部位ではないが、他の分子との結合部位となる結合残基」、 「結合残基を構成しないアミノ酸残基」の3つに分類した。その上で、デコイ上Hashimoto 結合部位に含まれる実際の結合残基数ごとに、デコイ数を計測した。これによって、どちらの既存ツールが正解の構造に近いデコイを選出しているかを評価した。

3. デコイ構造のクラスタリング

1つのタンパク質上に複数のタンパク質結合部位をもっている場合、タンパク質結合部位を構成するアミノ酸残基は一般に似通った性質をもつと考えられているため、結合部位を構成するアミノ酸残基から特定のサブユニットが結合する部位を選択することは難しい。そこで本研究では、結合するサブユニットどうしのタンパク質分子表面の構造を考慮することで、複数のタンパク質結合部位を有するサブユニットどうしであっても、結合部位を予測できるのではないかと考えた。実際には、複合体の候補構造であるデコイの全体構造を分類し、どのようなクラスターに含まれるデコイが既知構造と類似の構造であるか調べた(図2-2)。

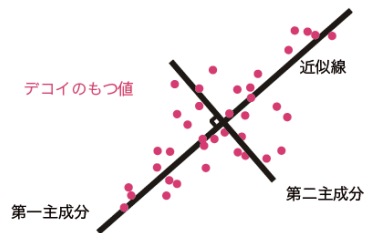
3-1. デコイの構造の数値化

ある2つのタンパク質の立体構造が似ているかどうかを比較するには、タンパク質どうしの原子座標を用いて平均二乗偏差の平方(RMSD)を算出する。RMSDの値は比較する構造のズレを示しており、0に近いほど2つの構造が類似していることを示す。本研究では、デコイの形状はそのデコイのCa原子の座標を用いることで数値として得た。また、あるデコイが他のデコイとどの程度類似しているかは、自分自身を含めたすべてのデコイとのRMSDを計算することで、2000個の数値として得た。

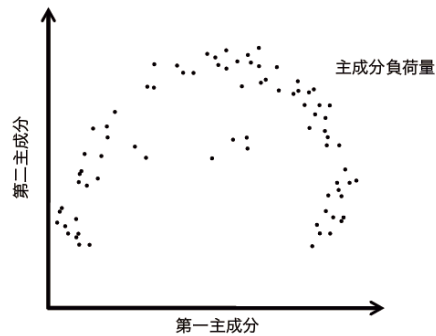
他の構造とのRMSDを計算することでデコイの構造を数値化



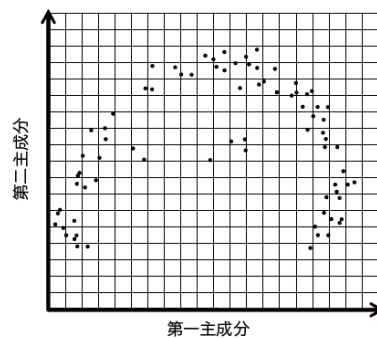
主成分解析



「デコイのもつ値」を最も分離できる軸で、二次元に投影



デコイの結合パターンによるクラスタリング



クラスターに含まれるデコイ数をヒートマップ様に表示

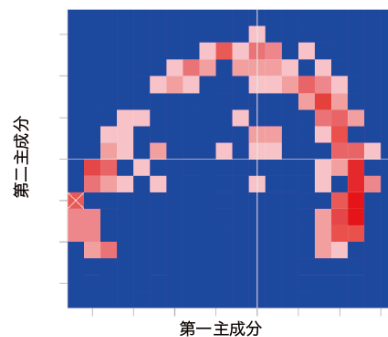


図2-2. 本研究で行ったクラスタリングの手順

本研究で行ったクラスタリング方法の手順を概念図として示す。ひとつの構造を表す値は、2000個のデコイを総当たりで重ね合わせて求めたRMSD値の配列であるため、2000個の座標をもつ1点として扱い、主成分解析を行った。

この計算は全てのデコイに対して行い、既知構造が不明な場合は2000×2000、既知構造が明らかな場合はその構造を含めた2001×2001の数値データを作成し、このマトリクスデータをクラスタリングに用いた。

3-2. クラスタリングとその表示方法

得られた2001個のデータセットは、統計解析言語Rのprcomp関数を用いて主成分解析によるクラスタリングを行った。ヘテロ三量体のすべてのクラスタリングにおいて、第一主成分と第二主成分で累積寄与率の70%以上をカバーしており、それぞれの固有値は1.0を超えていた。これによって元のデータセット以上の情報量を有することが確認できたので、この2つの主成分をXY座標にして二次元プロットを作成し、各クラスターをそのクラスター含まれるデコイ数に応じて色分けしてヒートマップ様のグラフで結果を示した。

4. クラスタリングスコアを用いた既存の複合体予測手法の改良

4-1. 結合パターンにより分類した各クラスターのスコアリング手法

3で作成したヒートマップを用い、複合体の立体構造が既に分かっている分子の複合体立体構造予測を行った。その結果、デコイの数が少ないクラスターに既知構造が含まれる場合、この既知構造は自由度の少ない結合、つまり凹凸のある部位で結合を行なっているということが分かった。また、このような結合部位は結合部位予測ツールで予測しやすく、スコアが高くなりやすい。同じクラスターに含まれるデコイの数が少ないため、クラスタースコアも、他のクラスターに比べて高く算出されることが分かった。一方、既知構造がデコイ数の多いクラスターに含まれる場合、この既知構造の結合部位は結合部位になりやすい特徴をもつ残基で構成され、かつ平面的な結合部位であることがわかった。この

ようなクラスターに含まれる既知構造は高いスコアを有しているが、同じクラスター中に多くのデコイが存在するため、スコアの低いデコイも複数含まれ、クラスター全体としてはスコアの平均が低くなることがわかった。

そこで上記のどちらの場合であっても既知構造が含まれるクラスターのスコアが高くなるよう、クラスターに存在するデコイの数(密度)と各デコイがもつスコアから以下に示す式によりクラスタースコア(Gscore)を求めることとした。

$$G_{score} = \frac{1}{n} \sum_1^n (score) - T \left(\frac{1}{n_{all}} \right) \ln \left(\frac{n}{n_{all}} \right) \quad \dots (1)$$

n はクラスターに含まれるデコイの個数、 n_{all} はZDOCKによって生成した全デコイ数、 $score$ はそれぞれのデコイが持つKYG-proteinまたはZDOCKによって得たスコアを指す。

(1)における定数 T は、他のクラスターに比べ既知構造が含まれるクラスターのGscoreが高くなるように設定した。クラスタリング結果から既知構造に最も近い構造を誤りなく選出するためには、理想的には既知構造に最も近い構造が含まれるクラスターのみGscoreが高く、他の全てのクラスターのGscoreが低くならなければならない。そこで(1)の式の T 値を変化させたときのGscoreの推移を計算し、その推移から理想的なクラスタリングを作成できるような T 値を求めることとした。

既知構造が含まれるクラスターのGscoreが最も高い値になるとき、 T の値は0付近の値をとることが分かっていたため、既知構造が含まれるクラスター、最も多くのデコイを含むクラスター、KYG-proteinおよびZDOCKのスコアが最も高いクラスター、デコイを含むクラスター全体について T を-1.0から1.0まで変化させGscoreの値をプロットした。クラスター全体のGscoreはクラスターすべてのGscoreを平均して求めた。

4-2. クラスタスコア算出に使用する複合体のスコアリングツールの検討

クラスタスコアの算出にあたり、KYG-protein、ZDOCKのどちらの手法によりデコイごとのスコアを算出すれば、より尤もらしい複合体構造を選出できるかを判断することにした。KYG-protein、ZDOCKのそれぞれでスコアのみで複合体構造を予測した場合と、クラスタスコアにより複合体構造を予測した場合で、既知構造の順位を比較した。

5. 候補構造数の決定

候補構造の生成は、ZDOCK(Chen and Weng 2002, Mintseris et al. 2007)によって行った。ZDOCKは、主にタンパク質表面の形状の相補性をもとに剛体ドッキングを行うプログラムである。ChenらはZDOCKでの複合体構造予測において、49サンプル中44サンプルが2000個のデコイの生成で実際の複合体構造に十分似た立体構造を生成できるとしていた(Chen and Weng 2002)。しかし、49サンプルはすべて二量体であったため、三量体以上の構造を予測する場合と比較して、狭い複合体のタンパク質表面積を用いた複合体予測にしか対応していない可能性がある。そこで、分子表面積が広い複合体でも2000個のデコイで十分に複合体構造予測が可能であるか、ひとつのサブユニットのサイズが1000残基を超える比較的大きな分子を用いて、2000個のデコイ生成で十分であることを検討した。PDBから収集した互いにホモログでないヘテロ三量体の立体構造を検索し、9つのヘテロ三量体の既知構造を得た。そのうち、1000残基を超えるサブユニットを含む比較的大きな複合体は3つあった。ZDOCKは2つの立体構造を入力として複合体構造を予測するプログラムなので、三量体を各サブユニットに分割し、1000残基を超えるサブユニットと結合しているサブユニット、もしくは結合しているサブユニットのうち大きい方のサブユニットとの結

合構造予測において、2000個のデコイ作成の場合と4000個のデコイ作成の場合の2通りでクラスタリング図を作成し、クラスタリング図のパターンを比較した。

6. 結合パターンによるクラスタリングを取り入れた構造予測手法の精度

本研究で開発したクラスタリングによる複合体構造予測手法の精度を検討するため、以下のような比較を行った。各クラスターから、最もエネルギーの低いデコイを各クラスターを代表するデコイとして取り出し、それらをランダムに並べたものとクラスターのスコア順に並んだものと、それぞれのデコイの結合部位になっているアミノ酸残基数が既知構造のアミノ酸残基数とが、どのくらい一致しているかを計測して比較した。

結果および考察

1. 既存二量体予測ツールの検討

複合体予測ツールのテストデータとしたヘテロ三量体構造は、ヘテロ三量体の立体構造は2011年5月時点で11体得られた。このうち2体はオーソログであったため、全9体のヘテロ三量体を解析に使用した(表2-1)。

ZDOCKおよびKYG-proteinの2つの二量体予測ツールにおいて、スコアによる順位付けの上位200位ごとにデコイの予測結合部位を構成するアミノ酸残基中、実際の構造で結合部位を構成しているアミノ酸残基の数を比較したグラフを図2-3に示す。

サブユニットどうしの結合部位における実際の結合残基を計測した結果では、ZDOCKの方がKYGより高い順位に実際の立体構造に近いデコイがみられたが、ZDOCKはエネルギー的に安定な構造を高く評価する傾向が強く、多くのアミノ酸残基で結合している構造をより高く評価するため、実際の結合部位を担っているアミノ酸残基が少ない場合は結合部位を予測するのは困難であることがわかった。また、KYGはZDOCKと比較し予測精度が一見低く見えるものの、これはKYG-proteinの予測の特性が良くない方に現れた結果であることが分かった。KYG-proteinでは、デコイを形成するサブユニットそれぞれの結合部位を個別に評価しており、サブユニットの結合部位どうしの相性などはみていない。そのため、三量体以上の複合体を組むサブユニットには多くの結合部位が存在し、どのサブユニットの結合部位であるかという区別はできていないことが分かった。しかしKYG-proteinを用いた場合、結合部位の予測という点では予測精度は高かったため、2つ以上の相手分子があるタンパク質については、複数の結合部位から特定のサブユニットが結合する結合部位を予測する指標が必要であることが分かった。

表2-1. 複合体予測ツールのテストデータとしたヘテロ三量体構造

PDB ID	UniProt ID	分子名	生物種
1EF2	URE1_ENTAE URE2_ENTAE URE3_ENTAE	尿素分解酵素	<i>Enterobacter aerogenes</i>
1IE7	URE3_SPOPA URE2_SPOPA URE1_SPOPA	尿素分解酵素	<i>Sporosarcina pasteurii</i> (<i>Bacillus pasteurii</i>)
1KQF	FDNG_ECOLI FDNH_ECOLI FDNI_ECOLI	ギ酸デヒドロゲナーゼ	<i>Escherichia coli</i> (strain K12)
1OGV	RCEH_RHOSH RCEL_RHOSH RCEM_RHOSH	光合成反応中心タンパク質複合体	<i>Rhodobacter sphaeroides</i> (<i>Rhodopseudomonas sphaeroides</i>)
1Q16	NARG_ECOLI NARH_ECOLI NARI_ECOLI	硝酸レダクターゼ	<i>Escherichia coli</i> (strain K12)
1Z0M	AAKB1_RAT	AMP活性化タンパク質キナーゼ	<i>Rattus norvegicus</i> (Rat)
2A5Y	CED9_CAEEL CED4_CAEEL	アポトーシス調節因子 ced-9/ced-4複合体	<i>Caenorhabditis elegans</i>
2DF4	GATA_STAAW	tRNA依存アミド基(Asn/Gln)転移酵素	<i>Staphylococcus aureus</i> (strain MW2)
2IVF	CLRA_IDEDE	エチルベンゼンヒデヒドロゲナーゼ	<i>Ideonella dechloratans</i>
2V92	AAPK1_RAT AAPK2_RAT AAKG1_RAT	AMP活性化タンパク質キナーゼ	<i>Rattus norvegicus</i> (Rat)
3H0R	GATA_AQUAE GATB_AQUAE GATC_AQUAE	tRNA依存アミド基(Asn/Gln)転移酵素	<i>Aquifex aeolicus</i> (strain VF5)

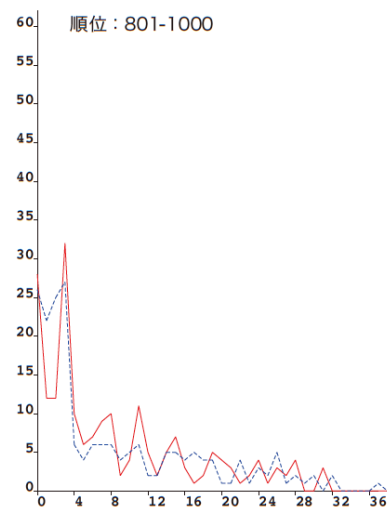
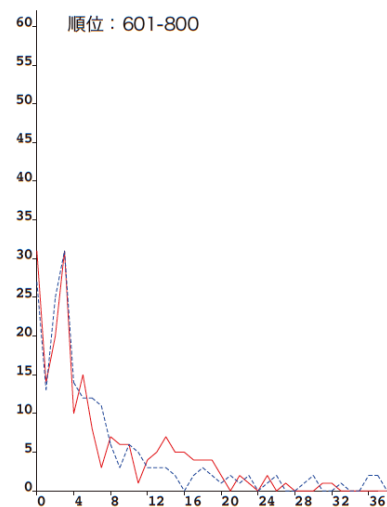
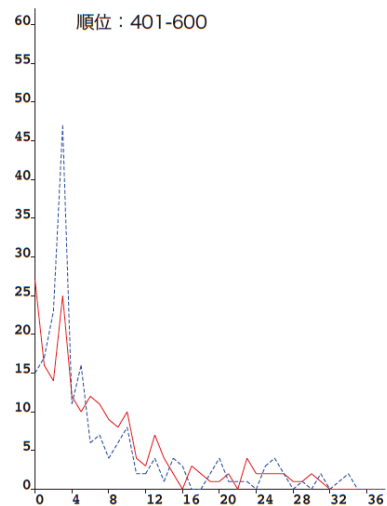
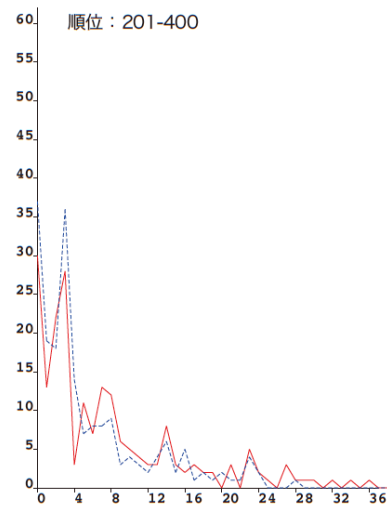
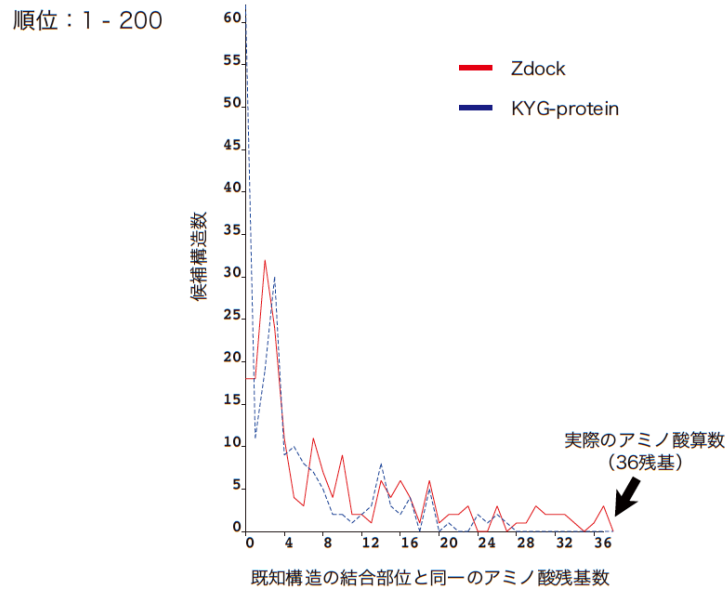


図2-3. 既存二量体予測ツールの検討

デコイの予測結合部位を構成するアミノ酸残基中、実際の構造で結合部位を構成しているアミノ酸残基の数を算出し、200位ごとにグラフ化し1000位まで表示した。横軸が実際の結合残基数、縦軸がその残基数で結合しているデコイの数、青がKYG-protein、赤がZDOCKの値を示す。

2. デコイ構造のクラスタリング

結合パターンによってデコイをクラスタリングした結果、実際の立体構造が比較的広い面を互いに使用した平面的な結合をしている場合は、クラスタリングを行なうだけでも実際の立体構造に近い構造を予測することができることが分かった(図2-4. a,b)。これは、ZDOCKがデコイを作成するときに、結合面の広いエネルギーが低い構造から2000個作成するプログラムであることが大きく寄与している。クラスタリング結果を示したヒートマップ様のグラフでは、X軸、Y軸はZDOCKに入力として与えた2つの構造(仮にサブユニットA、サブユニットBとする)のうち、サブユニットAの立体構造のタンパク質表面上の位置を示しており、ヒートマップの原点付近はタンパク質内部、もしくはタンパク質表面の凹みでの結合を示す。このため、タンパク質表面の凹凸が大きい部位、もしくはタンパク質が折れ曲がった形状でない限り、原点付近のクラスターにはデコイが含まれない。また、1つのクラスターの中にはサブユニットAの同一結合部位におけるサブユニットBの立体構造の結合パターンのバリエーションが含まれる。一般的にはタンパク質間結合部位は平面を形成していると考えられているので(Chakrabarti and Janin 2002)、予測したい複合体構造の結合部位が自由度の高い結合部位であれば、このヒートマップの結果から、少なくともサブユニットAの構造の結合部位がどこにあるか予測をできることが分かった。しかし、サブユニットどうしが自由度の少ない結合様式、つまり凹凸が多くはまり込むような結合を行っている場合は、その結合部位における結合パターンのバリエーションが限られてしまうため、そもそもデコイの生成数が限られてしまう。このことから凹凸の多い結合部位で結合している複合体構造の予測は、このヒートマップでは難しいことが分かった(図3 c,d)。そこで、そのような場合にも結合部位を予測できるようにするため、形状によるクラスタリングに加え、既存二量体予測ツールのスコアを用いることにした。

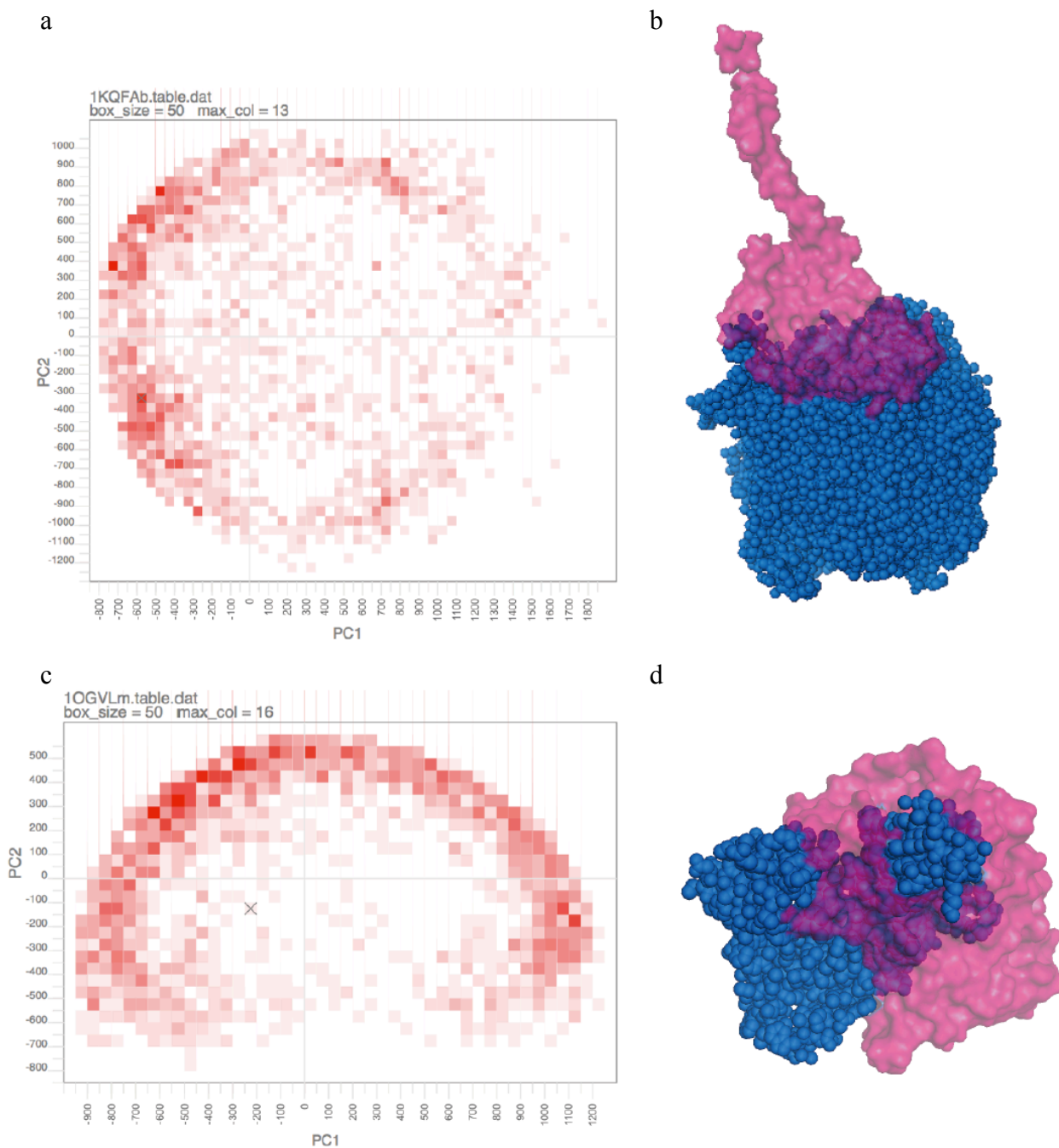


図2-4. 複合体立体構造の類似性によるデコイのクラスタリング

a, cに示したヒートマップはデコイの構造を主成分分析を用いたクラスタリングにより分類した結果であり、第二主成分までに70%以上の固有値を含んでいたため、第一主成分をX軸、第一主成分をY軸として構造をクラスタリングし、1マスで示されるクラスターに含まれるデコイの数を赤色の濃淡で表示している。最もデコイの数が多いクラスターを赤で示しており、デコイの数が少ないほど薄い赤色で表示している。また白色のクラスターはそのクラスターに含まれる構造が存在しないことを示す。“X”は既知構造が属するクラスターを示している。b, dに示す二量体構造はそれぞれa, cのクラスタリングに用いたデコイの実際の立体構造。比較的平らな結合部位を持つ複合体の既知構造(b)はデコイの多いクラスターに分類され(a)、自由度の少ない結合部位を持つ複合体の既知構造(d)はデコイの少ないクラスターに分類される(c)。

3. 式(1)によるクラスタースコア(Gscore)の算出

3-1 式(1)によるGscoreの最適化

Gscoreを推移させたときのTの値を求めた結果、既知構造が含まれるクラスターのGscoreが最も高い値をとる時、Tの値は極小値をとることが分かった。また、KYG-proteinを用いたヒートマップの結果では、すべてのクラスターでTの値は既知構造を含むクラスターと同じ値で極小値をとることが分かった。そこでクラスター全体のGscoreの平均を求め、そこからT値を得て、予測ごとにT値を変更してGscoreを再計算するプログラムを作成した。しかし、ZDOCKではGscoreの値は直線的に変化したので、適切なT値を算出することができなかった(図2-5)。

3-2 KYG-proteinおよびZDOCKの比較

結果を図2-6に示した。スコアのみでの比較においては、KYG-proteinの得意とする残基数の少なく凹凸の大きな結合の場合以外は、ZDOCKの結果が良いことが分かった(図2-7. a)。しかし結合面積が小さい構造や結合部位が平面ではないような構造では、ZDOCKの検出頻度は悪かった。一方、式(1)を用いた処理を行うと、KYGの検出頻度はZDOCKと同程度になり、ZDOCKでは検出頻度に大きな変化は見られなかった(図2-7. b)。このことから、式(1)を用いたスコアリングとの相性がいいのはKYG-proteinによるスコア付けの方であると考えられたため、CPSF複合体の組み上げの際にはKYG-proteinを用いることとした。

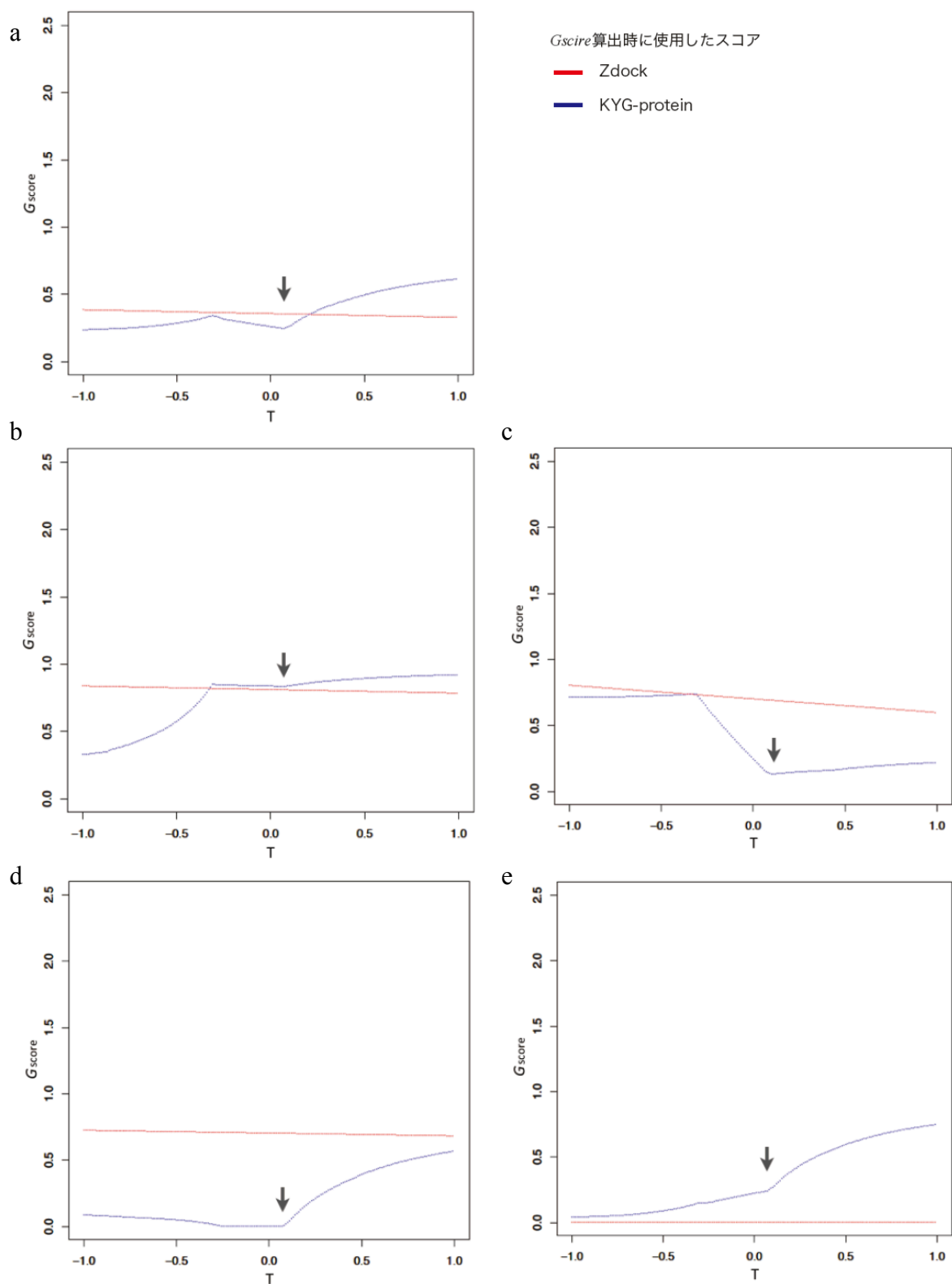


図2-5. T値を変化させたときのGscoreの値の推移

赤線はスコアにZDOCKを利用したヒートマップのGscoreの値を、青線はKYG-proteinを利用したときのスコアの推移を示す。Gscoreの推移は既知構造を含むクラスターのGscore(a)、デコイを含むクラスター全てのGscoreの平均値(b)、最も多くのデコイを含むクラスターのGscore(c)、クラスター内に含まれるデコイのKYG-proteinのスコアの平均が最も高いクラスター(d)、クラスター内に含まれるデコイのZDOCKのスコアの平均が最も高いクラスター(e)の5つの条件でクラスターを選択してグラフ化した。全てのグラフで同じT値が極小値となっているので、この値を用いてヒートマップを作成した。

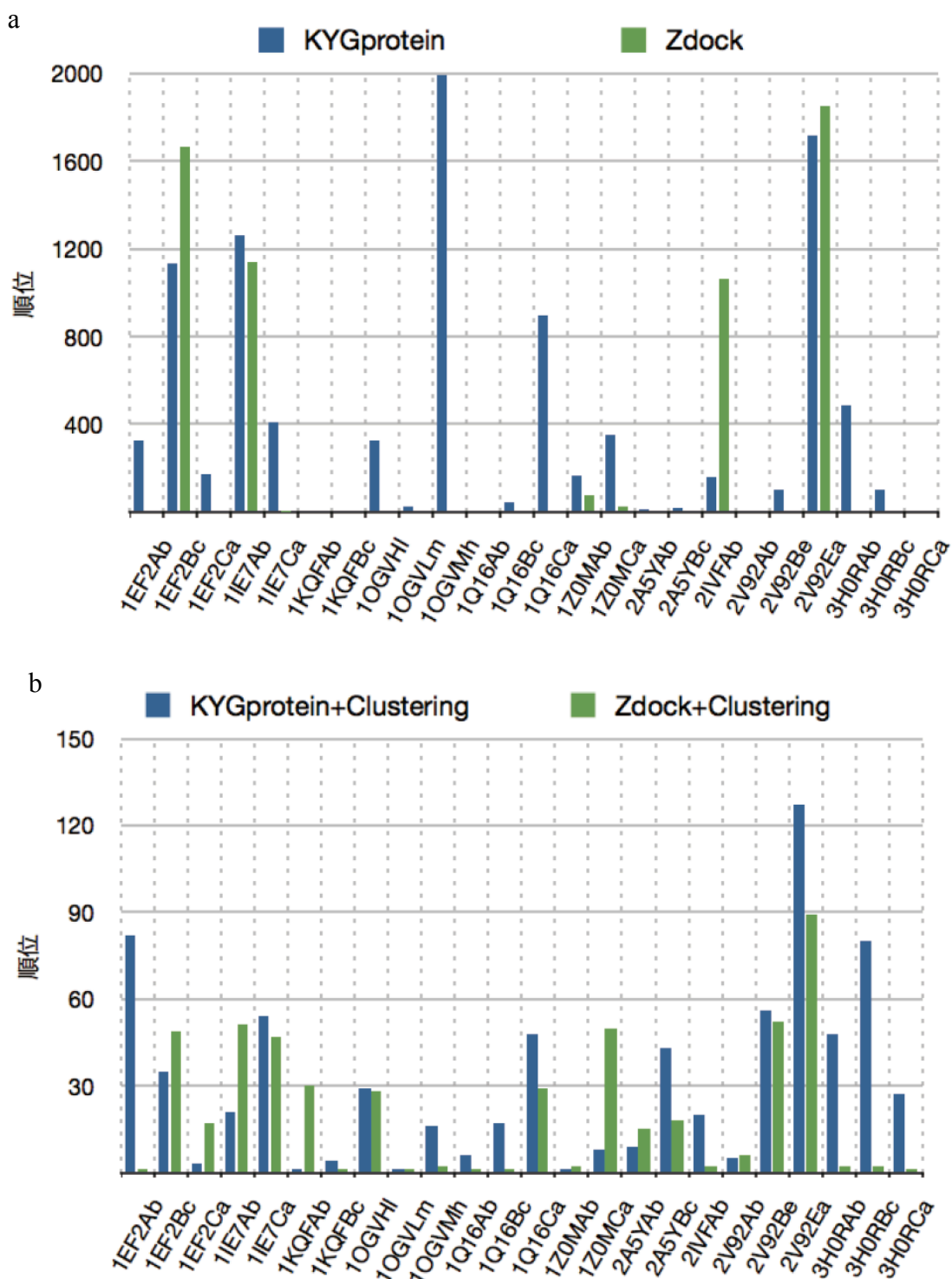


図2-6. KYG-proteinとZDOCKの予測精度の比較

既知構造に最も近い構造(a)および既知構造が含まれるクラスターの順位を比較した。横軸には、複合体予測を行った構造名を示した。aではZDOCKによる予測スコア(黄)とKYG-protein(赤)による予測スコアを示している。bはZDOCKによる予測スコア(緑)とKYG-protein(青)による予測スコアを式(1)に用いてクラスター自体の順位を出したもの。bではクラスターの数が構造によって異なるので、順位を割合で表示している。

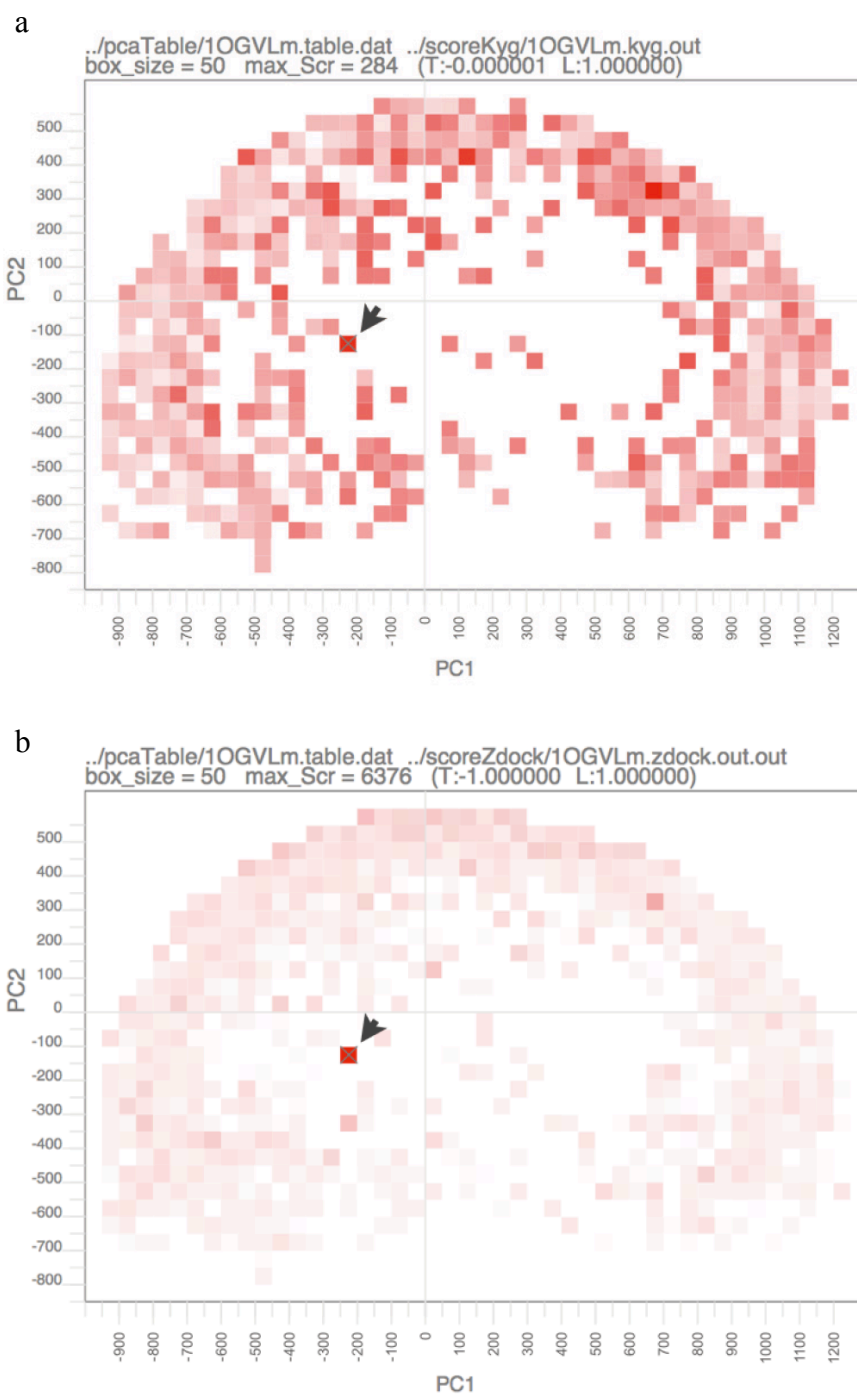


図2-7. 式(1)を利用したクラスタリング

aはKYG-protein、bはZDOCKのスコアを使用した結果。どちらも最も値の高いクラスターが既知構造を含むクラスターであった。図2-4.c, dで示したような結合を行う複合体構造の場合でも予測ツールのスコアを用いることで予測が可能になった。

4. 候補構造数の決定

分子量が100を超えるサブユニットを少なくとも1つ含む既知構造を用い、デコイ2000体(図2-8. a)、およびデコイ4000体(図2-8. b)でクラスタリングを行ったところ、クラスタの分布および各クラスタに含まれるデコイ数は、総デコイ数に関わらず同様の結果を示していた。よってCPSF複合体組み上げの際も、デコイを2000体で予測を行うこととした。

5. 式(1)を用いた構造予測手法の精度

図2-9に、デコイをランダムに並べた場合と式(1)によるスコア方法でトップ20%のデコイに既知構造に最も近い構造が含まれる割合を示す。この結果、本研究で開発した予測手法は、クラスタリング後にランダムに順位を出したものと比較して、ZDOCK、KYG-proteinとも複合体52個中40個で精度が上昇した。予測精度は最大でも5倍であったが、これはデコイを発生させる際に結合残基が既知構造と75~80%一致する構造が5個程度であることが影響しており、予測精度が極めて高い場合でも5倍以上にはならないためである。

式(1)に代入したKYG-proteinとZDOCKのスコアの比較においては、同じデコイの集団に対して5倍以上の精度を示した構造は、ZDOCKでは52構造中11個、KYG-proteinでは52構造中13であり、2つのツールを用いた結果にほとんど差はなかった。しかし、ZDOCKによるスコアリングは複合体構造全体の安定性によるものであり、クラスタリングも複合体の全体構造をもとにその類似性を分類するものである。一方KYG-proteinは、結合部位に現れるアミノ酸残基種の統計量に基づく予測手法である。種類の異なる特徴を用いて複合体構造を予測する方が、様々な結合の複合体に対応できる予測手法になると考え、本研究では、KYG-proteinのスコアを用いて式(1)に示すクラスタースコアを算出することとした。

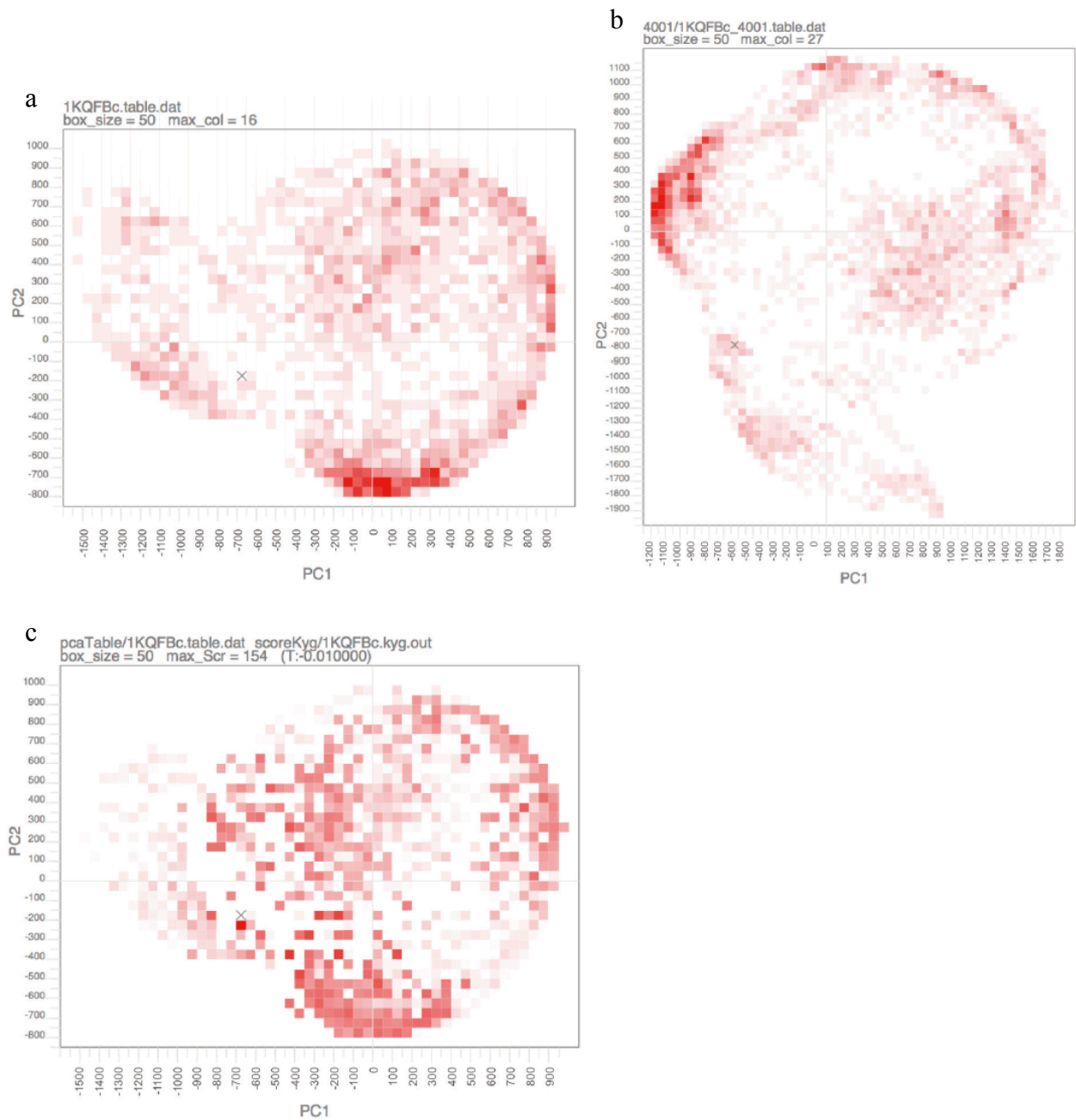


図2-8. 本予測に適正なデコイ数の決定

aはデコイを2000個発生させてクラスタリングを行ったもの、bは倍の4000個のデコイを発生させてクラスタリングした結果。a, bのヒートマップでは各クラスターの色の濃淡はクラスターに占めるデコイの数を示しており、赤いほど、多くのデコイを含むクラスターを示す。aとbを比較すると第一主成分と第二主成分が入れ替わっているが、ヒートマップにおけるクラスターの描く形状および、各クラスターに含まれるデコイ数の分布は同じパターンを描いており、デコイ4000個の場合とデコイ2000個の場合で同様の結果が得られている。cはKYGスコアを利用した式(1)を利用したヒートマップである。

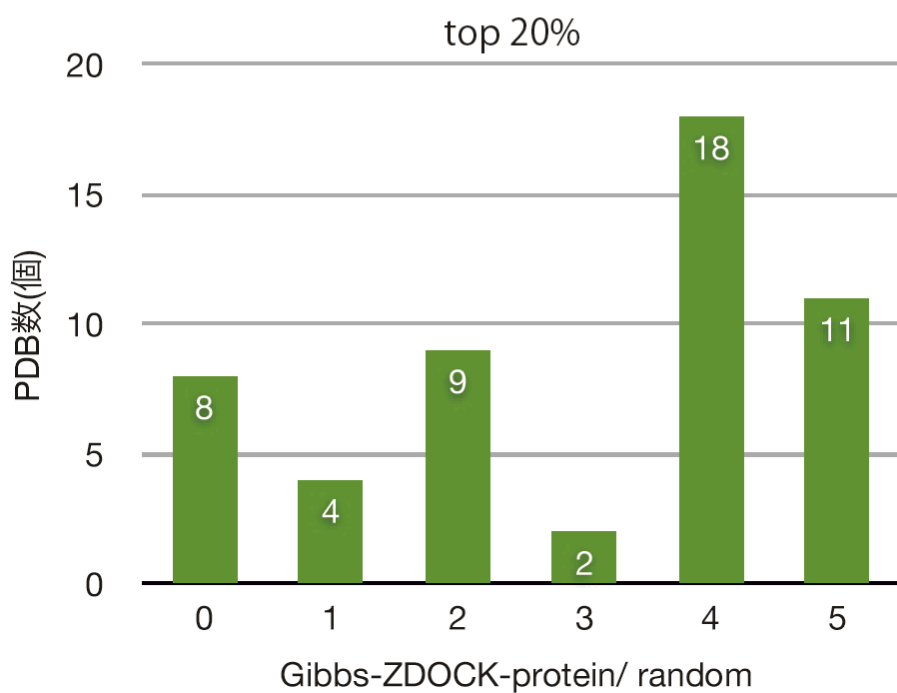
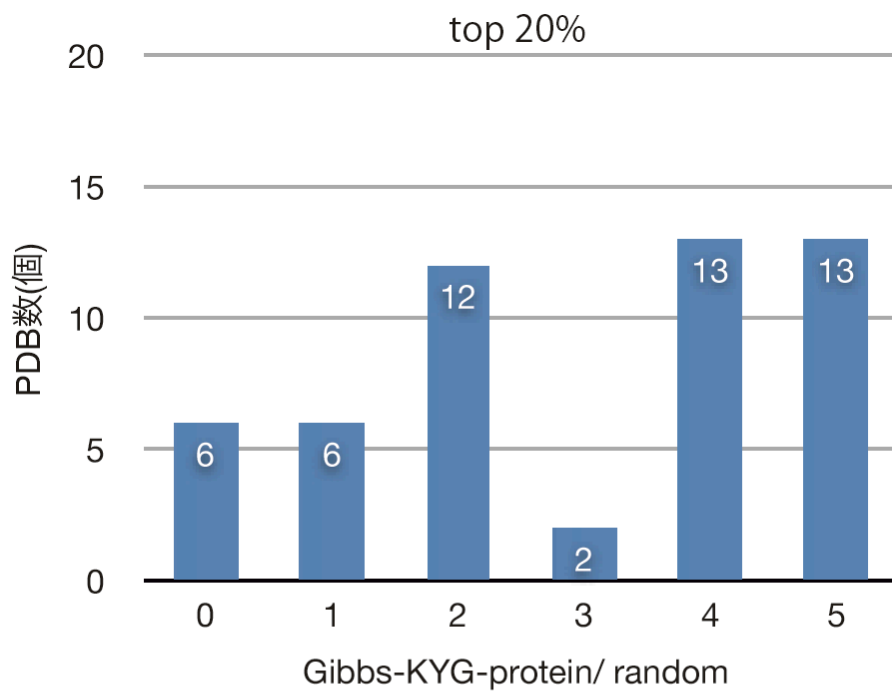


図2-8. KYG-proteinおよびZDOCKにおけるクラスタリングと予測ツールによるスコアリングによる複合体構造の予測手法の精度比較
 デコイをランダムに並べた場合、および式(1)によるスコア方法でランキングした場合においてトップ20%のデコイに既知構造に最も近い構造が含まれる複合体構造数を比較した。

6. 小括

本章では、3つ以上の分子が結合する複合体構造の予測手法について検討した。まず、デコイ作成に用いたZDOCKと本研究室で開発したKYG-proeinで既知ヘテロ三量構造の構造予測を行い、その精度を比較した。その結果、どちらのツールを用いた場合でも、スコアの高い順にランキングした上位10%のデコイには様々な結合パターンの複合体構造が含まれており、ツールの結果だけから実際の構造に近い形状を持つデコイを選択するのは困難であることが分かった。そこで本研究では、多数あるデコイを形状でクラスタリングし、そのクラスターごとにスコア付けをすることで、尤もらしい形状で結合しているデコイを絞り込むという複合体予測手法の開発を行った。

形状によるクラスタリングを行なった結果、各クラスターのもつ特性には結合部位の位置が大きく影響していることが分かった。デコイが多く含まれるクラスターは、比較的広い平らな面で結合しているパターンのデコイを内包することになる。タンパク質結合部位は一般的に平らである程度の広さが必要であることから、デコイが多く含まれるクラスターを選択するだけでも、既知構造に近い結合パターンのクラスターが選択できる場合があることが分かった。

しかし、凹凸の大きな領域で結合している構造については、この手法では予測が不可能であることがわかった。凹凸の大きな領域で結合している複合体は、自由度の少ない噛み合うような結合をしている構造が多くみられた。そこで、アミノ酸残基のタンパク質への結合のしやすさを求めることによりタンパク質間結合部位を予測するKYG-proteinのスコア、および複合体構造の安定さを示すエネルギーの低さから複合体構造を予測するZDOCKのスコアを用いて、自由度の低い構造でも予測できるように2つのプログラムのスコアをクラスタリングと組み合わせ、両者の結果を比較した。

その結果、KYG-proteinを用いた予測においてスコアのみでは既知構造に最も近い立体構造がランキングの上位50%に入らなかった構造が4構造あったのに対し、クラスタリングと組み合わせたことで、ランキングの上位50%に入らなかった構造は1つに減っていた。ZDOCKではクラスタリングを用いたことで、ランキングが大きく向上した構造は1つだけであった。このことからクラスタリングの手法にはKYG-proteinのスコアリングの方が有効に働くと考え、KYG-proteinのスコアを用いたクラスタリング方法を用い、CPSF複合体の複合体構造を予測することとした。

第三章

複合体におけるサブユニット界面の特徴抽出

背景

前章で複合体立体構造の予測手法の開発を行ったが、その際にKYG-proteinのスコアを用いた。このスコアは、結合部位が存在しやすいアミノ酸残基の統計量を用いて算出される。つまり、一般化された結合部位の特徴を反映させるスコアである。しかし、CPSF複合体は、CPSFサブユニットのパラログが構成する複合体とは、サブユニット構成が異なる。また、CPSFサブユニットのパラログには、ほかのCPSFサブユニットとは全く異なる立体構造の分子と結合することが明らかとなっているものも存在する(Scrima et al. 2008)。一般的には、結合部位を構成するアミノ酸残基は、アミノ酸配列や立体構造に類似性のあるパラログ間では、保存されていることが多く(Livstone et al. 2011)、サブユニット構成も同一であると考えられている。つまり、CPSFサブユニットは、KYG-proteinのスコアでは予測が困難な特徴的な結合部位を有する可能性があった。

パラログなタンパク質は進化の過程で生じた遺伝子重複から得られたと考えられており、このようなタンパク質が生体内で機能を発揮している場合、祖先分子の機能以外の新しい構造や分子機能を獲得している場合がある(Tatusov et al. 1997, Rao et al. 2014)。近年、このような分子は結合する相手分子を変えていたり(Aloy and Russell 2002)、挿入によって新たな結合部位を獲得している(Hashimoto et al. 2010, Nishi et al. 2011)ことが明らかとなってきている。従って、オーソログなタンパク質どうしの解析からは、進化的な背景をもとに祖先分子が有していたと考えられる共通した特徴を得ることができるが、パラログなタンパク質どうしの解析からは、そのパラログにしかみられない機能に関与するような特徴を読みとることができる。つまり、アミノ酸配列や立体構造に類似性のあるパラログであっても、新たな機能を獲得した際にもともともっていた結合部位に変化が生じ、結合

部位が消失したり、他の分子と結合できるようになったりすることが考えられる。そこで本研究では、パラログ間で共通しない結合部位をもつ複合体の立体構造をPDBから収集し、収集したパラログ間で異なる結合部位をもつ複合体のサブユニット界面にどのような特徴があるかを調べた。またこれらの特徴をもとに、CPSFサブユニットが祖先分子とは異なる新たな結合部位を獲得しているかを調べた。

方法

1. パラログ間で共有しない結合残基の特徴抽出

1-1 サブユニット構成を変化させた複合体およびそのパラログの収集

分子系統解析の結果、CPSFサブユニットはそれぞれのパラログと複合体構成が異なっており(図3-1)、それらのパラログとは異なる結合部位を有している可能性が高いことが分かった。このように進化の過程で変化した結合部位は、オーソログによる解析から見いだすことは困難である。そこで、CPSFサブユニットのようにパラログ間で明らかに異なる立体構造をもつ分子と結合する複合体の立体構造を収集し、その結合部位にどのような傾向があるか調べた。本章ではパラログとは異なるサブユニット構成を持つ複合体に着目し、研究を行った。共通祖先配列であるが互いに異なるサブユニット構成をもつ2つ以上のパラログを、パラログセットと呼ぶことにする。

立体構造情報を有するパラログセットを取得するため、PDB(Berman et al. 2003)に立体構造が登録されている全分子のアミノ酸配列に対し、BLASTClust(Altschul et al. 1997)を行った。この時BLASTClustにおいてホモログな配列のクラスターを作成する際の閾値は、配列一致度が25%、配列カバー率が90%とした。ホモログな配列のクラスターに含まれるすべての複合体について、生物種、分子名等のアノテーション、そしてサブユニット構成についての情報をPDBから得た。また、これらの情報がPDB上にない分子については、それらのアミノ酸配列をクエリとしてUniProtの配列データベースuniprot_sprot.datに類似性配列検索を行い、同種の配列から最も類似性の高い配列のアノテーションをその配列のものとして利用した。これらの方法により、進化の過程でサブユニットの構成および生物学的な機能を変化させ、かつ同一祖先由来と考えられるパラログセットを得た。

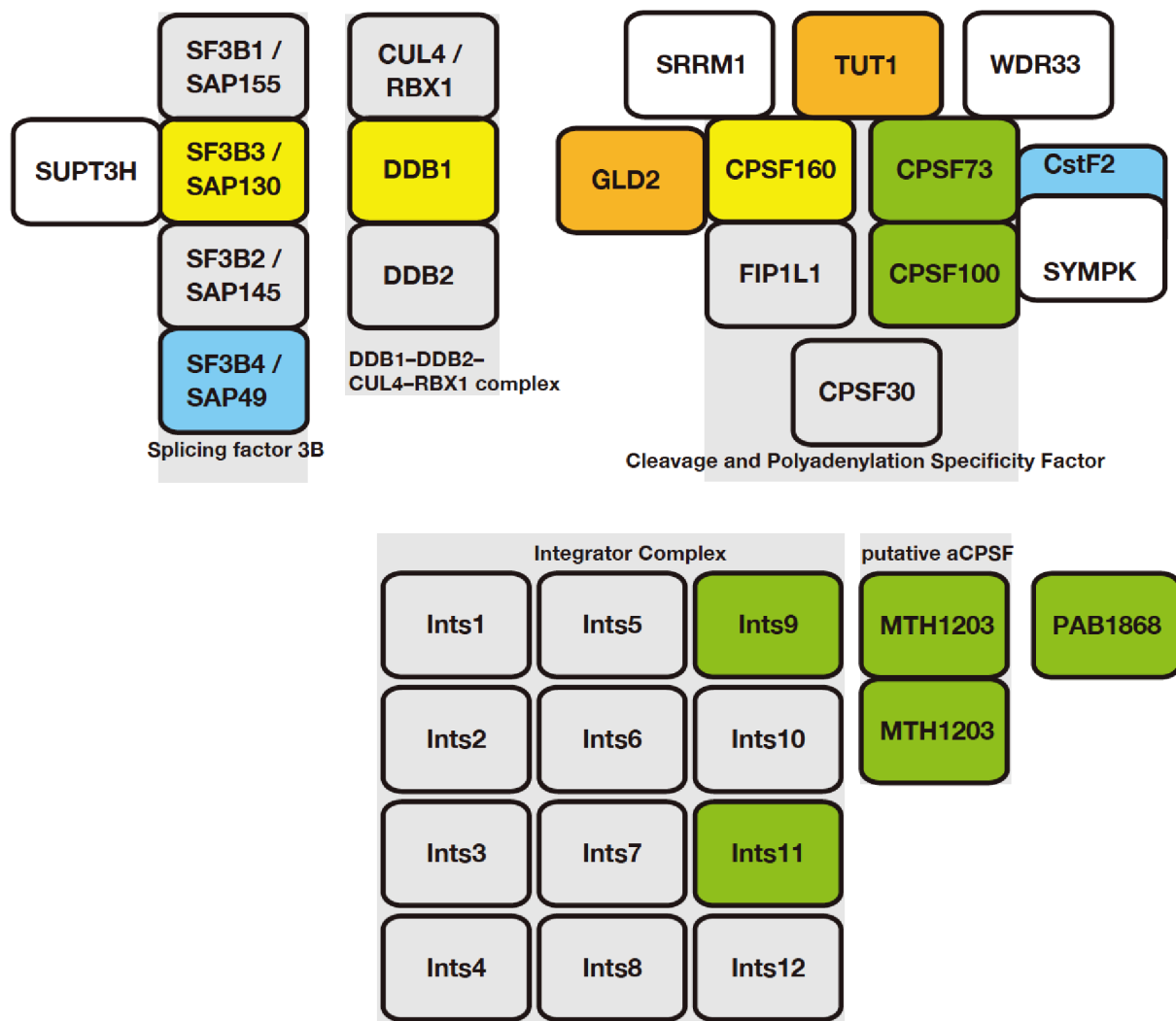


図3-1. CPSFサブユニットとそのパラログのサブユニット構成

CPSFのサブユニット構成と、そのパラログのサブユニット構成を示す。各複合体において、パラログな分子どうしを同じ色で示した。複合体を構成する分子数が異なるだけでなく、複合体を構成するサブユニットも複合体間で異なる。図はAoto and Yura (2015)より転載。

次に得られたパラログセットの各アミノ酸配列をクエリとし、改めてUniProtに対して類似性配列検索を行ない、E-value 10^{-4} 以下を示したアミノ酸配列を取得した。これらの配列から系統樹を作成し、ひとつの系統樹に含まれるパラログ、および各パラログに含まれるアミノ酸配列を特定した。さらに系統樹の枝から特定されるすべてのパラログに立体構造が明らかなアミノ酸配列が含まれているか調べ、ひとつの系統樹に含まれるすべてのパラログに立体構造が明らかなアミノ酸配列を含んでいる場合に限り、それらのパラログセットを解析対象とした。

また、UniProtのアノテーションに記載されている複合体構成とPDBに記載されている複合体構成数が一致しない場合は、これらのサブユニット数を結晶構造時の非対称単位でのサブユニット数、または機能的な複合体のサブユニット数と比較し、一致する複合体構成がなければ解析対象のパラログセットから外した(図3-2)。

1-2. 複合体時の結合部位を構成する残基の定義

結合部位を構成するアミノ酸残基は、単量体および多量体時の溶媒接触面積の差をとることで検出することができる(Kobayashi et al. 2013)。本研究では、単量体および多量体時の溶媒接触度をアミノ酸残基ごとに算出し、その差が0.003以上であった時、そのアミノ酸を結合部位を構成するアミノ酸残基と定義した。また、溶媒接触度の差が0.003以下で、多量体時の溶媒接触度が0.1以上であった場合、そのアミノ酸残基を複合体時の表面を構成するアミノ酸残基と定義した(Go and Miyazawa 1980)。

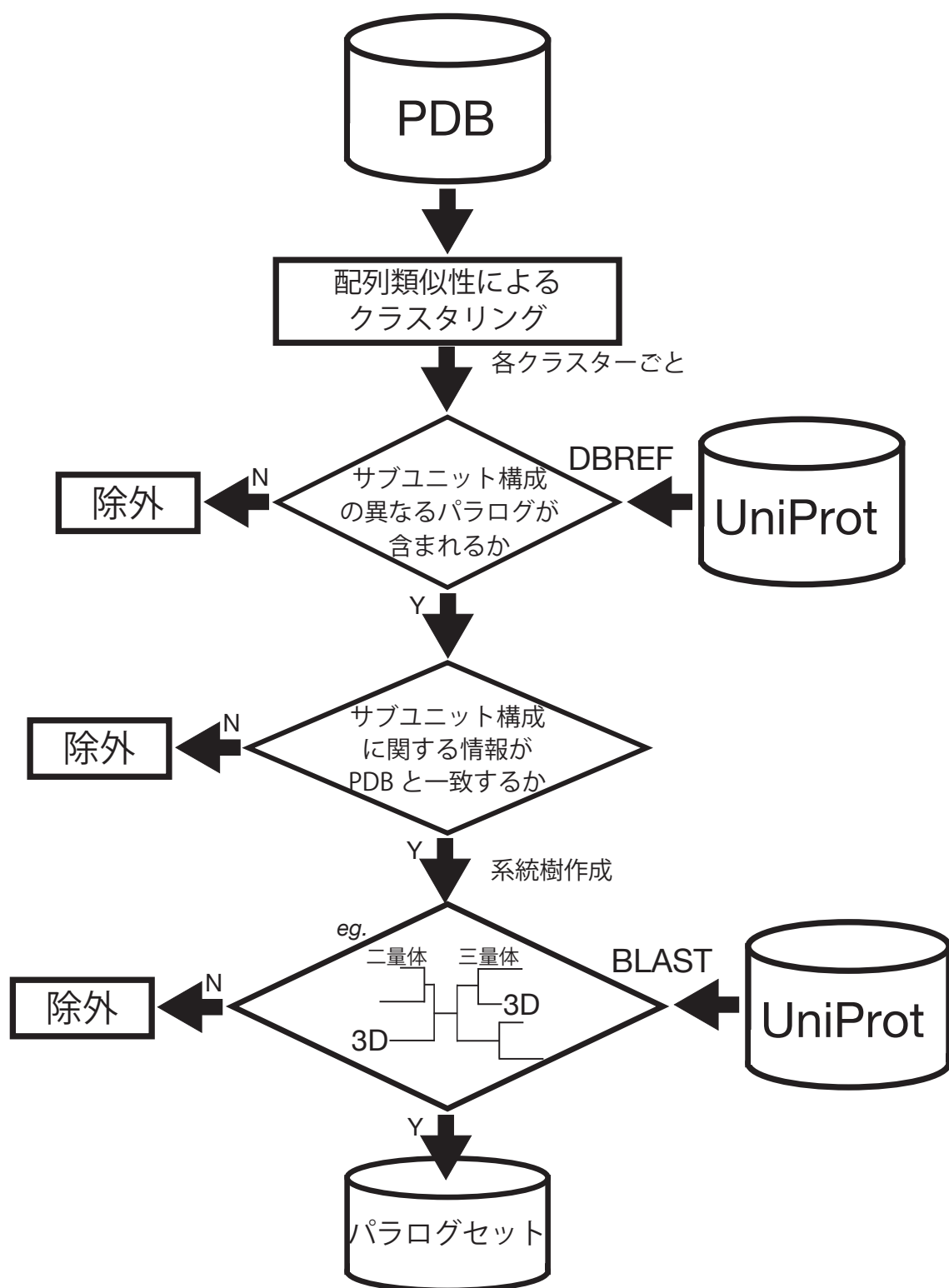


図3-2. サブユニット構成を変化させた複合体およびそのパラログの収集の流れ
 円柱はデータベース、長方形は手続き、菱形は条件分岐を示す。作業の流れは上から下へ向かって進む、条件分岐に示す特性を有する複合体をもつアミノ酸配列を収集した。図はAoto and Yura (2015)を改変して記載。

1-3 パラログセットで異なる結合部位の分類およびその定量化

パラログセット内のパラログ間で異なる結合部位を、6タイプに分類した(図3-3)。図3-3において、パラログセットにおけるパラログを比較したとき、サブユニット数もしくはサブユニットの種類が少ない方のパラログをParalog Aとし、もう一方のパラログをParalog Bとした。図3-3の結合部位の分類は、Paralog Aを基準とした時、Paralog Bの結合部位がどのように変化したかという視点で作成した。まずParalogBがホモ多量体かヘテロ多量体かで分類し、その2つについて結合部位の変化を、ParalogAが単量体でありParalogBが新規結合部位を獲得した場合(Invention)、結合部位の位置が変化した場合(switch)、ParalogAが複合体でありParalogBが新規結合部位を獲得した場合(add-on)の3パターンを考えて、計6つに分類した。

パラログセットから得られた結合部位の変化は、その結合残基の保存性と疎水度を示すHydropathy Indexとによりその変化を数値化した (Kyte and Doolittle 1982)。1つ目の指標とした結合残基の保存度は、オーソログ全体での保存度とパラログ内での保存度の2つを算出し、Shannon entropyにもとづいて作成した以下の公式(Schneider and Stephens 1990)により、そのばらつきを求めた。

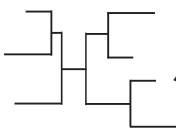
$$\Delta Conservation = -\sum_{n=1}^{21} p_n^i \times \log_2 p_n^i - \left\langle -\sum_{n=1}^{21} p_n^s \times \log_2 p_n^s \right\rangle$$

上記式の p^i は結合部位におけるアミノ酸残基種の出現頻度であり、 p^s はそのアミノ酸残基種のタンパク質表面での出現頻度を示す。 n は各アミノ酸残基種を示し、その範囲はアミノ酸20個とギャップを合わせた1~21の間の値である。この式では進化速度を各パラログどうしでそろえるため、タンパク質表面のアミノ酸残基の保存度を引いている。2つ目の指標であるHydropathy Indexの範囲は、-4.5から4.5の値をとる(Kyte and Doolittle 1982)。本

研究ではギャップを1つのアミノ酸残基として計算するため、ギャップのHydropathy Indexを-5.0と定義した。Hydropathy Indexの値は、パラログ内のすべての分子のアミノ酸配列をアラインメントし、各結合残基位置のHydropathy Indexの平均を取って得た値を用いた。得られた保存度とHydropathy Indexは、2次元プロット Δ Conservation-Hydropathy Index plotとして図式化した。

2. CPSFサブユニットへの適用

パラログどうしで異なる結合残基の分類では、CPSF160、CPSF100、CPSF73のサブユニットはすべてHeIタイプに分類される。パラログセットから得た結合部位の解析から得られたHeIタイプの結合残基に見られる傾向を元に、CPSFサブユニット特有の結合残基の予測を行った。配列保存度および疎水度の算出は、CPSF160では立体構造作成の際にテンプレート配列として用いた*Homo sapiens* DDB1(PDBID: 3EI1A)を含むDDB1のグループのマルチプルアラインメントとCPSF160のグループのマルチプルアラインメントを比較することで行った。同様に、CPSF100およびCPSF73は立体構造の決定されている*Homo sapiens* CPSF73(PDBID: 2I7VA)を含むCPSF73のマルチプルアラインメントと、CPSF100のグループのマルチプルアラインメントを互いに比較することにより配列保存度および疎水度の算出を行った。

パラログ A  パラログ B

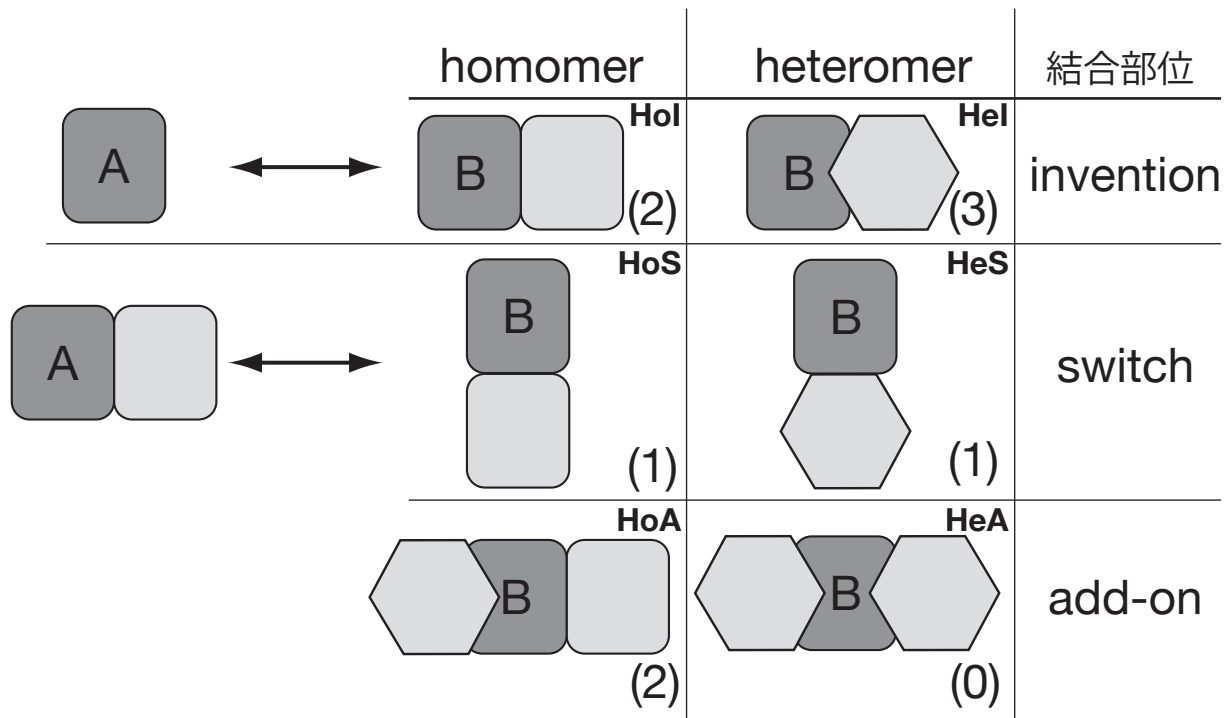


図3-3. パラログ間のサブユニット構成の分類

得られたパラログセットはまず解析対象としたサブユニット界面がホモ結合を行うものか、ヘテロ結合を行うものかで分類した。その中で、単量体が複合体に変化したものをinvention、サブユニット界面が変化したものをswitch、もともと複合体としてはたらくタンパク質が新たなサブユニット界面を獲得したものをadd-onとし、最終的に6つに分類した。図は Aoto and Yura (2015)より改変して記載。

結果

1. サブユニット構成を変化させた複合体およびそのパラログセット

PDBからサブユニット構成を変化させたと考えられる複合体およびそのパラログを9セット得た(表3-1)。これら9セットは図3-3に示した結合部位の変化のタイプに分類することができた。9セットのうち5セットは、Paralog Aが単量体、Paralog Bが多量体のタイプで、Paralog Bにはhomomer-invention(HoI)タイプとheteromer-invention(HeI)タイプが含まれていた。その他の4セットは、homomer-add-on(HoA)タイプが2つ、homomer-switch(HoS)タイプとheteromer-switch(HeS)タイプが1つずつに分類できた。

2. パラログセットから得られた結合部位の分類

2-1. Homomer-invention(HoI)タイプ

このタイプには、デオキシヌクレオシドキナーゼ(dNK)とデオキシシチジンキナーゼ(dCK)のパラログ対、およびピトリリシンとインシュリン分解酵素のパラログ対が分類された(表3-1)。

dNKは様々なデオキシリボヌクレオシドをリン酸化し、単量体でその活性を発揮する(Munch-Petersen et al. 1998)。一方、dCKはデオキシシチジン特異的なリン酸化酵素であり、ホモ二量体で機能する(Eriksson et al. 1991)。PDBには4つのdNKが結合した立体構造が登録されていたが(Mikkelsen et al. 2008)、先行研究の実験で得られている最も長いアミノ酸配列が決定されているサブユニットを選出し、代表構造とした。dCKの立体構造はホモ二量体として決定されており(Mikkelsen et al. 2008)、選出したdNKとdCKの立体構造を重ね合わせることで、35残基で構成される結合面のアミノ酸残基を特定した。dNKとdCKの残基の特性を図3-4.bに示す。グラフのX軸である Δ Conservationはアミノ酸残基の保存度を

表1-3. パラログ間で異なるサブユニット構成をもつタンパク質

Aoto and Yura. (2015)より転載

Paralogue A				Paralogue B				パラログ間の比較				
分子名*	多量体構成	UniProtID	PDB chain	分子名	identity (%)**	多量体構成	UniProtID	PDB chain	C α RMSD	結合部位***	Δ Conservation	疎水性度
デオキシヌクレオシドキナーゼ(5)	monomer	DNK_DROME	2LJ8 B	デオキシシチジンキナーゼ (8)	56±25	homodimer	DCK_HUMAN	2NO0 A	1.63Å (143 aa)	Hol	0.1	0.38
ピトリリンシン(11)	monomer	PTRA_ECOLI	1Q2L A	インシユリン分解酵素 (15)	40±23	homodimer	IDE_HUMAN	2G47 A	2.05Å (428 aa)	Hol	0.2	0.06
HlsA(535)	monomer	HIS4_THEMA	1QO2 A	HisF (529)	50±11	heterodimer	HIS6_THEMA	1GPW A	2.22Å (226 aa)	Hel	0.02	0.05
L-アスパラギン酸オキシダーゼ (37)	monomer	NADB_ECOLI	1KNP A	コハク酸デヒドロゲナーゼ SdhA (30)	66±13	heterotetramer	SDHA_PIG	3AE9 A	1.90Å (499 aa)	Hel	0.22	0.01
ミオグロビン (104)	monomer	MYG_HUMAN	3RKG A	ヘモグロビン β サブユニット (289)	62±16	heterotetramer	HBB_HUMAN	4HHB B	1.47Å (145 aa)	Hel	0.08	0.04
スレオニン合成酵素 (20)	homodimer	THRC_MYCTU	2D1F A	シスタチオニン β 合成酵素 (33)	36±13	homotetramer	CBS_HUMAN	1M54 A	2.69Å (290 aa)	HoA	0.06	0.2
D-システイン脱硫化水素酵素 (72)	homodimer	DCYD_SALTY	4D9F D	トリプトファン合成酵素 β サブユニット (364)	52±15	heterotetramer	TRPB_SALTY	1K7F B	2.73Å (197 aa)	HoA	0.03	0.49
MoeA (30)	homodimer	MOEA_ECOLI	2NQG A	グアイリン (5)	57±36	homotrimer	GEPH_HUMAN	1L1J A	2.87Å (135 aa)	HoS	0.11	0.4
GatD (34)	heterodimer	GATD_METTH	2D6F A	L-アスパラギナーゼ I (25)	34±15	homotetramer	ASPG1_ECOLI	2H1M A	1.96Å (307 aa)	HeS	0.01	0.3

* 括弧内の数値はUniProt中に登録されたホモログな配列の本数を示す

** 各分子ペアにおける配列一致度の二乗平均値とその標準偏差

*** 結合部位の分類は図5に示す

**** 図6から図12に示したグラフのX軸とY軸の分離率。T検定によって算出した

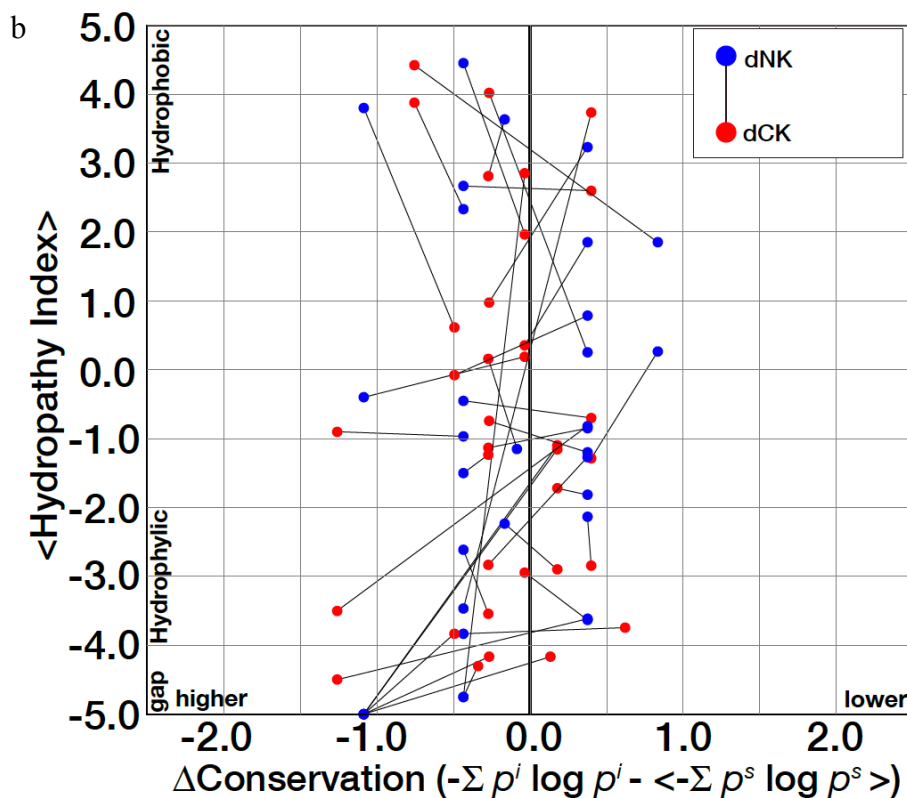
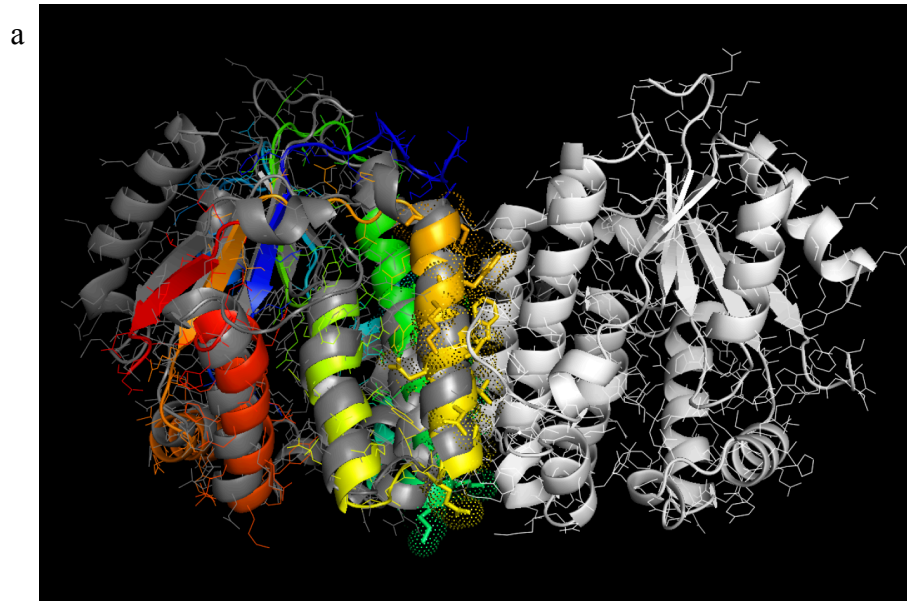


図3-4. dNKとdCKのサブユニット界面の比較 (HoIタイプ)

a. dNKおよびdCKの立体構造を重ね合わせた図。dNKをレインボーカラーで、dCKをグレーと白で示す。b. dNKおよびdCKの Δ Conservation-Hydropathy Index plot。赤いドットはホモ二量体であるdCKの結合部位に存在する残基を示し、青いドットはdNKのアミノ酸残基を示している。アラインメント上、同一位置のアミノ酸残基どうしを実線で結んで示した。dCKおよびdNKのドットの分布はプロットの Δ Conservation軸0.0付近に散らばっており、これはdNKとdCKの結合部位残基間の保存度に差がないことを示している。またhydropathy Index軸方向に両酵素とも大きく散らばっており、疎水的な傾向は見られない。図はAoto and Yura (2015)より転載。

示し、dCKの結合残基およびそれに相当するdNKのアミノ酸残基を示す点のほとんどが Δ Conservation軸の0.0付近に位置していた。これはホモ二量体のdCKと単量体のdNKの結合部位のアミノ酸残基の保存度が同程度であることを示している。dNKとdCKのグラフにおいては、10残基が Δ Conservationの値が-0.5未満にプロットされ、よく保存されていた。このうち赤い点で示される3残基は、ホモ二量体であるdCKに特異的に保存されていた。これらの残基は多量体構造の安定性などに関わる“hot spot”と呼ばれるアミノ酸残基と考えられる。また、残りの青い点で示されるdNKの7残基のうち5残基はY軸の値が0.0であり、ギャップであることが分かった。これらのことから、ホモ二量体dCKは単量体のdNKのもっていない5つのアミノ酸残基を結合残基として有していることが分かった。

結合部位においてアミノ酸残基の保存度が高いことは、多くの研究ですでに指摘されてきた(Janin et al. 2008)。しかし、本研究による解析からは単量体であるdNKの対応するアミノ酸残基もdCKと同程度に保存されていることが示されていた。また、結合部位は疎水性な傾向を持つことも多く指摘されているが(Janin et al. 2008)、また、dCKの結合残基は疎水性なアミノ酸残基に偏るような傾向は見られなかった。挿入による結合残基の重要性についてはすでに指摘されており(Hashimoto et al. 2010, Nishi et al. 2011)、dNK/dCKの解析においても結合部位の獲得に重要な役割を果たしていることが分かった。しかし、これらの挿入によって獲得したアミノ酸残基はあまり保存されていなかった。

*Escherichia coli*を始めBacteriaが有するピトリリシンは、小さいペプチドに特異的に働く亜鉛メタロエンドペプチダーゼの1つであり、単量体ではたらく。ピトリリシンの哺乳類のホモログであるインシュリン分解酵素は、ホモ二量体であることが知られている(Ding et al. 1992)。これらの酵素は、N末端側ドメインとC末端側ドメインの2つのドメインで構成されている。本研究で使用したピトリリシンの立体構造は、C末端側ドメインとN末端側

ドメインが蝶番のように開いていたため、インシュリン分解酵素(Shen et al. 2006)の1サブユニットと構造全体をかさねることができなかった。そこで、ホモ二量体のインシュリン分解酵素の結合部位があるC末端側ドメインがピトリリシンのC末端側ドメインと一致するように、2つの立体構造を重ね合わせた(図3-5.a)。 Δ Conservation-Hydrophathy Index plot(図3-5.b)をみると、2つのパラログ間に4つの明確な特徴が見られた。一つ目の特徴は、 Δ Conservation-Hydrophathy Index plotの左側に突き出た2つの“hot spot”残基である。この特徴はdNK/dCKのペアと類似している。二つ目の特徴は、インシュリン分解酵素の結合残基を示す青い点の多くがHydrophathy Indexの値が-5.0の位置にプロットされている点である。これはピトリリシンを基準とした場合、インシュリン分解酵素の結合残基は挿入によって得られたアミノ酸残基であることを示す。三つ目の特徴は、インシュリン分解酵素の結合残基を示す赤い点がグラフの上部には見られず、疎水性の傾向を示していない点である。このグラフでは、ピトリリシンのアミノ酸残基を示す青い点は、ギャップを除き、ホモ二量体であるインシュリン分解酵素の結合残基を示す赤い点よりグラフの上部にあり、インシュリン分解酵素の結合残基はピトリリシンのタンパク質表面のアミノ酸残基よりも疎水的ではないことを示している。四つ目の特徴は、インシュリン分解酵素の結合残基を示す赤い点がグラフの右側に位置している点であり、これは結合部位を構成している残基の保存度が結合部位ではないアミノ酸残基よりも低いことを示している。

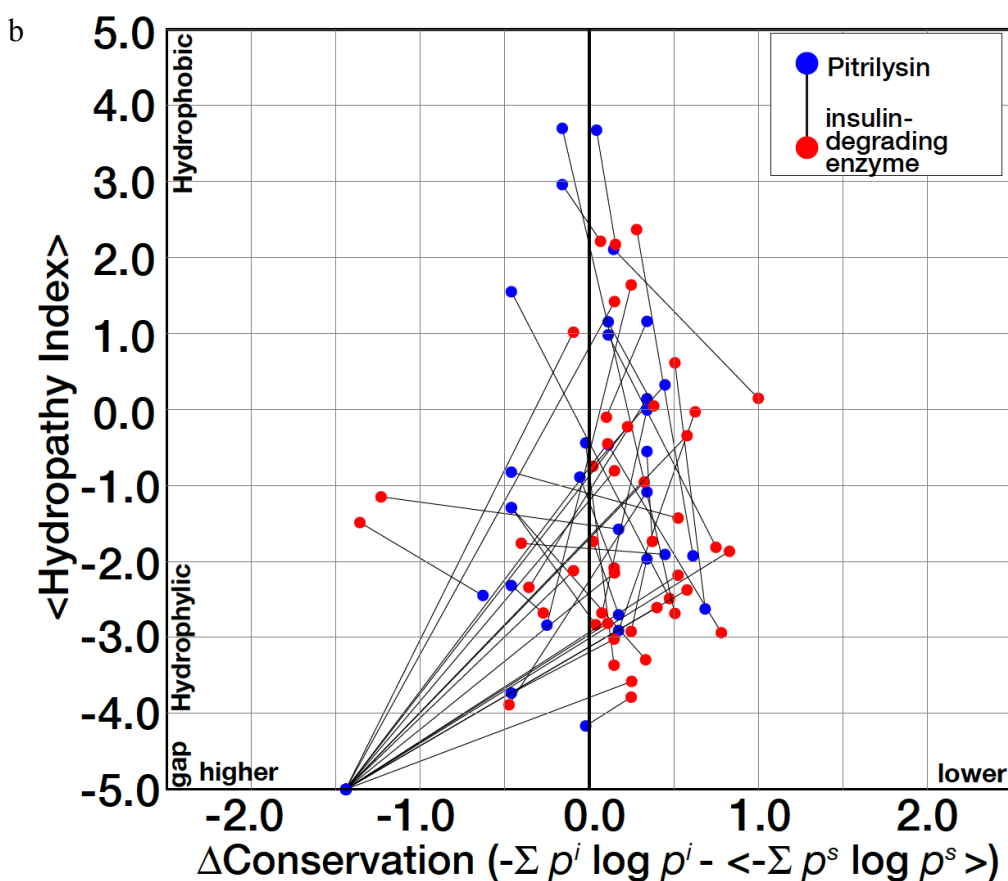
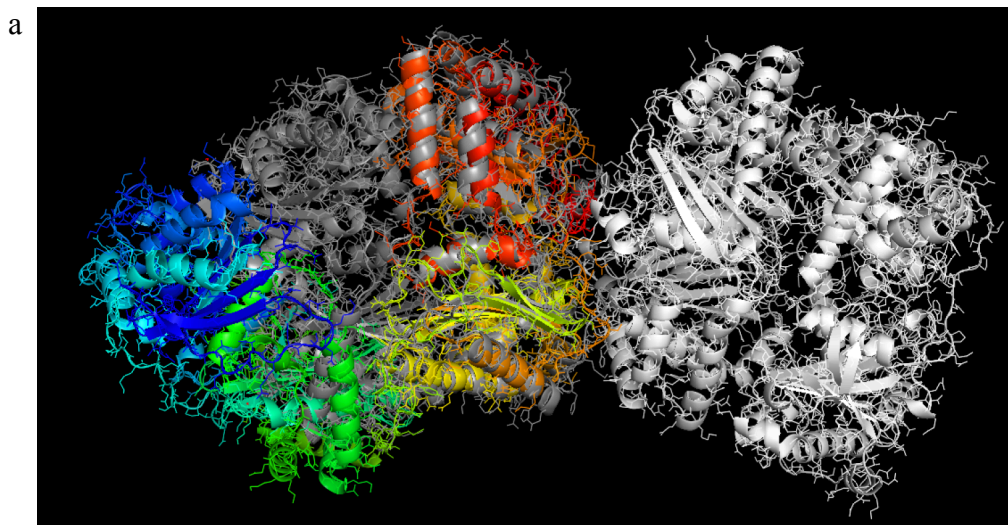


図3-5. ピトリリシンとインシュリン分解酵素のサブユニット界面の比較 (HoIタイプ)
 a. ピトリリシンとインシュリン分解酵素の立体構造を重ね合わせた図。ピトリリシンをレインボーカラーで、インシュリン分解酵素をグレーと白で示す。ピトリリシンの立体構造はインシュリン分解酵素に比べC末端側ドメインとN末端側ドメインが開いていたため、インシュリン分解酵素の結合部位が存在するC末端側ドメインを重ね合わせた。b. ピトリリシンとインシュリン分解酵素の $\Delta\text{Conservation}$ -Hydropathy Index plot。図はAoto and Yura (2015)より転載。

2-2. Heteromer-invention(HeI)タイプ

このタイプには、HisAとHisF、L-アスパラギン酸オキシダーゼとコハク酸デヒドロゲナーゼ、ミオグロビンとヘモグロビン β の3セットが分類された。

HisAはホスフォリボシルフォルミミノ-5-アミノイミダゾールカルボキシアミドリボタイドイソメラーゼの1つであり、ヒスチジン合成経路の4つ目のステップではたらく。HisAは $(\beta\alpha)_8$ バレルと呼ばれる基本構造をもち(Lang et al. 2000)、*Escherichia coli*および*Thermotoga maritima*では単量体としてはたらく(Henn-Sax et al. 2002)。HisFはヒスチジン合成経路の5つ目のステップでイミダゾールグリセロールリン酸合成を触媒する。この酵素はHisHとヘテロ二量体を構成することで反応を触媒することが示されており(Klem and Davisson 1993)、複合体として立体構造が決定されている(Douangamath et al. 2002)。HisAとHisFの立体構造を重ね合わせると、HisHはHisAが重ならない部分でHisFと結合していた(図3-6.a)。 Δ Conservation-Hydrophathy Index plotでは、HoIタイプとは異なる特徴が見られた。HisFで結合部位を構成している36個のアミノ酸残基は、HisAのアミノ酸残基に対し Δ Conservation-Hydrophathy Index plotの上側に存在しており、疎水性が高い傾向が見られた。このグラフにおいて、HisHとHisAの点の分布はウィルコクソン検定においてp値0.001以下を示し、2つの構造のアミノ酸残基は有意に傾向が異なることが分かった。さらに、赤い点は左に突き出しているものが多く、結合残基は疎水的な残基であり、保存度が高いことが分かった。これらの結果より新たに結合部位を獲得した場合、その結合部位の傾向はホモ複合体とヘテロ複合体で大きく異なることが分かった。

L-アスパラギン酸オキシダーゼはコファクター生合成経路に関わるフラボプロテインの1つであり、*Escherichia coli*において単量体として機能し(Mortarino et al. 1996)、その立体構造も単量体として決定されている(Bossi 2002)。コハク酸デヒドロゲナーゼはクエン酸

回路に関与する巨大な酵素の複合体であり、フラボプロテインサブユニット(SdhA)、コハク酸デヒドロゲナーゼ鉄硫黄サブユニット Succinate dehydrogenase iron-sulfur subunit (SdhB)、および膜貫通領域である2つのシトクローム b₅₆₀ サブユニット(SdhC, SdhD)の4つのサブユニットで構成される(Stenberg et al. 2005)。コハク酸デヒドロゲナーゼは四量体を構成し、その全体の立体構造は既に決定されている(Stenberg et al. 2005)。SdhAとL-アスパラギン酸オキシダーゼはホモログスであり、その立体構造はよく似ている(図3-7.a)。

SdhAはSdhBとのみ直接結合していることが明らかであったため、SdhAとSdhBのヘテロダイマーとして、その結合残基とそれにアラインメントされる単量体のL-アスパラギン酸オキシダーゼのアミノ酸残基の傾向を比較した。ΔConservation-Hydrophathy Index plot (図3-7.b)において、結合部位を構成するアミノ酸残基およびその対応するアミノ酸残基のペアは78個あり、L-アスパラギン酸オキシダーゼを示す青い点は中心付近に位置しており、その周囲にSdhAを示す赤い点が散らばっていた。グラフ上部には明らかに赤い点が青い点より上に位置しており、結合部位が疎水性であることを示していた。

グロビンファミリーのタンパク質については大規模に研究されており、グロビンファミリーのタンパク質はわずかなアミノ酸残基の違いによってサブユニット構成が劇的に異なっていることが発見されている(Shionyu et al. 2001)。本研究では、単量体のグロブリンタンパク質であるミオグロビンとミオグロビンに最も近縁なホモログであるヘモグロビンβサブユニット間の結合部位の差異を解析した(図3-8. a, b)。ヘモグロビンβサブユニットと他のサブユニットの結合部位には22のアミノ酸残基が存在し、他の HeIタイプと同様に、結合残基は疎水性であり、保存度の高い傾向を示した。しかし、統計的に有意な差を得ることはできなかった。

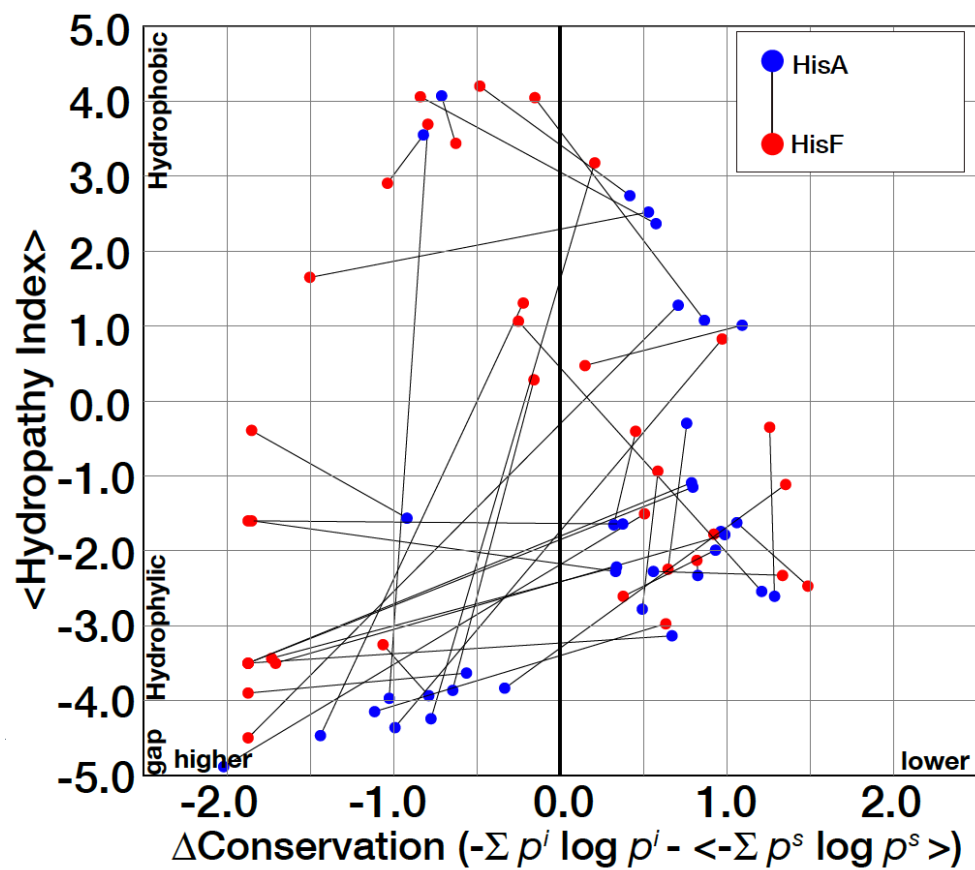
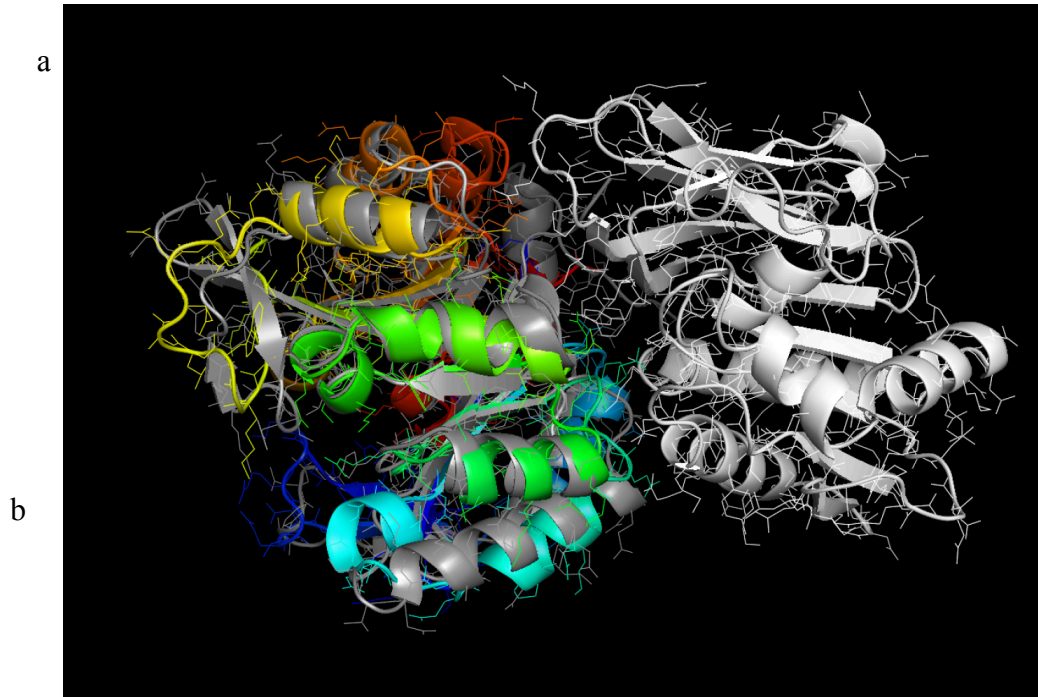


図3-6. HisAとHisFのサブユニット界面の比較 (HeIタイプ)

a. HisAとHisFの立体構造を重ね合わせた図。HisAをレインボーカラー、HisFをグレーと白で示す。b. HisAとhisFの Δ Conservation-Hydropathy Index plot。図はAoto and Yura (2015)より転載。

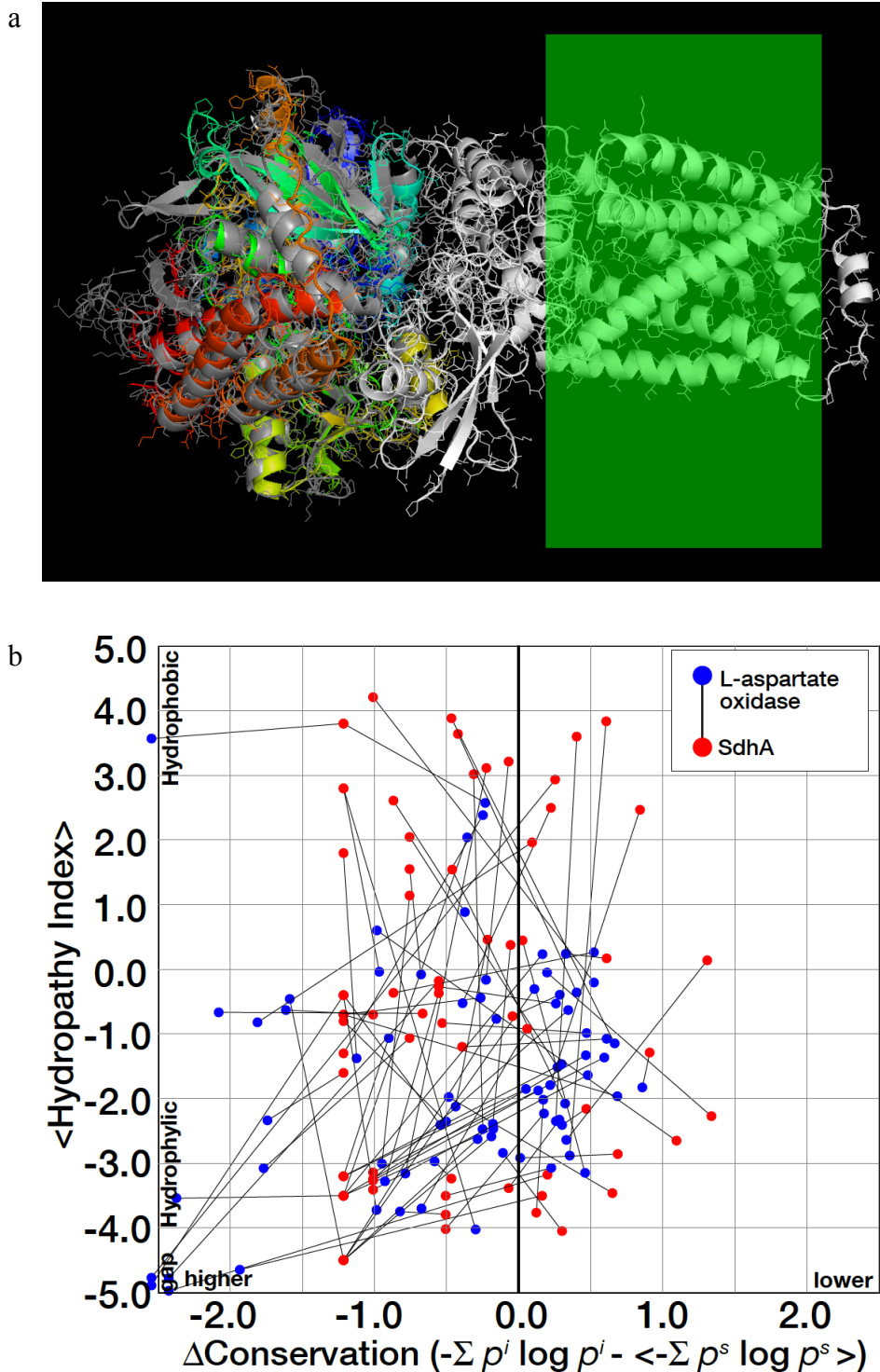


図3-7. L-アスパラギン酸オキシダーゼとコハク酸デヒドロゲナーゼのサブユニット界面の比較 (HeIタイプ)

a. L-アスパラギン酸オキシダーゼとコハク酸デヒドロゲナーゼを重ね合わせた図。L-アスパラギン酸オキシダーゼをレインボーカラーで、SdhAをグレー、SdhBを白で示した。半透明の緑色のボックスは細胞膜の存在する位置を示す。b. L-アスパラギン酸オキシダーゼとコハク酸デヒドロゲナーゼの $\Delta\text{Conservation}$ -Hydropathy Index plot。図はAoto and Yura (2015)より転載。

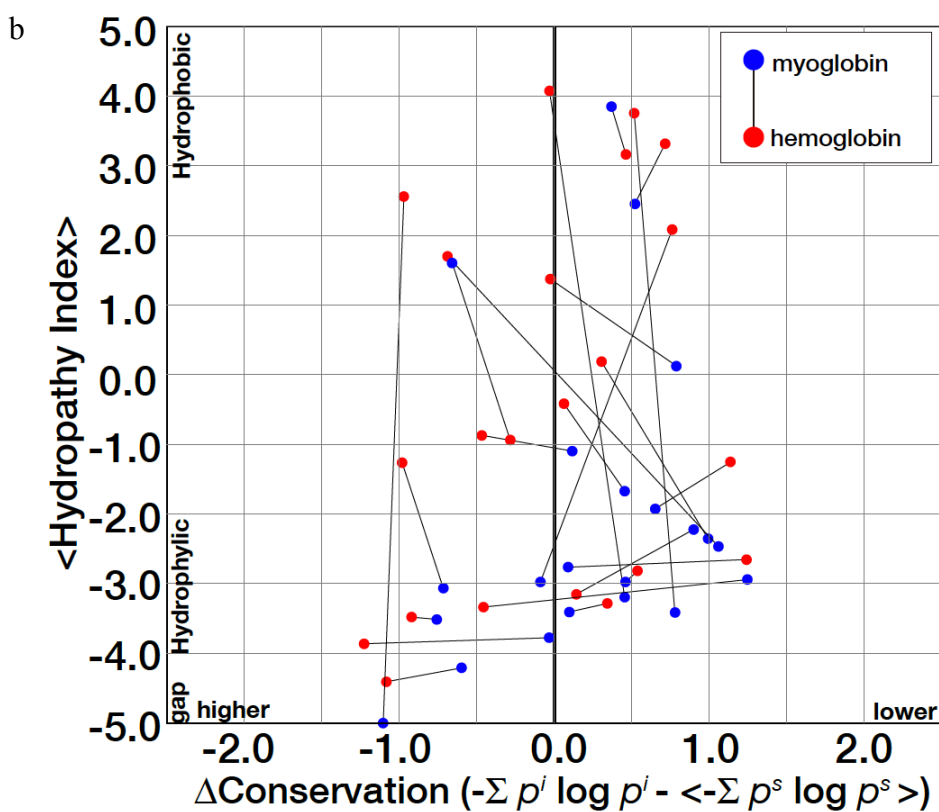
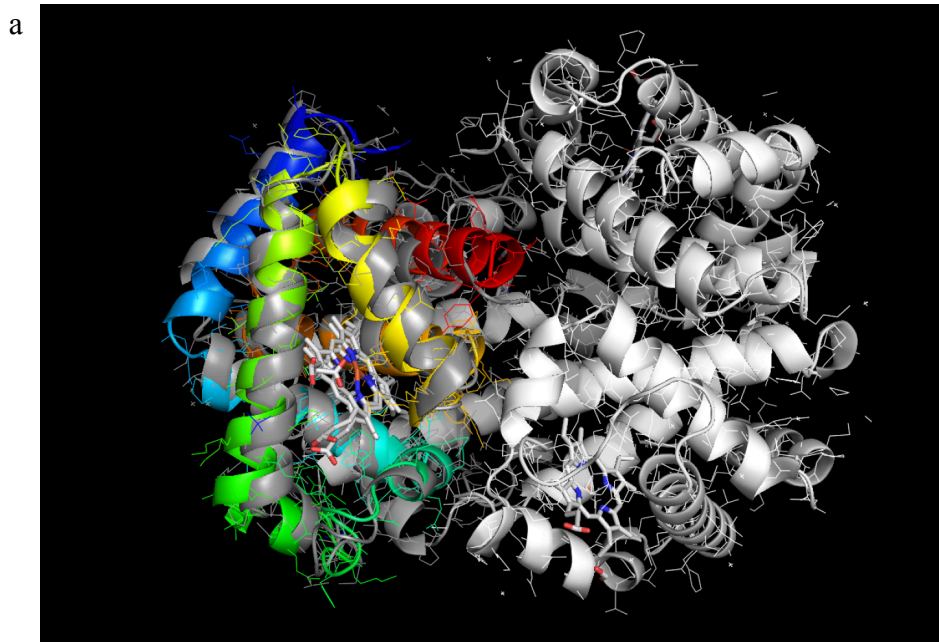


図3-8. ミオグロビンとヘモグロビンβサブユニットのサブユニット界面の比較 (HeIタイプ)

a. ミオグロビンとヘモグロビンβサブユニットを重ね合わせた図。ミオグロビンをレインボーカラーで、ヘモグロビンβサブユニットをグレー、ヘモグロビンαサブユニットを白で示した。b. ミオグロビンとヘモグロビンβサブユニットの $\Delta\text{Conservation}$ -Hydropathy Index plot。図はAoto and Yura (2015)より転載。

2-3. Homomer-add-on(HoA)タイプ

このタイプには、スレオニン合成酵素とシスタチオニン β 合成酵素、D-システイン脱硫化水素酵素とトリプトファン合成酵素 β サブユニットの2セットが分類された(表4)。スレオニン合成酵素、シスタチオニン β 合成酵素、D-システイン脱硫化水素酵素、トリプトファン合成酵素 β サブユニットの4つは同じPLP依存酵素スーパーファミリーに属するが、アミノ酸配列の類似性により2つのセットに分けて解析を行なった。

スレオニン合成酵素は、Bacteriaの1種である*Corynebacterium glutamicum*では単量体であるが(Malumbres et al. 1994)、*Mycobacterium tuberculosis*ではホモ二量体 (Covarrubias et al. 2008)、また*Arabidopsis thaliana*でもホモ二量体であることが示されている (Laber et al. 1999)。一方、シスタチオニン β 合成酵素はホモ四量体を構成している(Meier et al. 2001)。本研究においては、ホモ二量体のスレオニン合成酵素とホモ四量体のシスタチオニン β 合成酵素に着目し、このパラログをHoIタイプではなく、HoAタイプとして扱い、解析を行った(表4)。二量体スレオニン合成酵素の1つのサブユニットと四量体のシスタチオニン β 合成酵素(Taoka et al. 2002)の1サブユニットを重ね合わせると、二量体スレオニン合成酵素のもう一方のサブユニットは、シスタチオニン β 合成酵素の一方のサブユニットと重なることが分かった(図3-9.a)。このことは、シスタチオニン β 合成酵素はスレオニン合成酵素にはない新しい結合部位を獲得したことで、二量体から四量体に移行したと考えることができる。シスタチオニン β 合成酵素の結合部位を構成するアミノ酸残基とそれに対応するスレオニン合成酵素のタンパク質表面のアミノ酸残基の Δ Conservation-Hydrophathy Index plotを図3-9.bに示す。また、スレオニン合成酵素もシスタチオニン β 合成酵素もホモ多量体であるため、図3-9.cではシスタチオニン β 合成酵素とスレオニン合成酵素双方がもつ結合部位を構成するアミノ酸残基の Δ Conservation-Hydrophathy Index plotを示した。グラフの左側

にとびでてプロットされている“hot spot”と考えられる2つのアミノ酸残基があることが分かったが(図3-9.b)、シスタチオニンβ合成酵素とスレオニン合成酵素双方がもつ結合残基のΔConservation-Hydrophathy Index plotと比較したところ、疎水度と保存度に明らかな傾向は見られなかった。

スレオニン合成酵素とシスタチオニンβ合成酵素の結果に反して、D-システイン脱硫化水素酵素とトリプトファン合成酵素βサブユニットには、結合部位を構成するアミノ酸残基は特徴的であった。D-システイン脱硫化水素酵素は、BacteriaがもつPLP依存性のD-システイン分解反応を触媒する酵素であり(Nagasawa et al. 1985)、D-システインの蓄積に対する生体防御機構と考えられている。この酵素はホモ二量体ではたらいっていると考えられ、結晶構造もホモ二量体で決定されている。トリプトファン合成酵素βサブユニットは、αサブユニットとともにトリプトファン生合成系の最終段階でインドールとL-セリンからL-トリプトファンを合成する反応を触媒する。αサブユニットとβサブユニットがヘテロ二量体を構成し、それだけで反応を完了することができるが、α-β-β-α四量体は1~2桁反応が早く、この四量体が機能的な複合体のサブユニット構成として認められている(Hyde et al. 1988)。これらの2つの複合体構造を比較すると、D-システイン脱硫化水素酵素のホモ二量体とトリプトファン合成酵素βサブユニットの二量体はほぼ同じ形状をしている。したがってトリプトファン合成酵素βサブユニットは、D-システイン脱硫化水素酵素の単量体構造と相当な構造が新たな結合部位を得たことで、トリプトファン合成酵素αサブユニットと結合できるようになったと考えることができる。またトリプトファン合成酵素βサブユニットは、αサブユニットと結合するための新たな結合部位を得たと捉えることができる(図3-10.a)。これを踏まえ、ヘテロ結合部分の結合部位を対象とし、これらの分子をHeIタイプとして解析した(図10.b)。トリプトファン合成酵素βサブユニットは

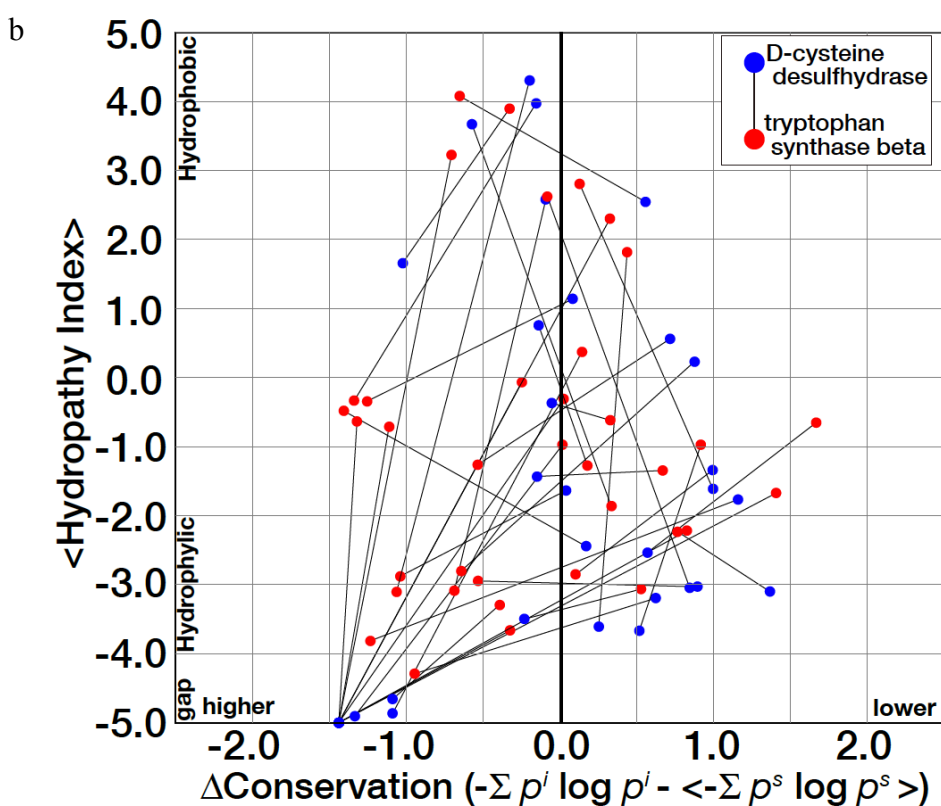
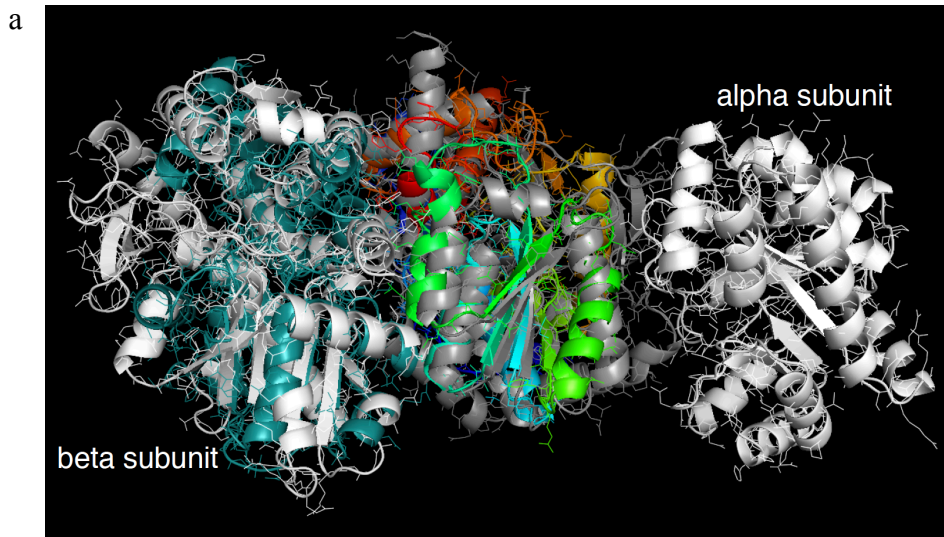


図3-9. ホモ二量体スレオニン合成酵素とホモ四量体のシスタチオニンβ合成酵素のサブユニット界面の比較 (HoAタイプ)

a. ホモ二量体スレオニン合成酵素(レインボーとシアン)とホモ四量体のシスタチオニンβ合成酵素(グレーと白)についてレインボーとグレーで示したサブユニットを重ね合わせた図。二量体スレオニン合成酵素の1つのサブユニットと四量体のシスタチオニンβ合成酵素の1サブユニットを重ね合わせると、二量体スレオニン合成酵素のもう一方のサブユニットはシスタチオニンβ合成酵素の他のサブユニットと重なった。そのため自然に重なり合ったサブユニットは表示していない。b. シスタチオニンβ合成酵素の結合部位を構成するアミノ酸残基とそれに対応するスレオニン合成酵素のタンパク質表面のアミノ酸残基の $\Delta\text{Conservation}$ -Hydropathy Index plot。c. シスタチオニンβ合成酵素とスレオニン合成酵素双方が持つ結合部位を構成するアミノ酸残基の $\Delta\text{Conservation}$ -Hydropathy Index plot。図はAoto and Yura (2015)より転載。

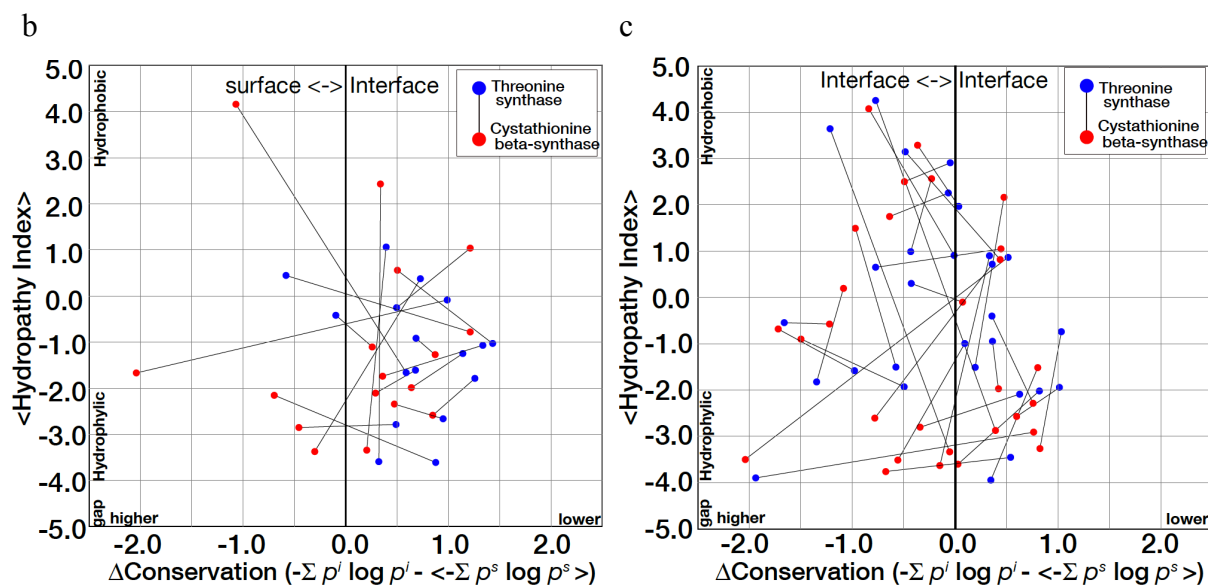
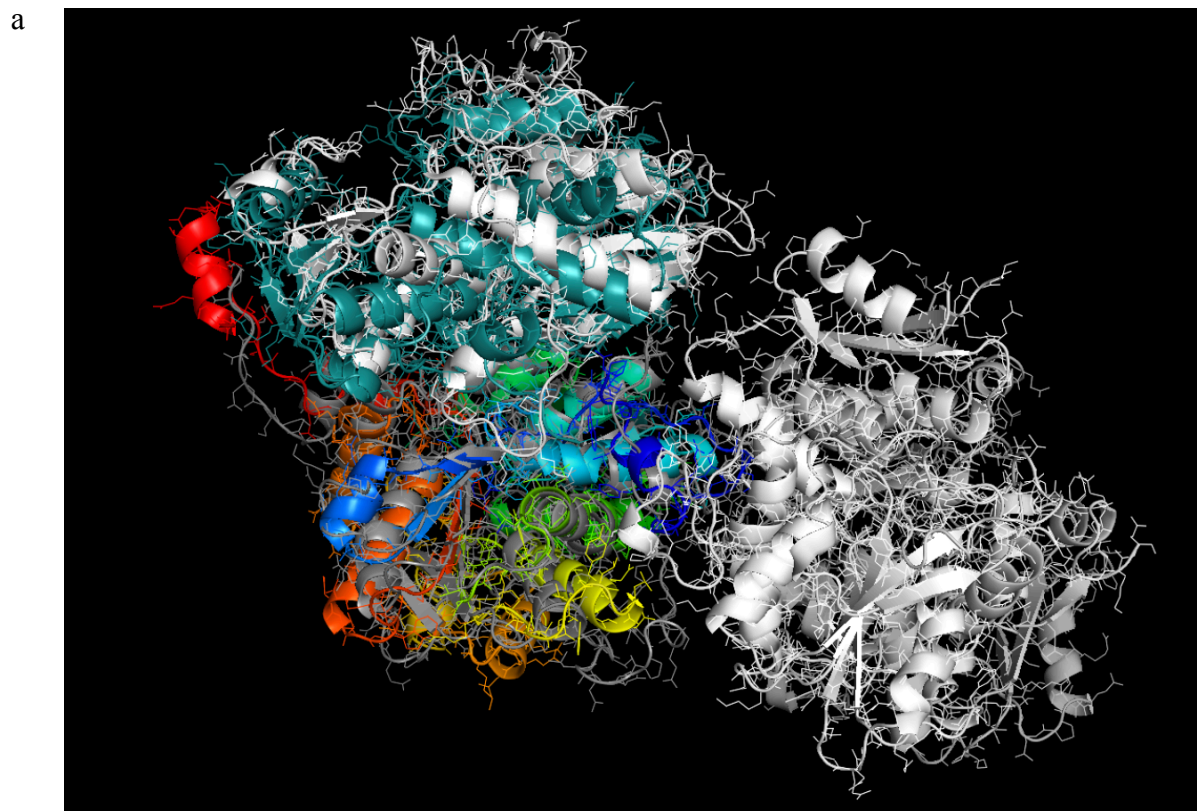


図3-10. D-システイン脱硫化水素酵素のホモ二量体とトリプトファン合成酵素βのサブユニット界面の比較 (HoAタイプ)

a. D-システイン脱硫化水素酵素のホモ二量体とトリプトファン合成酵素βのヘテロ二量体を重ね合わせた図。D-システイン脱硫化水素酵素をレインボーカラーとシアンで、トリプトファン合成酵素βサブユニットbをグレーとサブユニットaを白で示す。b. D-システイン脱硫化水素酵素とトリプトファン合成酵素βサブユニットbの Δ Conservation-Hydropathy Index plot。図はAoto and Yura (2015)より転載。

D-システイン脱硫化水素酵素に比べ疎水度が高く、ギャップを示すX軸上にプロットされた青い点がHoIタイプのものより目立っていた。

2-4. Homomer-switch(HoS)タイプ

このタイプにはMoeAとゲフィリンの1つのパラログセットだけが該当した。MoeAはモリブデン酸塩がモリブドプテリンに挿入される反応を触媒し、その際ホモ二量体を形成する(Xiang et al. 2001)。ゲフィリンは多機能タンパク質であり、初め細胞骨格に関連していると考えられていたが、後にグリシン受容体を阻害するはたらきが発見された(Schwarz et al. 2001)。ゲフィリンはアミノ酸配列全長の立体構造がホモ二量体として(Kim et al. 2006b)、C末端側ドメインのGドメインのみの立体構造がホモ三量体として決定されており(Sola et al. 2004)、これら2つの複合体のサブユニット構成は双方とも機能的に正しいと考えられた。ホモ二量体のゲフィリンの立体構造はMoeAとほぼ同じ形状であるが、三量体の結合様式は明らかに異なっていた。ゲフィリンのホモ二量体構造とホモ三量体構造を対応する部位で重ね合わせたところ、Gドメインはどちらの構造でも結合に関与していたが、二量体と三量体では結合部位が異なっていた(図3-11.a)。この違いは、HoIタイプで見られたようなアミノ酸残基の挿入または欠失による結合部位の制御であると解釈することができ、ゲフィリンのN末端側にアミノ酸残基が挿入もしくは欠失が起こることにより、サブユニット間の結合部位の位置が変化したと考えられる。

2-5. Heteromer-switch(HeS)タイプ

このセットには、ArchaeaのグルタミニルtRNAアミドトランスフェラーゼのD鎖(GatD)とL-アスパラギナーゼの1つパラログセットだけが該当した。GatDはtRNAのもつグルタ

ミンをグルタミン酸に転移する反応を触媒する酵素であり、L-アスパラギナーゼIはアスパラギン酸をアスパラギンに転移させる酵素である。GatDはGatEとヘテロ二量体を形成し(Tumbula et al. 2000)、tRNAと複合体を構成する(Oshikane et al. 2006)。一方、L-アスパラギナーゼIはホモ二量体を構成し、さらにL-アスパラギナーゼIIのホモ二量体と結合して、ヘテロ四量体を形成する(Yun et al. 2007)。GatDとL-アスパラギナーゼIの立体構造を重ね合わせると、GatEと重ね合わせていない方のL-アスパラギナーゼIの立体構造は一部がかさなっているようであったが(図3-11.b)、結合部位の大部分は重なっていなかった。またGatDとGatEの結晶構造は二量体が二つ結合した状態で決定されていたのだが(Oshikane et al. 2006)、GatDどうしの結合部位はL-アスパラギナーゼIのホモ四量体で見られたものとよく似ており、GatDのホモ二量体立体構造とL-アスパラギナーゼIのホモ二量体立体構造は重なることがわかった(図3-11.c)。これらの結果から、GatDとGatEの結合部位とL-アスパラギナーゼIとL-アスパラギナーゼIIの結合部位は一致しておらず、パラログ間で結合部位の位置を変化させていることが分かった。しかし、その疎水度とアミノ酸配列保存度に傾向は見られなかった。

上記の結果より、明らかな傾向がみられたHeIタイプであれば、得られた疎水度およびアミノ酸残基の保存度の傾向から、パラログとは異なる結合残基を予測することができる。CPSF複合体はヘテロ複合体であり、そのサブユニット構成は各CPSFサブユニットのパラアログのものとは明らかに異なっているため、HeIタイプの結合部位を構成するようなアミノ酸残基を予測できると考えられた。

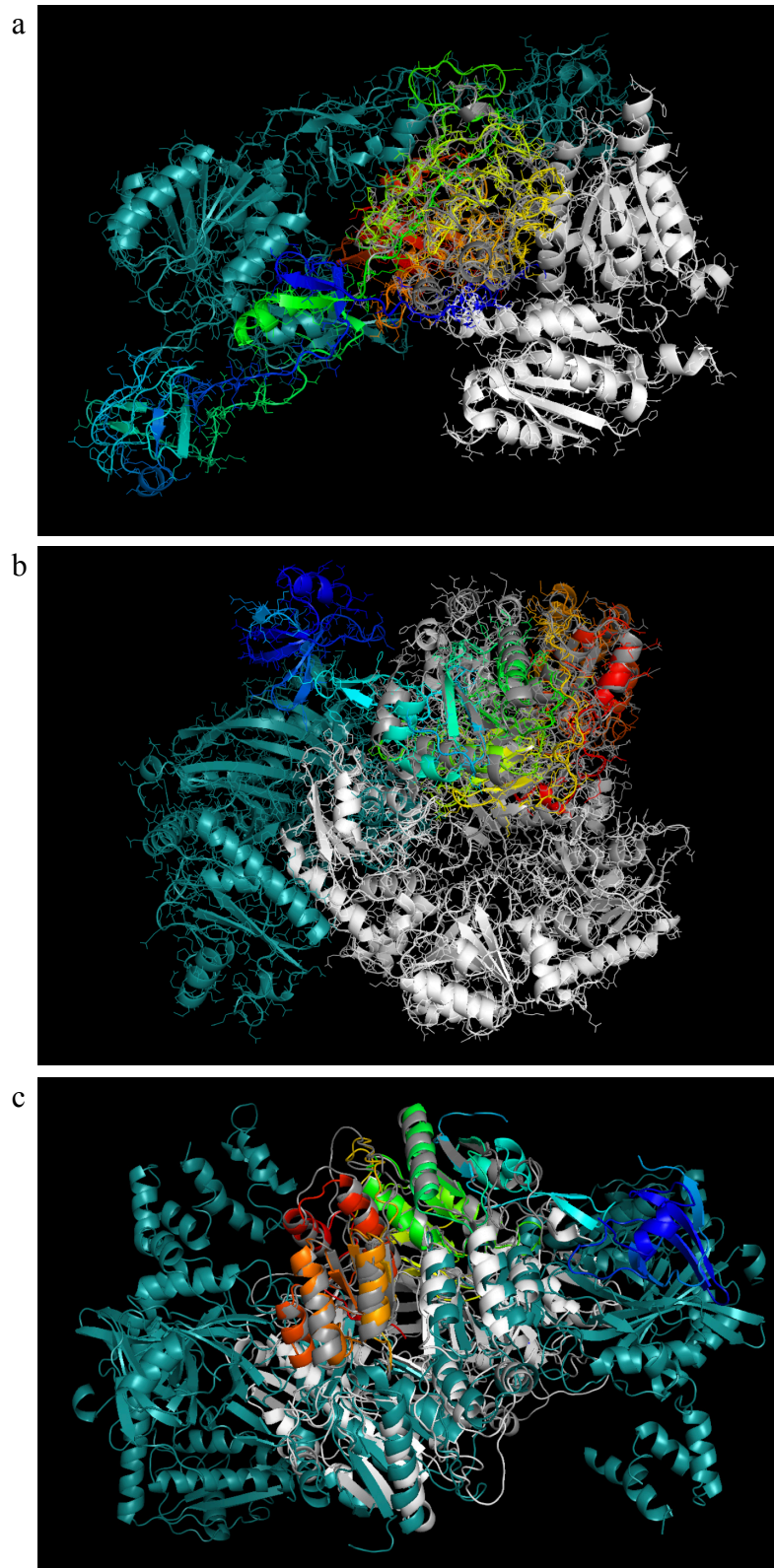


図3-11. HoSタイプのサブユニット界面の比較

a. MoeA(レインボーカラーとシアンで示す)とゲフィリン(グレーと白)の立体構造を重ね合わせた図。b. ヘテロ二量体であるGatD(レインボーカラーとシアン)とヘテロ四量体(ヘテロに両体が2つついた形状)のL-アスパラギナーゼI(グレーと白)の立体構造を重ね合わせた図。GatDの結合部位と比較しやすいように、L-アスパラギナーゼIはヘテロ二量体として表示している。c. GatDのヘテロ四量体(レインボーカラーとシアン)とL-アスパラギナーゼI(グレーと白)の立体構造を重ね合わせた図。図は Aoto and Yura (2015)より転載。

3. CPSFサブユニットにおけるパラログ間で異なる結合残基の予測

パラログ間で異なる結合部位の解析から得られたHeIタイプのHydropathy Indexと Δ Conservationの値を満たすCPSFサブユニット上の結合残基の特定を行なったところ、CPSF160は17残基、CPSF100は13残基、CPSF73は9残基のアミノ酸残基がCPSFサブユニットに特異な結合に関与することが分かった(図3-12)。CPSF160はC末端側のプロペラドメインに偏って結合残基が予測された(図3-12. a, b)。CPSF73の予測結合残基は活性部位周辺ではなく、CPSF73のエンドヌクレアーゼ活性を妨げない結果を得た(図3-12.c)。一方、エンドヌクレアーゼ活性をもたないとされるCPSF100では、CPSF73では活性部位にあたる領域に特に多くの結合部位が予測された(図3-12. d)。

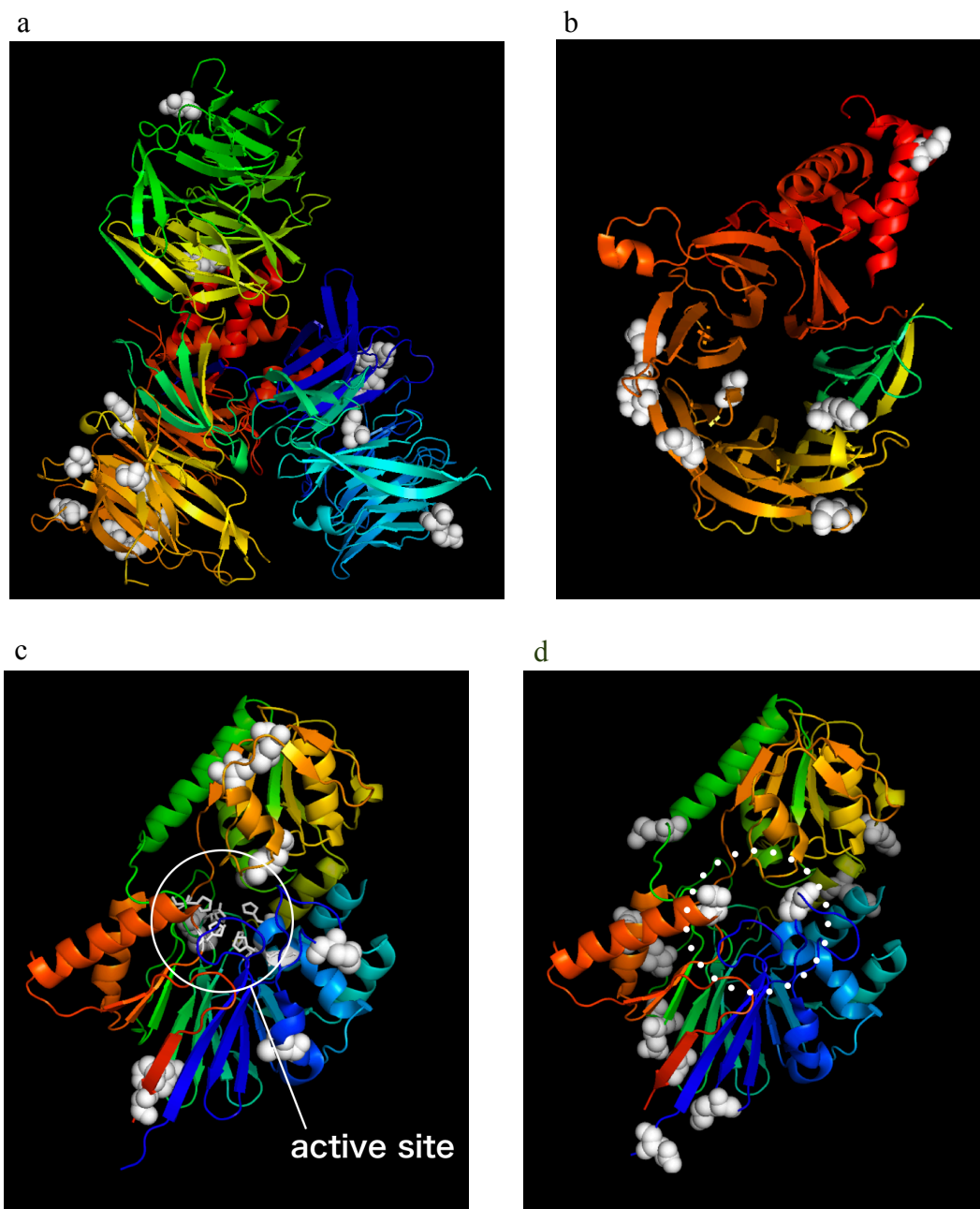


図3-12. HeIタイプの傾向を元に予測した各サブユニットの予測結合残基。白い球で示した位置が予測結合残基位置である。a. 比較に用いた*Homo sapiens* DDB1の立体構造 (PDBID: 3IE1A)上にCPSF160の予測結合残基を表示している。b. 特に結合残基が多く予測されたC末端側のプロペラドメイン。c. CPSF73の予測結合残基。立体構造は既知構造である*Homo sapiens* CPSF73 (2I7VA)。d. CPSF100の予測結合残基。立体構造は既知構造である*Homo sapiens* CPSF73 (2I7VA)。図は Aoto and Yura (2015)より転載。

考察

PDBからサブユニット構成を変化させたと考えられる複合体およびそのパラログを収集したところ、各立体構造においてUniProtのアノテーション、PDBのSUBUNITアノテーションが一致したものは9セットであった。

UniProtのサブユニット構成についてのアノテーションは、明らかにオーソロガスな関係にある分子どうしであっても異なっていることが多く、実際には複合体はサブユニット構成が常に変化しているか、もしくは状況によってサブユニットどうしの相互作用は変化するものであると考えられた。本研究では、PDBに登録されている複合体が実際に機能を発揮できると考えられているサブユニット構成であるかを文献をもとに調べ、機能的単位であると考えられた複合体のみを研究対象とした。

現在までの包括的な結合部位の研究によって、結合部位を構成するアミノ酸残基は分子表面のアミノ酸残基に比べ、疎水的であり、保存性が高いことが示されてきた(Janin et al. 2008)。ホモ多量体においては挿入による結合部位の獲得が見られること(Hashimoto et al. 2010, Nishi et al. 2010)、またホモ多量体の結合部位ではヘテロ多量体の結合部位に比べ、疎水的でなく、結合部位の面積が狭いことが指摘されている(Larsen et al. 1998)。

本研究では、パラログセットにおいて変化がみられた結合部位をもつ複合体をホモ多量体(Homomer)とヘテロ多量体(Heteromer)にわけ、さらに結合部位の変化について invention(新規結合部位の獲得)、switch(結合部位の位置の変化)、add-on(結合部位数の変化)に分類した6グループに分けて解析を行った。パラログどうしで結合残基と同位置にアラインメントされる結合残基ではないアミノ酸残基を比較した結果、結合部位を構成するアミノ酸残基の特徴である、疎水性残基が多く保存度が高いという傾向は、HeI(Heteromer

Invention)タイプに顕著にみられた。また、ホモ多量体で指摘されていたパラログセットで変化のあった結合部位における挿入配列は、ヘテロ多量体でもみられた。さらに本研究では、パラログセットで変化がみられた結合部位において、“hot spot”と考えられる保存性と疎水度の高いアミノ酸残基を特定した。“hot spot”残基は多量体の安定性に有効に働いていると考えられ、HoI(Homomer Invention)タイプに顕著に見られた。Hashimoto *et al.* (2010)や Nishi *et al.* (2010)で既に指摘されているように、特にホモ多量体においてはこれらの挿入部位は結合部位の獲得に重要であるようだったが、ヘテロ多量体ではその寄与は低いようであった。ホモ多量体の結合部位では1残基の挿入は、結合するタンパク質どうしの双方の結合部位が挿入を受けることを意味するため、その影響はヘテロ多量体より大きくなる。このことがホモ多量体とヘテロ多量体の結合部位における挿入残基による傾向の差を生み出している可能性がある。しかし、本研究では9つのパラログセットしか得られず、統計的な解析を行えなかった。そのため、今回得られたホモ多量体とヘテロ多量体と結合部位の特徴の違いは、解析を行ったパラログセットが少ないことが影響して得られた結果である可能性は否定できない。しかし今回の結果から見れば、ホモ多量体の結合部位およびヘテロ多量体の結合部位は明らかに異なる傾向があり、結合部位予測は少なくともホモ多量体とヘテロ多量体の2つの場合に分けて行なう必要がある。

結合部位予測や複合体構造予測の精度の向上は、新たな結合部位の特徴を予測手法に追加することで可能出ると考えられるが、新たな結合部位の特徴を得るためには、相互作用様式を考慮して結合部位の特徴を調べるのが重要となると考えられる。

これらの成果をもとにして、HeIタイプの傾向を利用し、CPSFの各サブユニットがそのパラログにはない結合部位を構成するアミノ酸残基を有しているかの予測を行った。その結果、CPSF73の活性部位付近には結合残基は予測されなかった。一方、CPSF73とホモロ

ガスなタンパク質であるCPSF100では、CPSF73の活性部位にあたる位置にHeIタイプの傾向を持つ結合残基が予測された。CPSF100はエンドヌクレアーゼ活性がないことが示唆されており、CPSF73と分子進化の過程で機能を分けた際に、この部位を結合部位として利用するようになった可能性が考えられる。CPSF160では、C末端側のプロペラドメインにHeIタイプの結合残基が多く予測された。CPSF160の結合残基予測を行なうために比較に用いたCPSF160のパラログの立体構造は、DDB1のものを用いた。DDB1は、C末端側のプロペラドメインとN末側のプロペラドメインで挟み込むようにして、DDB2と結合している。このことから、CPSF160がDDB1と分岐し、DDB2ではない他の分子と結合するようになるにあたって、C末端側のプロペラドメイン部分のアミノ酸残基が大きく変化したと考えられる。

第四章

CPSFサブユニットの複合体予測

背景

多数のサブユニットから形成される複合体の全体構造を予測するとき、どのサブユニットどうしが直接結合しているかを特定することは、最も重要なポイントのひとつとなる(Pierce et al. 2007)。

直接相互作用するサブユニットを明らかにする手法としては、イースト2ハイブリッド法(Fields and Song 1989)、ファーウェスタン法、プルダウンアッセイに代表される免疫沈降法(Dellis et al. 2004, Klupp et al. 2005)が一般的に用いられる(Pierce et al. 2007)。イースト2ハイブリッド法では、タンパク質間結合をイーストの増殖度、もしくは死滅によって見分けることができる。そのためこのアッセイでは、タグなどの本来Yeastに無い分子を強制発現させることによる死滅と、目的とするタンパク質どうしが結合しなかったことによる死滅の見分けが困難である場合がある(Wang et al. 2013)。そのため、プルダウンアッセイや変異挿入などと組み合わせて、相互作用の確認を行う必要がある(Li et al. 2006)。後者の2つに大別されるようなタンパク質のアフィニティーをタグなどで分離する*in vitro*における相互作用解析は*in vivo*の実験系に比べて多量の試料が必要となる(Brymora et al. 2004)。ほかに、ファージディスプレイ法(Mullaney and Pallavicini 2001)、プロテインチップ(Zhu et al. 2001)などによる手法があるが、いずれもタンパク質どうしの結合が直接か間接かを確認するには、複数の実験を組み合わせる必要がある。

cytoplasmic polyadenylationの複合体の構成因子の同定を目的とした先行研究では、プルダウンアッセイ法がもっとも使用されている(Kaufmann et al. 2004, Kyburz et al. 2006)。この方法では、mRNPsやCPSFのように相互に結合しあっている分子は、すべて結合した状態で検出されてしまう。そのため、mRNPsやCPSFのサブユニットを特定することはできて

も、それらのサブユニットが互いに直接相互作用しているのか、間にmRNAや第三のサブユニットが介在する間接的な結合なのかは区別することができない。直接結合しているサブユニットの実験的な検出には、イースト2ハイブリッド法やファーウエスタン法(Bowen et al. 1980)を用いる。

CPSFは複合体としてはよく研究されているが(Dominski et al. 2010, Dominski and Marzluff 2007)、各サブユニットのはたらき、および直接相互作用する相手分子などが不明である。cytoplasmic polyadenylationやnuclear polyadenylationの研究におけるCPSF複合体の免疫沈降法では、CPSF160、CPSF100、CPSF73、CPSF30のうち、どの分子で沈降させてもすべてのサブユニットが検出される(Kaufmann et al. 2004)が、直接結合している分子どうしの特定に有用なイースト2ハイブリッド法やファーウエスタン法を用いた解析は行なわれていない。

そこで本研究では、生命情報学的手法を用いて相互作用する分子の推定を行うこととした。分子系統学における研究により、直接結合するタンパク質分子どうしはその結合によって発揮される分子機能を維持するため、結合部位を始めとしたタンパク質表面が保存されていること(Huynen and Bork 1998, Pellegrini et al. 1999)、また、結合する分子どうしは分子系統樹のトポロジーに相似性があり、分子進化のパターンが似通る(Pazos and Valencia 2001)ことが示されている。これらの原則を利用してCPSFサブユニットのオーソログを解析することで相互作用するサブユニットペアの推定を行なった。

分子系統解析を行うことで、複数のオーソログな分子の情報を得ることができると、複合体構成に関する情報や相互作用相手の分子名や特徴を得ることができる。アミノ酸配列や立体構造に類似性の高いタンパク質のうち、相互作用についての情報がすでに明らかかな分子があれば、同様の相互作用をすると考えることができる(Livstone et al. 2011)。

このようなオーソログを利用した相互作用の予測は、アミノ酸配列の情報を用いる相互作用部位の予測として最も一般的な手法であり(Brenner 1999)、アラインメントと結合残基を比較することで結合する相手分子を予測することも可能である(Aloy and Russell 2002)。この手法はある生物種において見られる分子およびその相互作用は、他の種でも基本的には存在し、機能しているはずであるという考えに基づいている(Livstone et al. 2011)。

分子量が1000kDaを超えるタンパク質複合体は、リボソームやスプライソソーム、プロテアソームに代表され、その多くは遺伝子発現の制御に重要な役割を持つ(Aderem and Ulevitch 2000, Adams 2003, Rice et al. 1999)。このような分子機械のはたらきやその構造変化は、各サブユニットの機能発現や構造変化が連鎖的に起こることで引き起こされる。各サブユニットの機能は、各分子のアミノ酸残基側鎖で起こる化学反応に集約される。そのため、分子機械のはたらきを厳密に理解するためには、その働きを担う複合体立体構造を原子レベルで明らかにする必要がある。近年ではX線結晶構造解析、NMR、原子顕微鏡による複合体構造を実験的に決定する試みが多く行われている(Webb et al. 2015, Nguyen et al. 2015)。

しかし、これらの構造は巨大なだけでなく、低分子や塩基を含むことで複合体として安定して存在している場合もあり(Ramakrishnan 2002)、複合体そのものを実験的に取り扱うことは非常に難しく、構造生物学実験における十分な技量が必要となる。また、サブユニットを単離して行う実験方法は、複合体の分子構成の変化を明らかにするには優れた方法であるが、複合体全体の変化を俯瞰することは非常に困難である。そこで本研究では生命情報学的なアプローチを用いて、母系mRNPsの貯蔵型複合体構造において、翻訳活性化の際に相互作用の変化の中心となると考えられるCPSF複合体の三次元複合体構造の予測を行った。予測したCPSF複合体の立体構造をもとに、そのタンパク質界面の形状やアミ

ノ酸残基の特性を用いてCPSF複合体の周囲に存在するタンパク質分子との相互作用について検討し、CPSF複合体が母系mRNPs中でどのようなはたらきを担っているかを考察した。

CPSFは1つの分子のように振る舞うよう記載されていることも多いが、それ自身が少なくとも4つのサブユニットをもつ複合体タンパク質であり、cytoplasmic polyadenylation, nuclear polyadenylationでのサブユニット構成とヒストンmRNA上ではたらく際のサブユニット構成は異なり、常にすべてのサブユニットが結合した状態で存在しているわけではない(Sullivan et al. 2009)。また、CPSFは母系mRNAのような細胞質での転写抑制を行うmRNAに結合するCPEBに直接結合する。CPEBはmRNAを翻訳する際には複数箇所にリン酸化を受けることが分かっている(Lapasset et al. 2005)。タンパク質のリン酸化はタンパク質表面の電荷を大きく変化させるため、CPEBは母系mRNPsの貯蔵型から翻訳活性型への複合体構造の変化の際に、相互作用を大きく変化させていると考えられる。また、母系mRNPsのポリA鎖伸長は細胞質内で特異的にはたらくpolyA polymerase(PAP)であるGld2が担っているが(Wang et al. 2002)、Gld2もCPSF複合体と相互作用することが示唆されている(Rouhana et al. 2005)。

これらのことから、本研究では母系mRNPsの翻訳抑制機構を解明するにあたり、CPSF複合体立体構造およびCPSF複合体とCPEB、Gld2の三次元的な相互作用部位を原子レベルで予測し、CPSF複合体の母系mRNAの転写抑制におけるはたらきについて考察することとした。

方法

1. 複合体立体構造予測に用いたサブユニット構造

本章では、CPSFサブユニットの分子系統解析、およびCPSF複合体の立体構造予測を行った。その際、*Homo sapiens* CPSFサブユニットの立体構造、およびそのアミノ酸配列を利用した。ここで用いた*Homo sapiens* CPSFサブユニットの立体構造は第一章で行ったコンパラティブモデリングにより予測した立体構造である。

2. 分子系統解析によるCPSFサブユニットの結合ペア予測

2-1. 類似配列検索

CPSFサブユニットの類似配列はBLASTを用いて検索した(Altschul et al. 1990)。

BLASTの中でもアミノ酸配列をクエリとしてアミノ酸配列を検索するblastpを用い、Non-Redundant Database (GenBankおよびProtein Data Bank、SwissProt、PIR、PRFのいずれかに登録されているアミノ酸配列から重複を排除し、一つにまとめたアミノ酸配列データベース)に登録されている全アミノ酸配列に対し検索を行い、E-valueが 10^{-4} 以下のアミノ酸配列のみを収集した。

類似配列検索では、mRNPsの翻訳制御を研究するにあたり、この分野の研究において最も研究が進んでいた*Xenopus leavis*のアミノ酸配列を用いた。しかし*Xenopus leavis*のCPSF160はアミノ酸配列全長が得られていなかった。そのため、アミノ酸配列全長が決定されており、もっともアノテーションが豊富であった*Homo sapiens*のCPSFサブユニットのアミノ酸配列をクエリとして、CPSF160の類似配列の検索を行った。CPSF160の系統樹と、検索条件を等しくするため、CPSF100、CPSF73およびCPSF30については、*Xenopus leavis*

のアミノ酸配列をクエリとした類似性検索で得られた*Homo sapiens*のアミノ酸配列を用い、類似配列の再検索を行った。

2-2. 分子系統樹の作製

Homo sapiens CPSFの各サブユニット160kDa、100kDa、73kDaの類似配列検索より得たアミノ酸配列群から、それぞれのサブユニットの分子系統樹を作製した。タンパク質間結合部位となるタンパク質分子の表面に関する情報を反映させた系統樹を作成するため、前項で収集した重複のないアミノ酸配列に、CPSFサブユニットのモデリングに用いたテンプレート配列を加えて系統樹を作成した。系統樹作製に用いたアラインメントにはALAdGAPsを用いた (Hijikata et al. 2011)。ALAdGAPsは、立体構造を元にギャップペナルティを計算してアラインメントを行うアラインメントプログラムであるが、そのアラインメントに基づき、近隣結合法(Saitou and Nei 1987)によって描画した系統樹も出力する。このALAdGAPsによって出力された系統樹により、相互作用予測の解析を行った。

2-3 分子系統樹を利用した相互作用予測

直接結合するタンパク質分子どうしはその結合によって発揮される分子機能を維持するため、結合部位を始めとしたタンパク質表面が保存されている(Huynen et al. 1998, Pellegrini et al. 1999)、また、結合する分子どうしは分子進化のパターンが似通る(Pazos et al. 2001)、という分子系統学において経験的に得られてきた事象をもとに、CPSF160、CPSF100、CPSF73の分子系統樹のトポロジーを比較することで、どのCPSFサブユニットが直接結合するか予測した。また、各CPSFサブユニットのパラログがどのようなサブユニ

ット構成で、実験的に確認されているか調べ、予測した直接結合するサブユニットペアが機能的であるか判断した。

また本研究では、アミノ酸配列類似性の高いタンパク質どうしは、類似の立体構造、または類似の相互作用部位（活性部位）をもつ、という経験則をもとに系統樹を利用しているため、類似配列であるか、相同配列であるかは明確に区別していない。そのため、パラログかオーソログであるかは、分子系統樹のトポロジーより次のように区別した。オーソログは、明確な分子名のアノテーションが一種類のみである枝に含まれるアミノ酸配列群と定義した。また、パラログは、同一系統樹上の異なるオーソログどうしの関係と定義した。

3. CPSF複合体立体構造の予測

CPSF複合体はCPSF160、CPSF100、CPSF73、CPSF30の4つのサブユニットをもつが、このうちCPSF30は、第一章で行ったCPSFサブユニットの立体構造予測において、配列全長の三分の一の立体構造を予測することができなかった。そのため本研究では、CPSF30を除いた3つのサブユニットの複合体構造予測のみを行なうこととした。ドッキングで用いたZDOCKは1分子対1分子の結合構造を生成するツールであるため、始めに直接結合すると結論づけたCPSF73とCPSF100のドッキングを行ない、つぎにCPSF73とCPSF100の複合体とCPSF160のドッキングを行なうことで、三量体の複合体立体構造予測を行った。候補構造の生成は、CPSF100とCPSF73のドッキング、CPSF100-CPSF73ヘテロ複合体とCPSF160のドッキング、ともに2000体のデコイの生成を行った。これらのデコイに対し、第二章で開発した複合体構造予測手法を適用し、第二章の式(1)を用いたクラスタリング図を作成した。

4. 候補構造の選出

クラスタリングによって尤もらしい構造を含むクラスターを選出することができるが、選出したクラスターに多くの候補構造が含まれている場合がある。そこで、本研究では既知立体構造はエネルギーが低いことを利用し (Callebaut et al. 1999)、クラスター中で最もエネルギーの低い構造を候補構造として選出することにした。

5. 候補構造の評価

5-1. FireDockによるCPSF複合体の作成

予測したCPSF複合体構造が尤もらしい予測結果であるかを評価するため、近年、タンパク質複合体構造予測の国際的なコンテストであるCAPRI(Critical Assessment of PRediction of Interactions)において好成績を残しているFireDock (Andrusier et al. 2007) によって予測した予測複合体構造と比較することとした。FireDockは本研究手法と同様、スコアリングに特化したツールであり、ZDOCKなどの他のソフトウェアによって生成したデコイを、スコアリング、およびランキングすることで複合体構造を予測する。またウェブ上にソフトウェアが公開されており、簡易に予測を行うことが可能である。

FireDockは最大で1度に100個の複合体構造の評価を行なうことができる。しかし、本研究手法では候補構造を2000個作成し複合体の評価を行なったため、本研究で用いた手法とFireDockの予測精度を比較するためには、比較に用いる候補構造を100個に減らす必要があった。そこでクラスタリングを行い、各クラスターから最もエネルギーの低い候補構造を選出することで、比較する候補構造数を100個以下に減らした。各候補構造のエネルギーの低さは、ZDOCKでドッキングを行った際に出力される値を参照した。得られた構造

はFireDockによってスコアリングを行い、その結果に基づいて尤もらしい予測構造を選出し、本研究で得られた予測構造同様、その構造について評価した。

5-2. ホモログな分子どうしの対称性を利用した結合による評価

CPSF73とCPSF100はパラログな分子であり、配列保存性が約25%ある。このような相同な分子同士では、分子の形状やタンパク質間結合部位を始めとした機能部位に類似性があるため、ホモログなドメインでの結合、または複合体全体からみるとサブユニットどうしが対称性のある結合をしている(Goodsell and Olson 2000)。そこで、CPSF73とCPSF100の予測複合体構造においては予測結合残基が存在するドメインの配置を比較し、本予測方法によって得られたCPSF73とCPSF100の複合体構造が対称性のある構造を形成しているか考察した。

5-3. 結合部位上のアミノ酸配列の保存性による評価

CPSFは高度に保存されている複合体であり、少なくとも動物界に属する生物の多くでサブユニット構成が類似であることが確認されている(Millevoi and Vagner 2010)。このような複合体を構成するサブユニットは、複合体構造を保つことによって機能を発揮するため、他のサブユニットと結合ができなくなると機能が十分保持できなくなる。このことは、結合部位での変異はその個体の生存に影響を与える確率が高く、このような変異は集団内に蓄積されないことを意味するため、1つの分子中でも他分子との結合に関与する界面はそうではないタンパク質分子表面に比べ進化速度が遅くなる。そのため一般的には、結合部位は他のタンパク質表面より保存度が高いと考えられている(Valdar and Thornton. 2001)。そこで本研究で予測された複合体構造の表面および結合部位を形成するアミノ酸

残基の保存性を調べた。アミノ酸残基ごとの保存度の算出には、前章で作成した分子系統樹作成時と同様のアラインメントを使用した。前章の系統樹ではパラログな分子も含まれているが、各CPSFサブユニットごとに保存されたアミノ酸残基を特定することが目的であるため、系統樹上でCPSF160、CPSF100、CPSF73のグループに含まれる配列群のみを用いて、各分子ごとに再度アラインメントを行い、アミノ酸残基位置ごとに保存度を算出した。

6. CPSF複合体と結合する分子の結合部位予測

CPSF複合体の立体構造予測と同様に、多くのサブユニットをもつ複合体構造であっても、基本的にはサブユニットを1つずつ追加していくことで、複合体構造全体を予測することが可能である。本研究では貯蔵型母系mRNPsの全体構造に少しでも近づくため、CPSFがmRNA内でCPEB、Gld2およびmRNAとどのように結合しているかを、予測したCPSF複合体構造を使用して、タンパク質結合部位およびRNA結合部位の予測を行った。

CPEBとGld2はCPSF複合体と少なくとも2つ以上のCPSFサブユニットを介して結合していることが分かっている (Laishram and Anderson 2010)。そこで予測CPSF複合体の構造において、CPEBとGld2が2つのサブユニットに結合することが可能かを調べた。

6-1. 結合残基予測に用いた一般的な結合部位の特性

CPSF複合体上のCPEBとGld2の結合部位の予測は、次のような流れで行った。CPSFサブユニットとの結合を満たすような領域のアミノ酸残基のうち、結合部位の特徴である比較的平らな面を形成し、かつ結合部位を構成しやすいアミノ酸残基を特定する。さらに、それらのアミノ酸残基が結合部位を形成するのに十分な面積を形成している領域を検索する

ことで行った。結合部位の総面積は結合している両分子の結合部位を合計して1200~2000Å²、アミノ酸残基数は36-54残基であり (Chakrabarti and Janin 2002, Lo Conte et al. 1999)、その結合面はホモダイマーよりヘテロダイマーの方が比較的平らであるとされている (Jones and Thornton 1996)。タンパク質表面の凹凸はConcave finder:Cx (Concave finder: <http://cib.cf.ocha.ac.jp/bitool/CONCAVE/Cx.html>, Pintar et al. 2002)を用いて特定した。また、KYG-proteinを用いて結合部位になりやすいアミノ酸残基を予測した。これら2つのツールの結果から、KYG-proteinの予測スコアが高く、比較的平らな面を構成しているアミノ酸残基が集中している領域を予測結合部位とした。今回の結合部位予測においては、CPEBは立体構造が得られていないため、実際の結合表面積の算出は行えない。そこで2つのツールの結果から予測結合部位のアミノ酸残基数を計測し、結合部位を構成するに十分な面積を形成しているかを判定した。

6-2. CPSF複合体上のCPEBとの結合残基予測

CPEBはCPSF160とCPEF100と結合することが示唆されており (Papin et al. 2008)、予測CPSF複合体構造においてCPEBがCPSF160とCPSF100に結合するためには、CPSF160上のCPEBの結合部位とCPSF160のCPEB結合部位は近接していなければならない。そこでCPSF160とCPSF100が最も近くに存在することになるCPSF160とCPSF100の結合残基、およびその周辺の結合残基においてCPEBが結合することが可能な領域が存在するかを調べた。

CPEBの立体構造は未だ決定されていない上、立体構造既知のタンパク質中に十分な配列長の類似構造をもつタンパク質も存在しないため、ドッキングにより結合部位の推定を行なうことはできなかった。そこでCPEBの全アミノ酸配列からCPEBの分子量を算出し、

CPEBの体積を推定した。分子量の計算はSwiss Institute of Bioinformatics(SWISS)が提供するサイトExPASyの分子量計算ツール(http://web.expasy.org/compute_pi/)を用いて行った。CPEBは水溶性タンパク質であることから、CPEBを球で近似したときの体積を求め、この球の半径をCPEBの半径とみなした。CPEBと同体積の球の直径から、CPEBがCPSF160とCPSF100に結合できるアミノ酸残基を抽出し、さらにこれらのアミノ酸残基についてKYG-protein、Concarve finder: Cxによるタンパク質間結合部位を構成しやすいアミノ酸残基の予測を行ない、タンパク質間結合部位の平均的な面積を形成している領域を特定し、その領域をCPEBの結合部位として特定した。

6-3. CPSF複合体上のGld2との結合残基予測

Poly A polymerase (PAP) には様々な種類が存在し、細胞質で母系mRNAのポリA鎖伸長を行うはたらくPAPはGld2である(Rouhana et al. 2005)。 *Caenorhabditis elegans*のGld2については、アミノ酸配列全長1113残基のうち364残基の立体構造が明らかにされている(PDBID: 4ZRL, Nakel et al. 2015)。 *Homo sapiens* Gld2の全長は484残基(UniProtKB/Swiss-Prot: Q6PIY7.1)であるが、 *Caenorhabditis elegans*において立体構造が決定されている領域が *Homo sapiens* Gld2の配列とアラインメントできる領域であったため、 *Caenorhabditis elegans*のGld2の構造を基に *Homo sapiens* Gld2の立体構造をモデリングした。

Gld2がCPSF73およびCPSF160の双方と相互作用する(Laishram and Anderson 2010)ためには、CPSF73とCPSF160の距離がGld2の直径と等しいか、それより短いことが条件となる。さらに、結合部位としてはたらくためには、結合部位としてはたらくアミノ酸残基は600~1000Å²程度の表面積を形成する必要があると考えられている(Jones and Thornton 1996)。これは、結合部位を正円と仮定した場合、半径約17Åの領域である。そこで、

CPSF160-CPSF73間の距離が40Å以内の2種類のアミノ酸残基群を第二の予測結合部位候補として抽出した。このとき、CPSF160-CPSF100間の結合残基、またはCPSF100-CPSF73間の結合残基になっているために複合体では露出していないと考えられるアミノ酸残基は、予測結合部位候補から除外した。

さらに、KYG-proteinを用いたタンパク質間相互作用予測、Concarve finder: Cxによるタンパク質表面の平面度予測を行ない、タンパク質結合部位に存在する可能性の低いアミノ酸残基、および凹凸の激しい部分を形成しているアミノ酸残基を排除した。最終的に残った結合残基の候補から、結合部位として機能しうる表面積を形成しているアミノ酸残基の部位を選択し、予測結合部位とした。

6-4. Gld2上の結合残基予測およびCPSFとGld2の結合様式予測

Gld2が貯蔵mRNP複体内でどのように結合しているか調べるために、ホモロジーモデリングによって得た*Homo sapiens* Gld2の予測立体構造上のタンパク質結合部位を予測した。また、この結果をCPSF複合体構造上のGld2の予測結合部位と照らし合わせることで、Gld2がCPSF複合体とどのように結合しているか考察した。

7. mRNAの結合部位予測

CPSFサブユニット複合体はmRNAと直接相互作用し、poly(A) tail付加部位の認識(Zhao et al. 1999)、及びmRNAの3'末端の切断を行う。そこでKYG-RNA(Kim et al. 2006a)を用い、予測CPSF複合体立体構造のmRNA結合部位を予測した。

8. CPSF160における一塩基変異(SNV)検索

タンパク質間結合部位にその組成が大きく変化するようなアミノ酸置換が起こると、その分子は結合しにくくなるだけでなく、分子機能を失う可能性がある (Fryxell 1996)。進化医学の領域においては、結合部位には結合親和性が大きく変化するようなアミノ酸置換はない、もしくはそのようなことが起こった場合、致死などの表現型として現れると考えられるのが一般的である (Kumar et al. 2011, David et al. 2012)。CPSF160はCPSFサブユニットのCPSF100、CPSF30と結合するだけでなく、CPEB、Gld2、mRNAとも結合することが示唆されている。またCPSF160は3つのプロペラドメインで構成されているが、このプロペラドメインを構成するWDドメインは足場タンパク質としてはたらく分子に多く見られる配列である(由良 and 高井 2013)。これらのことから、CPSF160のタンパク質表面は相互作用部位としての機能にすぐれた性質を有していると考えられる。そこで、*Homo sapiens* CPSF160の遺伝子上に見つかったnon-synonymous SNVは、CPSF160のタンパク質間結合にどのような影響を与えているかを調べた。CPSF160のnon-synonymous SNVはdbSNPデータベースより収集した。dbSNPはGenBank等のゲノム情報に対しSNVについての情報を補うように作成され、NCBIから提供されている世界最大のSNVデータベースである(Sherry et al. 1999)。得られたCPSF160のnon-synonymous SNVは、そのSNVがタンパク質表面に影響を与えるかを確認するため、CPSF160の予測立体構造上で、その変異がどこにアミノ酸残基置換を起こすかを特定した。CPSF160の結合部位に影響を及ぼすようなSNVであるかを類推するため、結合部位に位置するレファレンス配列のアミノ酸残基とnon-synonymous SNVによって、別の別のアミノ酸残基に置換されたアミノ酸残基について、KYG propensity値の変化について調べた。KYG propensityとは、タンパク質間結合部位の予測ツールであるKYG-proteinが予測の際に用いる、各アミノ酸残基種がどのくらいタンパク質結

合部位に存在しやすいかを算出したものであり (Yura and Hayward 2009)、この値が高ければ高いほど、タンパク質間結合部位を形成しやすい性質をもつアミノ酸残基であると推測できる。タンパク質結合部位となりやすいアミノ酸残基は、一般的に疎水性アミノ酸残基などである。このような性質のアミノ酸残基がタンパク質表面に存在すると、水分子と近づくことを避けるため、近くにいる比較的疎水性の物質つまりタンパク質と結合しやすくなる、もしくはタンパク質と離れることは水分子と近づくことになるので、それを避けるために結合が強固になると考えられる。立体構造作成に用いた *Homo sapiens* CPSF160 のアミノ酸配列の残基、および non-synonymous SNV によりアミノ酸置換が起きた残基について KYG propensity の値を求め、その値の変化から SNV によって生じる CPSF160 表面の性質の変化を数値化した。KYG propensity の標準偏差は 0.05~1.0 の間の値 (平均 0.06) をとることから、SNV によって生じる KYG propensity の変化が 0.1 以上あるものをタンパク質結合親和性が変化した SNV とし、0.4 以上の変化を極端にタンパク質結合親和性が変化した SNV とした。

9. CPSF100 の天然変性領域の機能予測

第一章におけるモデリングの際、CPSF100 の C 末端側から 80~180 残基の領域に約 100 残基の比較的大きな非構造領域が予測され、この予測非構造領域を含む C 末端側の約 180 残基の立体構造を予測できなかった (図 1-5.b)。

そこで CPSF100 の C 末端領域を球で近似し、予測した CPSF 複合体立体構造上のどこ位置するかを予測を行った。また、CPSF100 の C 末端領域が存在すると予測される部位と予測 CPSF 複合体立体構造上のタンパク質間結合部位の位置を比較することで、CPSF100 の C 末端領域の機能について考察した。

結果

1. CPSFサブユニットの類似配列検索

分子系統樹のトポロジーが似通るCPSFサブユニットどうしが存在するかを調べるため、まず、各CPSFサブユニットをクエリとして、全生物種のアミノ酸配列を対象とした類似配列の検索を行った。

1-1. CPSF160の類似配列検索

Homo sapiens CPSF160をクエリとして類似性検索を行った結果、590本の配列を得た(表4-1)。その中に*Xenopus leavis*の配列も含まれていたが、アノテーションは無く、363残基と160kDのタンパク質を形成するには配列長が短いものであった。

1-2. CPSF100の類似配列検索

CPSF160で*Homo sapiens*の配列を用いたため、*Xenopus leavis* CPSF100の類似性検索で得られた*Homo sapiens* CPSF100のアミノ酸配列を用い、類似配列の検索を行った。その結果、2420本のアミノ酸配列を得た(表4-2)。CPSF100のパラログとしてはInt9が取得された。また、この類似配列検索の結果には、“CPSF73”という分子名のアノテーションが付いたアミノ酸配列が含まれていた。

表4-1. CPSF160の類似配列検索に用いたクエリ、および類似配列検索により得られた類似アミノ酸配列の本数

クエリ	ID	類似アミノ酸配列 (本)
<i>Xenopus leavis</i> CPSF 160kDa subunit (LOC3989231 protein, partial)	AAH60475.1	241
<i>Homo sapiens</i> CPSF 160kDa subunit	Q10570.2	590

表4-2. CPSF100およびCPSF73に用いたクエリ、および類似配列検索により得られた類似アミノ酸配列の本数

クエリ	ID	類似アミノ酸配列 (本)
<i>Xenopus leavis</i> CPSF 100kDa subunit	NP_001081123	1146
<i>Xenopus leavis</i> CPSF 73kDa subunit	NP_001088278.1	2133
<i>Homo sapiens</i> CPSF 100kDa subunit	Q9P2I0.2	2420
<i>Homo sapiens</i> CPSF 73kDa subunit	Q9UKF6.1	1459

1-3. CPSF73の類似配列検索

CPSF100と同様、*Xenopus leavis*の類似配列検索で得られた*Homo sapiens* CPSF100のアミノ酸配列をクエリとして類似配列の検索を行い、1459本のアミノ酸配列を得た(表4-2)。この結果、Int11、Archaeal CPSFをパラログとして得た。また、この類似性検索の結果には、CPSF100のアミノ酸配列も含まれていた。これらの結果から、CPSF100とCPSF73はパラログであることが分かった。ここで収集したアミノ酸配列を用いて行う分子系統樹解析では、パラログの情報からCPSFサブユニットの分子機能を探ることを目的としているため、CPSF100およびCPSF73は同一の分子系統樹として作製した。

1-4. CPSF30の類似配列検索

Homo sapiens CPSF30は、全長269残基に5つのzinc finger配列と1つのzinc knuckle配列をもっている。これらのzinc finger配列とzinc knuckle配列は多くのRNA結合分子に保存された配列であり、収斂進化の結果、祖先配列を共通としないRNA結合分子にも広く散見される。CPSF30の全長269残基の半分以上が、zinc fingerモチーフとzinc knuckleモチーフを構成する。そのため、CPSF30をクエリとした類似配列検索では大量のアミノ酸配列が得られたものの、全てzinc finger配列部分またはzinc knuckle配列部分での類似性をもとに取得した

配列が多く、CPSF30にホモログなアミノ酸配列を選択することが困難であった。そのため、CPSF30をクエリとして収集したアミノ酸配列を用いても、正確な系統関係を示す分子系統樹を描くことが困難と判断し、CPSF30については分子系統樹の作製は行わないこととした。

2. CPSFサブユニットのアラインメントおよび分子系統樹の作成

分子進化のパターンが似通っているか調べるには、分子系統樹のトポロジーを比較することが最も簡易であるため、各CPSFサブユニットの分子系統樹を作成した。

2-1. CPSF160の分子系統樹作成

CPSF160は約1400残基のタンパク質であるが、類似性検索で得られた配列には300残基程度の配列なども含まれていた。これらはCPSF160の部分配列である可能性が高いが、配列長があまりに異なるアミノ酸配列を同時に系統樹として作図すると、進化距離が正確に計算できない可能性がある。そこで、1000残基以下のアミノ酸配列を除去した。また、同様の理由により、1500残基以上の配列も除いた上でアラインメントを作製した。最終的に系統樹作成に使用したアミノ酸配列は429本であった。

2-2. CPSF100およびCPSF73の分子系統樹作成

CPSF100とCPSF73は互いにパラログな分子であったため、同一の系統樹として作成する必要があったが、類似配列検索の結果から得られたアミノ酸配列の本数が多かったため、まず別々の系統樹として作成し解析を行った。この時、配列長が明らかにCPSF100かCPSF73より短い配列と長い配列は、それぞれ部分配列または遺伝子重複やオルタナティ

ブスプライシングによる転写産物と考えられたので、500残基以下および1000残基以上の配列は系統樹作成に用いるアラインメントから除去した。またBacteria由来のアミノ酸配列が作製した分子系統樹の様々な分岐位置に散見されたため、系統進化を判断しやすくするためにBacteria由来の配列をすべて除去した。その結果、CPSF73とCPSF100の類似配列の総数1189本中582本が同一の配列であった。

別々に作成したCPSF100とCPSF73の系統樹の結合は、次のように行った。CPSF73およびCPSF100の系統樹から、重複するパラログ部分として、Archaeal CPSFの枝に相当するアミノ酸配列を特定した。これによりCPSF73およびInt11、CPSF100とInt9、Archaeal CPSFの3つのアミノ酸配列セットを得た。この配列をもとに、CPSF73およびInt11、CPSF100とInt9、Archaeal CPSFの3つの系統樹を作成した。これらの3つの系統樹には、それぞれひとつ以上の立体構造のアミノ酸配列が含まれていたため、各系統樹の代表となる立体構造のアミノ酸配列をアラインメントすることで各系統樹間の距離を算出し、この距離をもとに3つの系統樹を結合した。このとき各系統樹を代表する立体構造として、CPSF73およびInt11の系統樹は*Homo sapiens* CPSF73の立体構造 (PDBID: 2I7V)、CPSF100とInt9の系統樹はCPSF100のYeast ホモログであるYDH1の立体構造 (PDBID: 2I7X)、Archaeal CPSFの系統樹はArchaeal CPSFの立体構造 (PDBID: 2XR1) を使用した。

さらに3グループに含まれている既知立体構造である*Homo sapiens* CPSF73、YDH1、Archaeal CPSFの3本のアミノ酸配列から系統樹を作成し、この系統樹をもとにCPSF73およびInt11の系統樹、CPSF100とInt9の系統樹およびArchaeal CPSFの系統樹を結合し、1つの系統樹とした。

最終的にCPSF100を主に含む枝で242本、CPSF73を含む枝で239本、Archaeal CPSFの枝で126本の配列を含む全607本のアミノ酸配列の系統樹を得た。

3. CPSFサブユニットの分子系統解析

分子系統樹のトポロジーを比較するため、各CPSFサブユニットが、そのパラログとどのような分岐をしているか、また、どのようなパラログをもつかを調べた。

3-1. CPSF160の分子系統樹

CPSF160の分子系統樹を図4-1に示す。CPSF160の分子系統解析の結果、CPSF160のパラログとして、Splicing Factor 3B subunit 3 (SF3B3)、cleavage factor 2 protein 1 (CFT1)、DNA damaged binding protein 1 (DDB1) が取得できた。Archaea及び、Bacteriaの配列は存在せず、CPSF160は真核生物のみが有するタンパク質であることがわかった。

SF3B3、DDB1、CFT1は、CPSF複合体同様、塩基に結合するタンパク質複合体の構成分子としてはたらくが、これらの複合体のサブユニットには、CPSF複合体のCPSF160以外のサブユニットとアミノ酸配列や立体構造が類似の分子はなく、直接結合している可能性の高い分子を推測することはできなかった。

3-2. CPSF100およびCPSF73の分子系統樹

CPSF100とCPSF73の分子系統樹には、Integrator complex subunit 9 (IntS9)、Integrator complex subunit 11 (IntS11)のグループが見られた。IntS9はCPSF100と、IntS11はCPSF73と、より配列類似性が高く (Dominski et al. 2005)、それらの枝とは別に、Archaeal CPSFのグループが見られた(図4-2)。動物界、植物界、菌界のそれぞれで、CPSF73およびCPSF100のそれぞれに対応するグループが見られたが、Archaeaでは一つのグループしか現れなかった。しかし、Archaeal CPSFはホモダイマー化する可能性が示唆されており (Mir-Montazeri

et al. 2011)、IntS11とIntS9はヘテロダイマーとなることでIntS11のエンドヌクレアーゼ活性の活性化もしくはsnRNAとの結合に関与していると示唆されている(Albrecht and Wagner 2012)。このことから、CPSF100とCPSF73は結合している可能性が高いと結論づけた。

また、系統樹によるCPSF73とCPSF100の分類により、XP_002174386.1、YP_023465.1、ZP_05570330.1の3つの配列のアノテーションが誤りであることが分かった。

XP_002174386.1はCPSF73とのアノテーションがあつたが、CPSF100の枝に分類された。

XP_002174386.1はMBLスーパーファミリーの分子としてほぼ全長と考えられる786残基の配列であるが、CPSF73に高度に保存されているはずのHXHXDHドメインを保持していなかった。これはCPSF73の機能であるエンドヌクレアーゼ活性を有していない可能性が高いことを示すため、CPSF73とのアノテーションは誤りであると考えられる。YP_023465.1

とZP_05570330.1はCPSF100とのアノテーションがなされていたが、Archaeal CPSFの枝に分類された。Archaeal CPSFはその特有のドメインとしてN末端側にKHドメインを有していることが分かっている。YP_023465.1およびZP_05570330.1の配列は共にKHドメインを持っており、Archaeal CPSFと考えられる。ZP_05570330.1の配列は、現在はNCBIのデータベースからその登録が削除されている。また、YSH1とアノテーションされたがIntS11の枝に見られたXP_002427672.1、CPSF73とアノテーションされたがIntS11の枝に見られた

XP_001862054.1、CPSF73とアノテーションされたがYSH1の枝に見られたXP_002373170.1の3つの配列についてもアノテーションが誤っている可能性があるが、系統樹上でごく近傍に位置するアミノ酸配列がアノテーションがないため、断定できなかった。

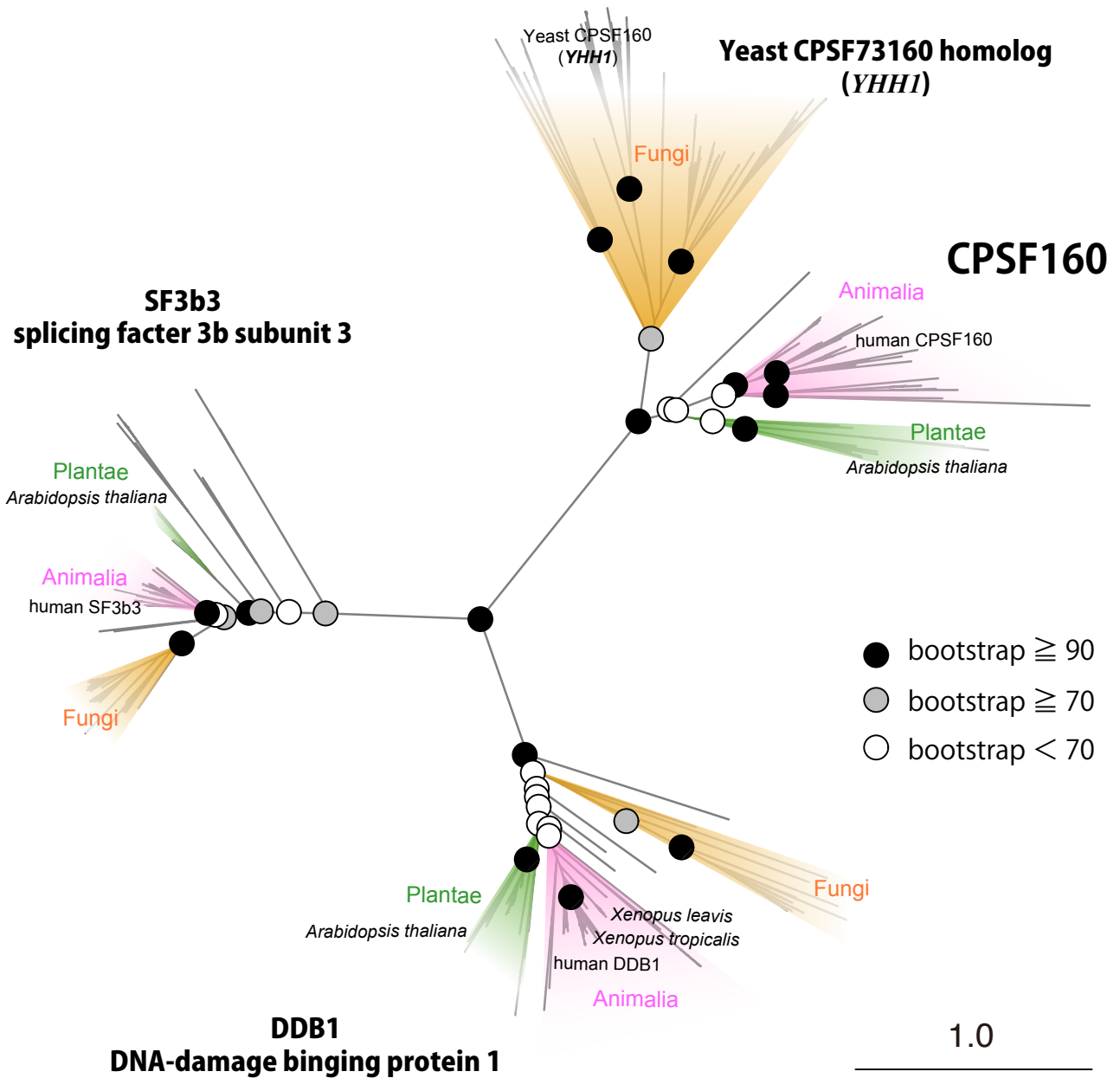


図4-1. CPSF160の分子系統樹

CPSFのパラログとして、SF3b3およびDDB1がみられた。Bacteria、Archaeaの配列は見られなかった。図中のバーは、1座位あたりのアミノ酸残基の置換数を示す。rotocistaの配列は、固まった枝としてみられなかったため、表示していない。

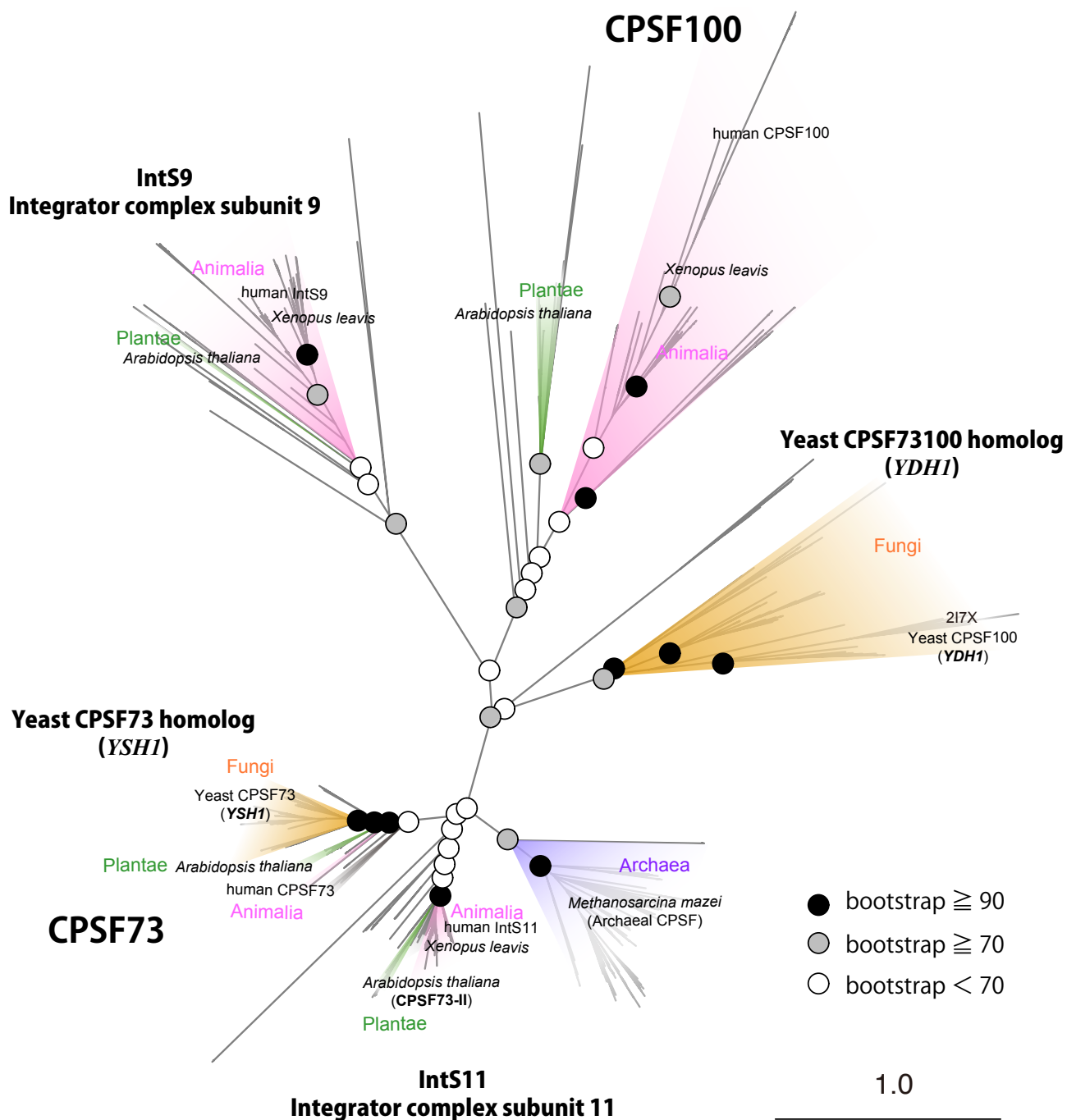


図4-2. CPSF100およびCPSF73の分子系統樹

CPSF73とCPSF100はパラログであったので、同一系統樹として作成した。図右下のバーは、1座位あたりのアミノ酸残基の置換数を示す。CPSF100に最も近いパラログとしてInts9が、CPSF73に最も近いパラログとしてIntS11が得られた。またこの系統樹では、Archaeaのひとつの枝としてみられた。またBacteria、Protoctistaの配列は、固まった枝としてみられなかったため、表示していない。

4. 候補構造の選出

CPSF複合体の立体構造の予測結果では、クラスタースコアが高いクラスターが3つ存在し、それらのスコアの値が近かったため、1つのクラスターに絞り込むことができなかった。そこで、クラスタースコアの高かった3つのクラスターに含まれる候補構造の中で最もスコアが高い構造をCPSFサブユニットの予測立体構造とした。

三量体以上の複合体である場合、サブユニットどうしは直接結合していない場合がある。ドッキングの行程では直接結合できるかどうかに関わらず結合構造を作成するため、直接結合するサブユニットどうしを選択してドッキングを行う必要がある。前章の分子系統樹解析により、CPSF73およびCPSF100はCPSF160に比べて幅広い種に保存されている分子であり、CPSF73とCPSF100のパラログがヘテロダイマーを形成していることから、CPSF73とCPSF100も直接結合していることが示唆された。よって、CPSF73とCPSF100を最初にドッキングし、次に1回目のドッキングで得られたCPSF73-CPSF100ヘテロ二量体とCPSF160をドッキングした。2度のクラスタリングの結果(図4-3. a, b)、および最終的に選出したCPSF複合体のモデル構造を図4-4に示す。

5. 候補構造の評価

候補構造の妥当性の評価は、特にCPSF73とCPSF100の結合に着目し、次の3つの手法により行った。

5-1. FireDockによるCPSF複合体の作成

本研究手法によって予測したCPSF複合体構造を図4-5.aに、FireDockによって予測したCPSF複合体構造を図5.bに示す。また、CPSF100およびCPSF73の結合残基を対象に、本研

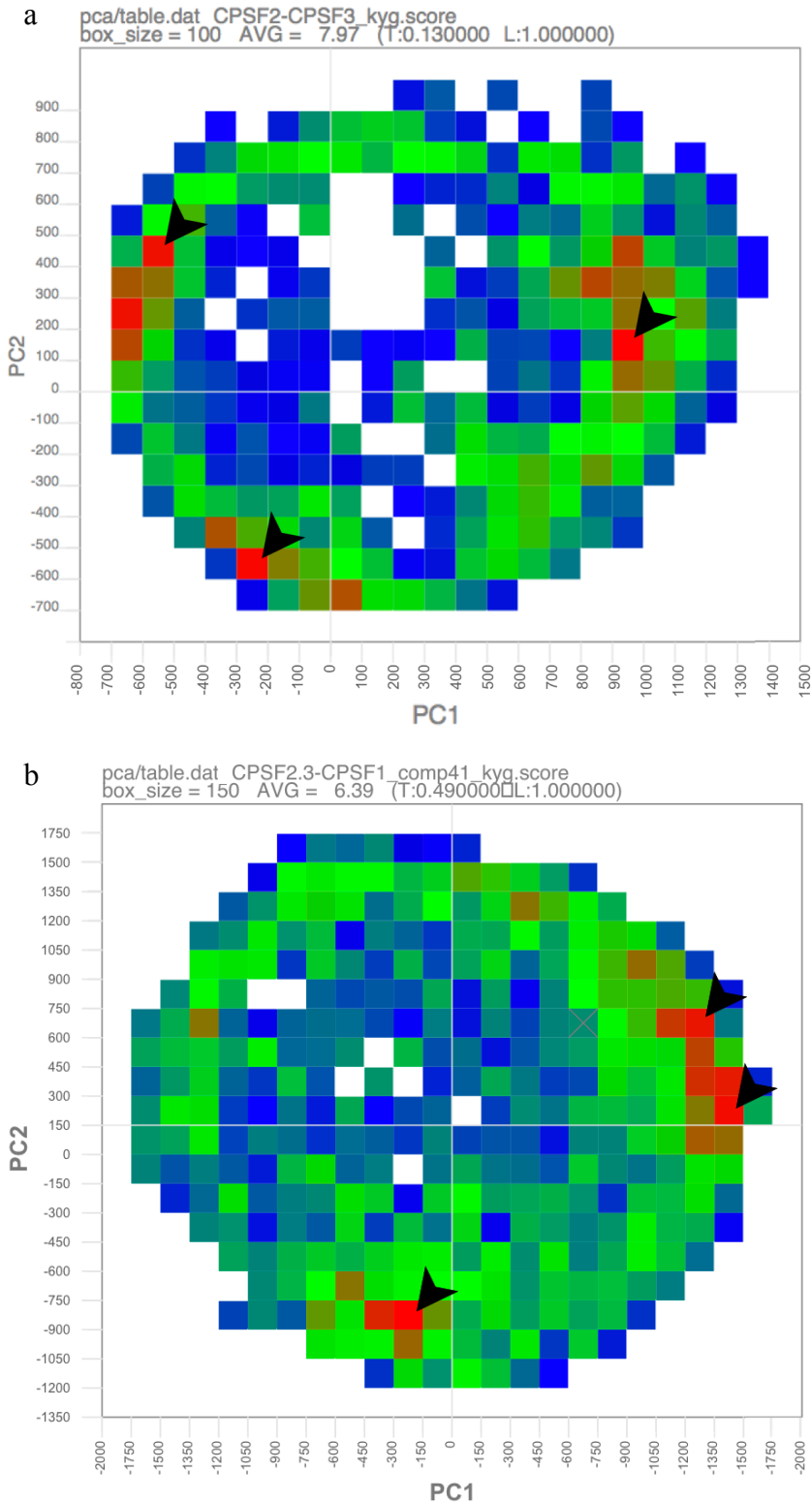
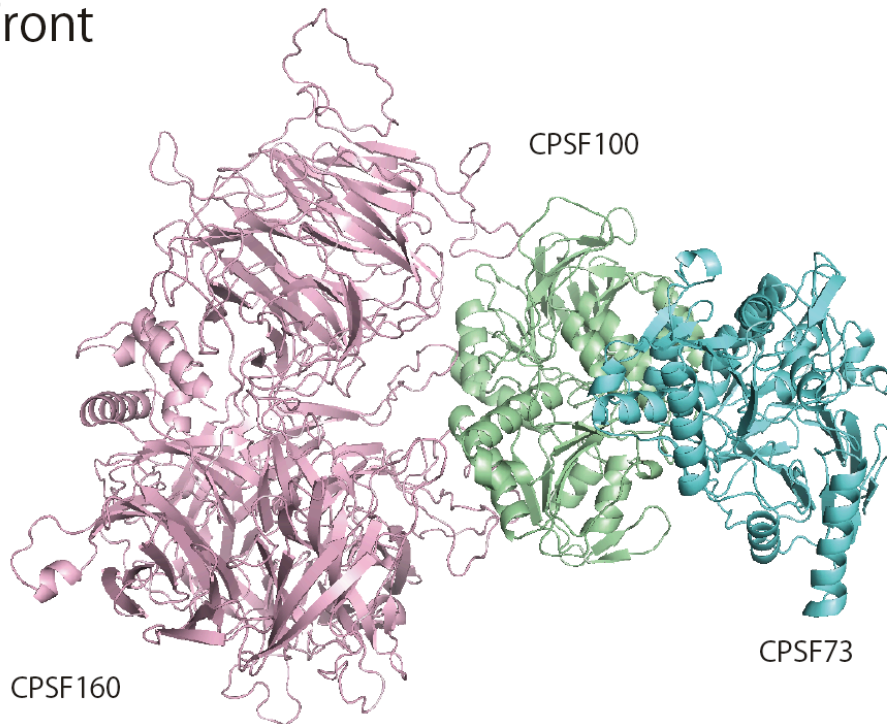


図4-3. 本研究で開発した複合体予測手法によるCPSFサブユニットの予測結果の出力。
 a. CPSF100とCPSF73の候補複合体構造のクラスタリング結果。b. 図1で選出したCPSF100-CPSF73複合体とCPSF160の候補複合体構造のクラスタリング結果。予測スコアの高いクラスターを赤、予測スコアが低いクラスターを青で、デコイが含まれないクラスターを白で示している。クラスタースコアが高い3つのクラスター(矢頭で示したクラスター)からKYG-proteinスコアが最も高い構造を予測構造として選出した。

Front



Top

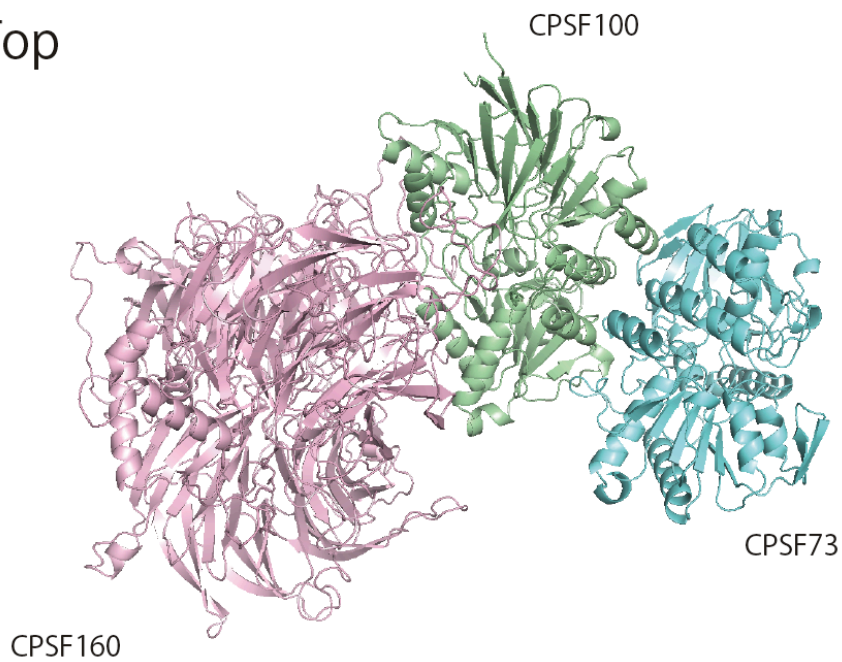


図4-4. *Homo sapiens* CPSF複合体の予測立体構造

予測したCPSF複合体構造では、エンドヌクレアーゼ活性をもつCPSF73と、mRNA認識を行うとされるCPSF160は、機能未知とされるCPSF100を介して結合していた(Front)。またその結合は、U字に似た形状を取っており、狭い領域で複数の分子が近接して結合できるような構造であった(Top)。

究手法およびFireDockの予測残基を比較した結果を図4-5. dに示す。赤の矢じりは本研究で予測した複合体構造の予測結合残基を、青の矢じりはFireDockで予測した複合体構造の予測結合部位を示し、二つのツールで予測したCPSF複合体構造は全く異なって見えるが、結合部位となっているアミノ酸残基の多くが一致している。

5-2. CPSF73とCPSF100のホモロガスな結合による評価

CPSF73とCPSF100は対称軸のある結合をしていた(図4-6)。CPSF73のオーソログは全生物を通して保存されているが、特に高度に保存されているのは配列の中央部に位置する β -Caspドメインである(Mir-Montazeri et al. 2011)。本研究で得た予測構造では、 β -Caspドメインに該当する領域にCPSF73とCPSF100の結合残基が予測された。AnimaliaやFungiでは、CPSF100やCPSF73とともに、C末端側に他のMBLファミリーの分子にはない配列をもっているが、Archaeal CPSFではこの配列はなく、むしろN末端側に長い配列をもつ。CPSF73とCPSF100のC末端領域およびArchaeal CPSFのN末端領域は、mRNAとの結合親和性を高める機能があると考えられており(Albrecht and Wagner 2012)、CPSF100とCPSF73間の結合部位を中心としたとき、CPSF73とCPSF100のC末端ドメインはその外側に位置すると考えられている(Silva et al. 2011, Albrecht and Wagner 2012)。予測したCPSF73-CPSF100のヘテロダイマーは、CPSF100とCPSF73の結合部位を中心とした時にC末端とN末端の両端を外側にした複合体構造を取っていた。

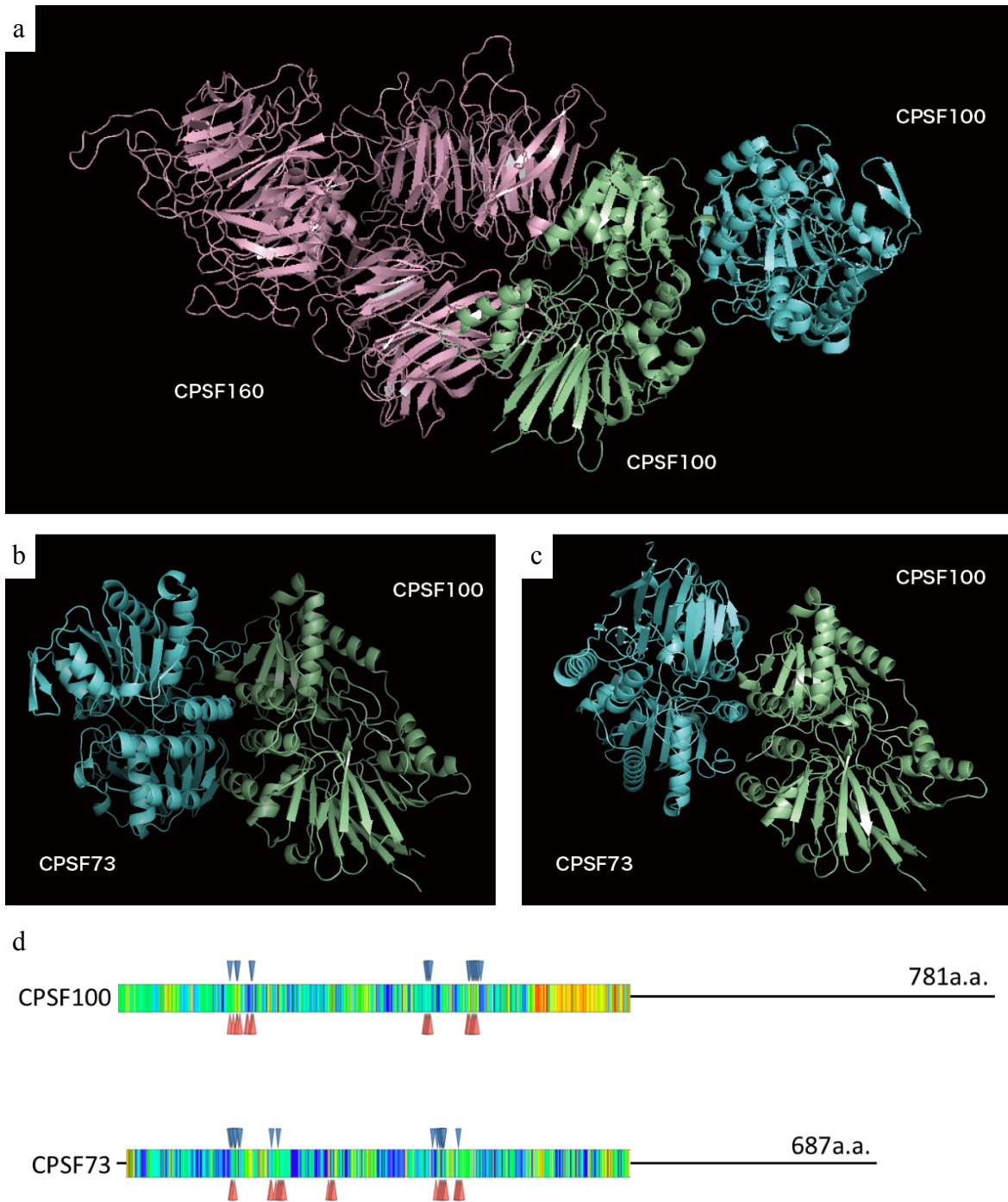


図4-5. 本研究で開発した複合体予測手法とFireDockによる複合体予測手法の比較
 a. FireDockを利用して作成した予測CPSF複合体構造。DDB1と似た結合様式をとる。CPSF73とCPSF100と二量体構造を用いて、本研究で開発した予測複合体構造 (b) とFireDockによる予測構造 (c) と比較した。d. CPSF73とCPSF100の予測結合残基の比較。青い矢尻はFireDockの予測した結合残基を、赤い矢尻はKYG-proteinで予測した結合残基を示す。予測した複合体立体構造は全く異なって見えるが、結合部位として選ばれているアミノ酸残基の位置はほぼ一致しており、同一ドメインの結合を予測していることが分かる。

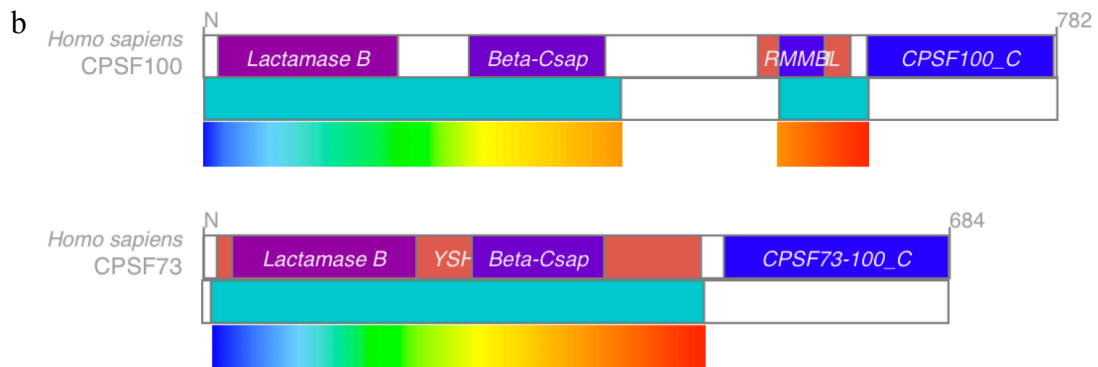
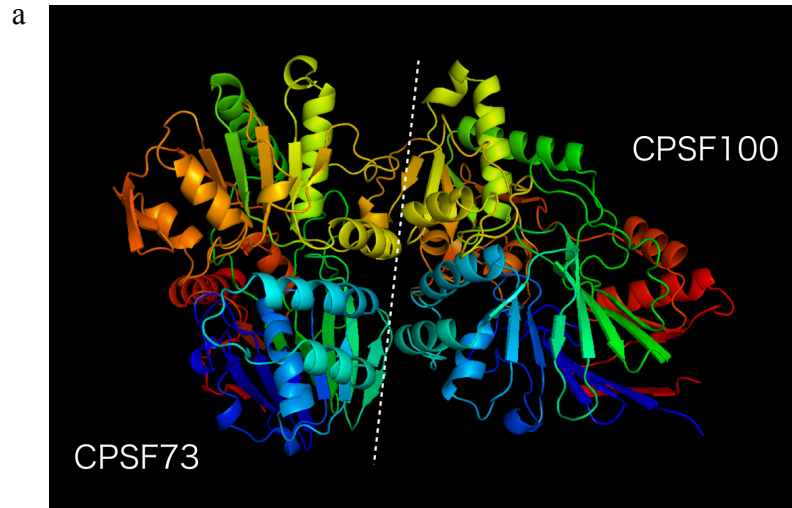


図4-6. 本予測構造におけるCPSF73とCPSF100の結合様式

a. 図2で示したCPSF複合体構造を下から見たときのCPSF73-CPSF100複合体構造。左がCPSF73、右がCPSF100。それぞれの分子骨格はN末端からC末端にかけて青~赤で示している。b. CPSF100およびCPSF73のドメイン構成(上)、モデリングできた領域(中:緑色で示す部分)、aの予測立体構造における部位とその色(下:レインボー)。

5-3. 結合部位上のアミノ酸配列の保存性による評価

CPSF100、CPSF73において、この2分子間の結合部位と予測されたアミノ酸残基は、特に周囲のタンパク質表面のアミノ酸残基と比較して保存されていた(図4-7)。また、CPSF160は一見保存性が低く見えるが、CPSF160のタンパク質表面全体の保存度がCPSF160の他のパラログと比較しても保存性が低く、タンパク質結合部位と予測された領域は、周囲のアミノ酸残基に比べ、保存度が高いことが分かった。

表4-3. CPEB、Gld2およびmRNA結合部位を特定するために使用した項目と予測構造の比較

論文から抜粋した条件	予測に用いた値	予測構造との比較	
		FireDock (図4-4)	本予測手法 (図4-5a)
mRNA前駆体上ではAAUAAA配列とCA部位間は約10~30 ntであり、それぞれCPSF160およびCPSF73に結合する。 Murthy and Manley 1995	1nt(リン酸結合部位から隣のリン酸結合部位までとした)は6.0~7.0ÅなのでAAUAAA及びCA間は210Å以下。CPSF160-CPSF73間の距離(d)も210Å以内でなければいけない。	d = 70~90Å ◎	d = 70~90Å ◎
PolyA polymerase(PAP)の一つであるGld2はCPSF160及びCPSF73と直接結合する。 Laishram and Anderson 2010	Gld2と予測構造の類似性の高いYeastのPAP(PDB : 2Q66)の短軸は約30Åであり、CPSF160とCPSF73間の距離(d)が25Åであれば左記条件を満たすものとする。	d = 0Å *CPSF160とCPSF73は直接結合している ◎	20Å ◎
CPSF100はCPE結合タンパク質であるCPEBと結合する。CPEはmRNA前駆体上のUUAUUUを認識して結合する。UUAUUUはAAUAAAから30 nt以内の近傍に存在する。 Papin et al. 2008	CPEBの立体構造は未だ決定されていないが、そのアミノ酸配列の分子量から同体積の球を作成することができる。その球の半径を測ったところ約26オングストロームであった。CPSF160の配列認識部位とCPSF100の距離(d)がCPEB直径と同程度の長さ、52Å程度であれば左記条件を満たす可能性が高い。	d > 70Å △	d < 25Å ◎

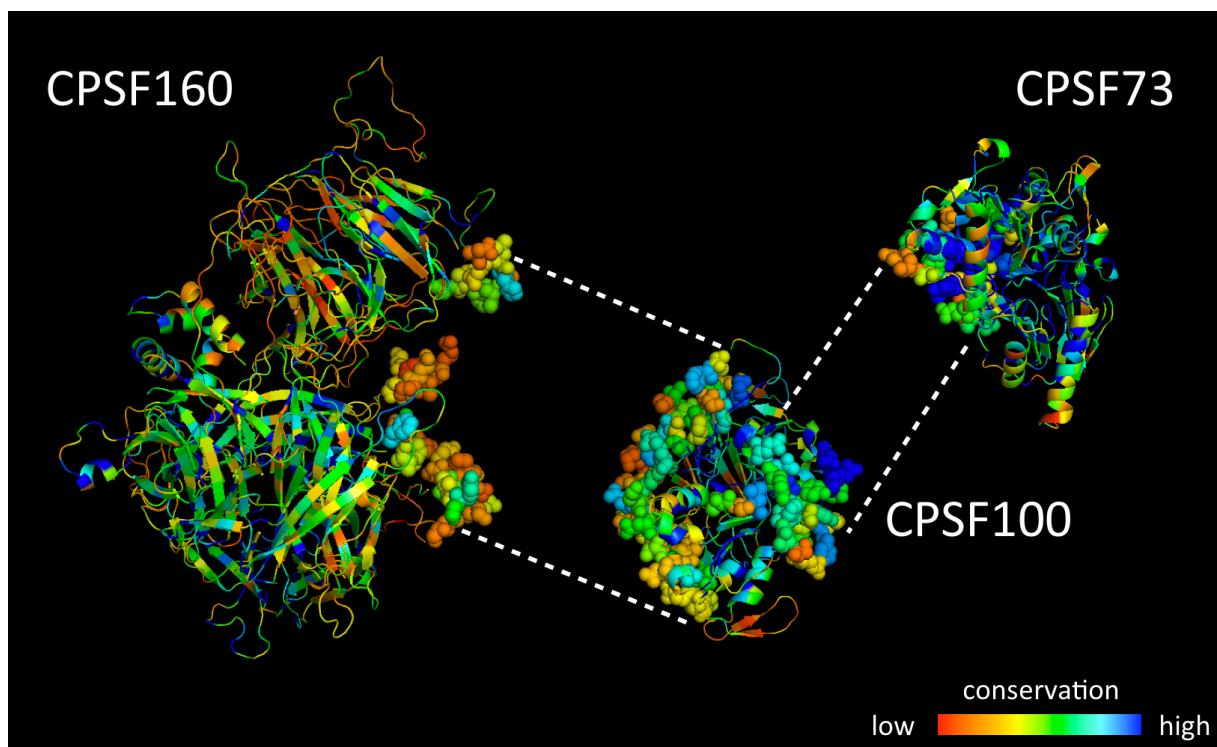


図4-7. 本研究で予測したCPSF複合体構造における結合部位のアミノ酸配列保存度
 CPSF予測複合体構造における結合部位。結合部位を構成するアミノ酸残基の原子を球で示し、結合する部位どうしを線で結んで示した。立体構造上のアミノ酸残基の色は、そのアミノ酸残基の保存度を示し、保存性の高いアミノ酸残基を青で示している。

6. CPSF複合体とCPEBおよびGld2の結合部位予測

CPSF複合体が周囲の分子とどのような相互作用を行っているか、予測するため、CPSF複合体の立体構造を用いて、CPEBとGld2がどのように結合するか、三次元空間における制限を利用して予測した。

6-1. CPEBの結合部位

CPEBの配列全長の立体構造が得られなかったため、CPSBと同体積の球の半径を求めたところ、CPEBと同体積の球の半径は約26Åであった。平均的なタンパク質間結合部位のサイズは600~1000Å²であることから(Chakrabarti and Janin 2002)、仮にCPSF複合体上のCPEBの結合部位を1000Å²の正円であるとする、その結合部位の半径は約17Åになる。そこで、CPSF160とCPSF100の結合面から15Å以内のアミノ酸残基および20Å以内のアミノ酸残基を抽出した。さらに、表4-4に示したCPSFサブユニットとの結合条件とConcave finder: CxおよびKYG-proteinにより予測した結合残基から、結合残基の絞り込みを行ったのち、結合部位を構成する平均結合残基数を利用して結合部位を特定した(表4-4)。

表4-4. CPSF複合体上のCPEB、およびGld2の予測結合部位を構成するアミノ酸残基数

ターゲット分子	CPSF複合体中の パートナー分子	パッチと考えられる領域に 存在していたアミノ酸残基数 (残基)
CPEB	CPSF160	20
	CPSF100	23
PAP(Gld2)	CPSF160	24
	CPSF73	23
janin, Pinakら	—	ターゲットとパートナー合わせて36~59残基

CPEBの予測結合部位では、CPSF160 から飛び出たループ部分がCPSF100との相互作用部位とCPEBとの相互作用部位を構成していた (図4-8. a (i, ii))。このCPSF160上の予測結合部位とされるループを構成するアミノ酸残基は、KYG-proteinの予測において高い予測値を示した(図4-8. a (iii))。一方、CPSF100ではCPSF160と比較するとKYG-proteinの予測値が低かったが、CPSF100は分子表面が全体的に予測値が低く、CPEBとの結合部位と考えられるこの部分は、他の部分と比較して高い予測値を示していた (図4-8. a (iv))。

CPEBは母系mRNAに結合する分子であり、そのmRNA上のCPEB結合部位であるCPEとCPSF160の結合部位であるHexは30nt以内に位置していなければ、翻訳抑制が行なわれないことが示唆されている(Piqué et al. 2008)。CPSF複合体を対象にしたRNA結合部位予測では、二カ所のRNA結合部位がCPSF160上に予測された(図4-9. a)。この2つの予測RNA結合部位とCPEBの予測相互作用部位の距離を測ったところ、それぞれ最短距離が22.2Å、76.7Åであった (図4-9. b)。

いくつかの母系mRNAでは、CPEB結合部位であるCPE配列とCPSF160のmRNA認識部位であるAAUAAA配列がオーバーラップしているものもみられる。このような母系mRNAであった場合、CPSF160上のmRNA認識部位とCPEB結合部位は、接するほど近くなければ同時に結合することができないことが予測される。貯蔵mRNPs中では、mRNAはCPEBおよびCPSFと結合していると考えられているため (Piqué et al. 2008, Radford et al. 2008)、本研究ではCPSF160上のmRNA認識部位とCPEB結合部位は非常に近接して存在していると考え、予測CPEB結合部位から22.2Åの距離にあるRNA予測結合部位をCPSF160のmRNA認識部位とした (図4-9. b)。

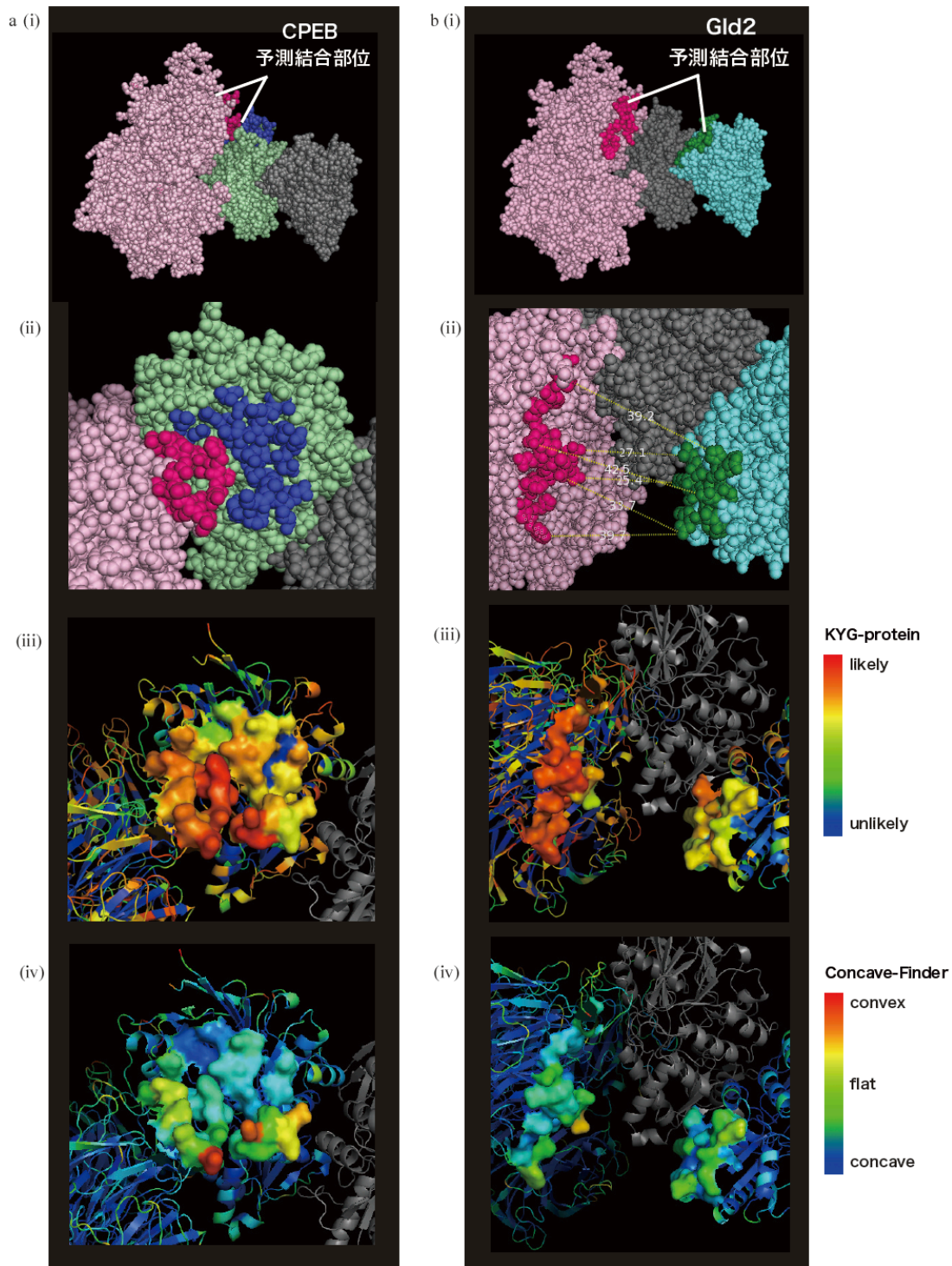


図4-8. CPSF複合体上のCPEBおよびGld2の予測結合部位

Gld2の予測結合部位(a)、およびCPEBの予測結合部位(b)。a(i, ii), b(i, ii)において薄ピンクはCPSF160、水色はCPSF73、黄緑はCPSF100、グレーは結合に関与していないCPSFサブユニットを示す。i.に示す構造は正面から見た構造であり、結合部位の拡大図をii.に示す。a. Gld2の予測結合部位。iii.では、タンパク質間結合部位予測、KYG-proteinスコアをレインボーカラーで示しており、赤がRNA結合部位になりやすいアミノ酸残基を示している。iv.では、Concave-Finderの結果を示す。緑色で示されるアミノ酸残基は平らな面を形成しており、低分子の結合部位よりタンパク質どうしの結合部位は平らであると考えられている。Gld2の予測結合部位では、CPSF160のループ部分がCPSF100と結合部位を形成していた。CPSF100側のKYG-proteinスコアは低いが、CPSF100およびCPSF73はCPSF160に比べて分子表面全体にわたって予測値が低かった。a(ii)において、破線で示したのは測定した残基どうしを示し、数値はその距離を示す(単位: Å)。予測結合部位間の距離は、最も近い部分で25Å程度。この距離の測定と表示には分子グラフィックツールPyMolの機能を利用した。

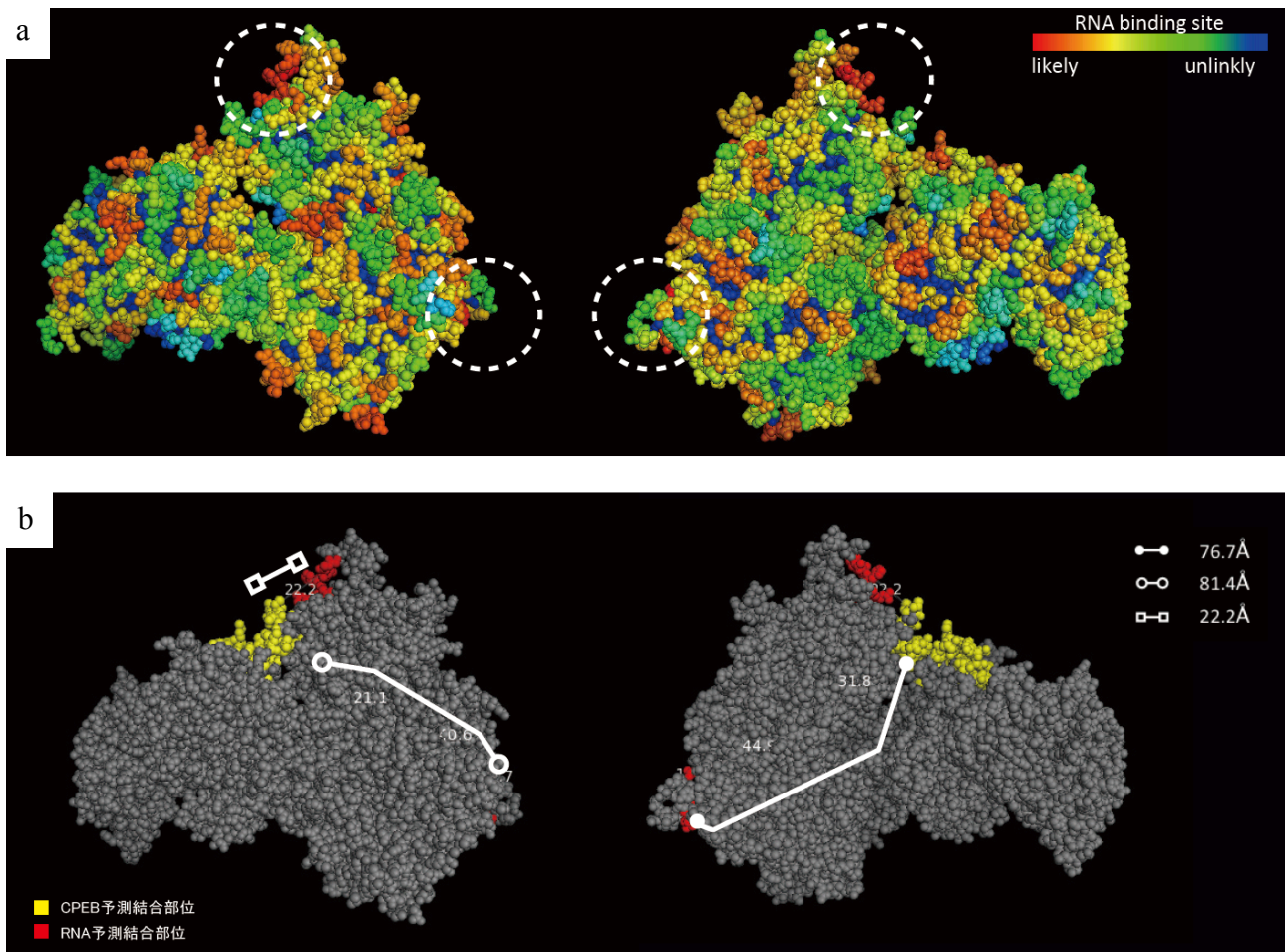


図4-9. CPSF複合体の予測mRNA結合部位

予測はKYG-RNAにより行った。赤色で示される部分は、RNA結合部位である可能性が高い(a)。KYG-proteinの結果より、結合部位が二カ所予測されたので、CPEB予測結合部位との距離を測定した(b)。二カ所とも、矛盾のない距離であったが、本研究ではCPE配列とHex配列が重なる場合を考慮し、CPEB結合部位と近い方を選択した。

6-2. Gld2の結合部位

Homo sapiens Gld2の予測立体構造は、直径より高さの低い円柱のような構造であった。

Homo sapiens Gld2は、CPSF73とCPSF160に結合するが、本研究で予測したCPSF複合体の立体構造では、CPSF73とCPSF160は直接結合していない。そこで、Gld2がCPSF73とCPSF160に同時に結合できる最も困難な場合を想定し、Gld2の短軸の両端にCPSF73とCPSF160の結合部位がある場合を考えることとした。そこで短軸を測定したところ、約30Åであった。CPSF160-CPSF73間の距離が30Å以内になる箇所があるか調べたところ、CPSF160-CPSF73間の最短距離は25.4Åであった。そこで、CPSF160-CPSF73間の距離が30Å以内のアミノ酸残基を予測結合部位候補とした。さらに、KYG-proteinおよびConcave-Finderの結果からタンパク質間結合部位を構成しそうなアミノ酸残基を予測し、結合部位を構成する平均結合残基数を利用して、結合部位の特定を行った(表4-1, 図4-8. b(i, ii))。CPSF160上の予測結合部位を構成するアミノ酸残基は、KYG-proteinによる予測において高い予測値を示した。CPSF73上の相互作用部位は予測値自体は低かったが、アミノ酸残基の凹凸の少ない比較的平面の領域を形成していた(図4-8. b(iii, iv))。

CPEB、Gld2とも、CPSF160との結合部位はCPSF160上の2番目のプロペラドメインに予測され、2つの予測結合部位は近接していた(図4-10. a,b)。このプロペラドメインは、CPSF160のパラログであるDDB1においてCullin 4Bの結合部位となっており(Fischer et al. 2011, PDB id: 4A0L)、第3章で行ったパラログセットによる結合残基の予測結果においても、3つのドメイン中最も多く結合残基が予測された。このことから、CPSF160のプロペラドメインにも結合部位が存在している可能性が高いと考えられる。

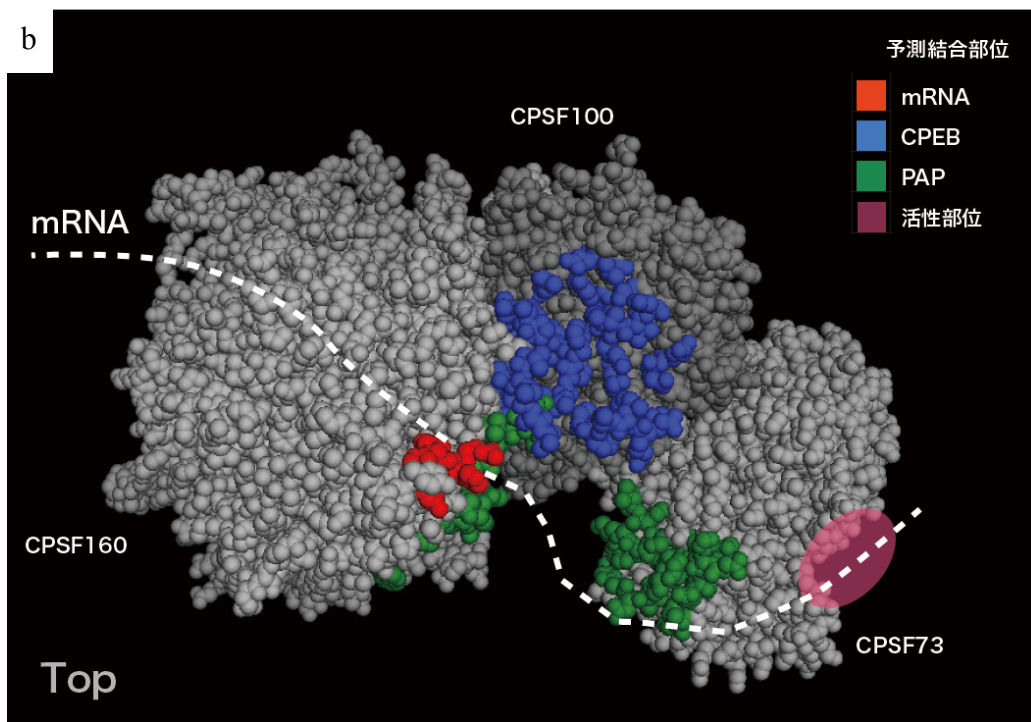
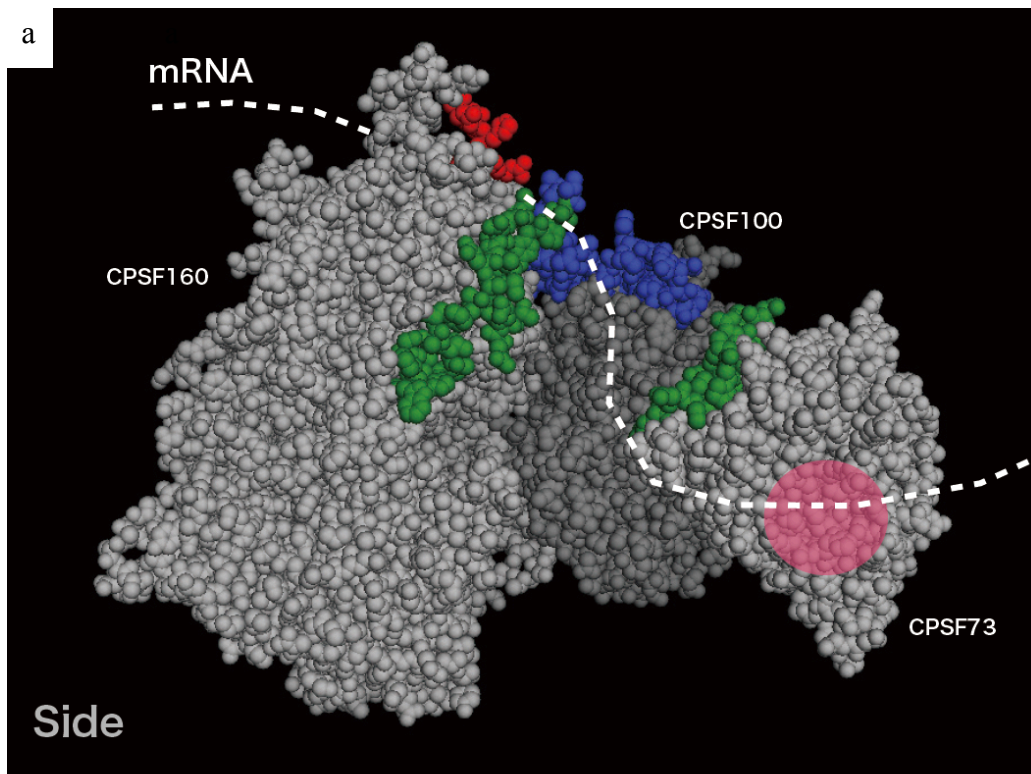


図4-10. CPSF複合体の予測構造とCPEB、Gld2およびmRNAの予測結合部位
 CPSF複合体の予測立体構造上の相互作用部位の正面図(a)、及び上面図(b)。mRNA, CPEB, Gld2の結合残基を、それぞれ黄色、青、緑で示す。

6-3. Gld2-CPSF複合体間の結合様式予測

Gld2が貯蔵mRNP複体内でどのようにCPSF複合体と結合しているかを調べるため、*Caenorhabditis elegans*のGld2の立体構造のアミノ酸配列をテンプレート配列とし、*Homo sapiens* Gld2のコンパラティブモデリングを行なった(図4-11. a, b)。この構造に対してKYG-proteinを適用し、タンパク質結合部位を予測した。その結果、タンパク質間結合部位になりやすい残基が集中している領域が、少なくとも2カ所見られた(図4-12. a)。この結果から、Gld2には複数のタンパク質結合部位があると考えられる。2つの予測タンパク質結合部位間の距離は、最も近い結合残基のC α 原子どうしで約25Å、2つの領域は約35Å前後離れていた。CPSF100とCPS160上に予測したGld2結合部位の距離も25~35Åであり、Gld2立体構造上に予測された結合部位は、CPSF100とCPSF160の結合部位である可能性が考えられる。

*S. cerevisiae*のpoly A polymerase / PAP1(PDB1D:2Q66, 図4-11. c)の立体構造は、基質であるRNAと共にその立体構造が決定されており、この構造からPAP1のRNA結合部位および活性部位がわかる。Gld2がPAP1と同様の部位でRNAと結合すると仮定し、CPSF複合体との結合様式を予測したところ、図4-12. bに示すように、Gld2はCPSF複合体のU字型の構造に埋まり込むように結合し、Gld2のRNA認識部位、および活性部位はCPSF複合体とは逆側を向いていると考えられた。また、Gld2がCPSF複合体のU字型の構造に埋まり込む様な結合を行っていた場合、Gld2とCPSF複合体の間にはmRNAが存在できるような空間がないと考えられる。このことから、既にGld2はCPSF複合体と結合していると同時に、mRNAと結合していると考えられた。またこの複合体構造においてGld2の活性部位があると考えられる位置とCPSF73の活性部位間の距離は、30~50Åと推測された。

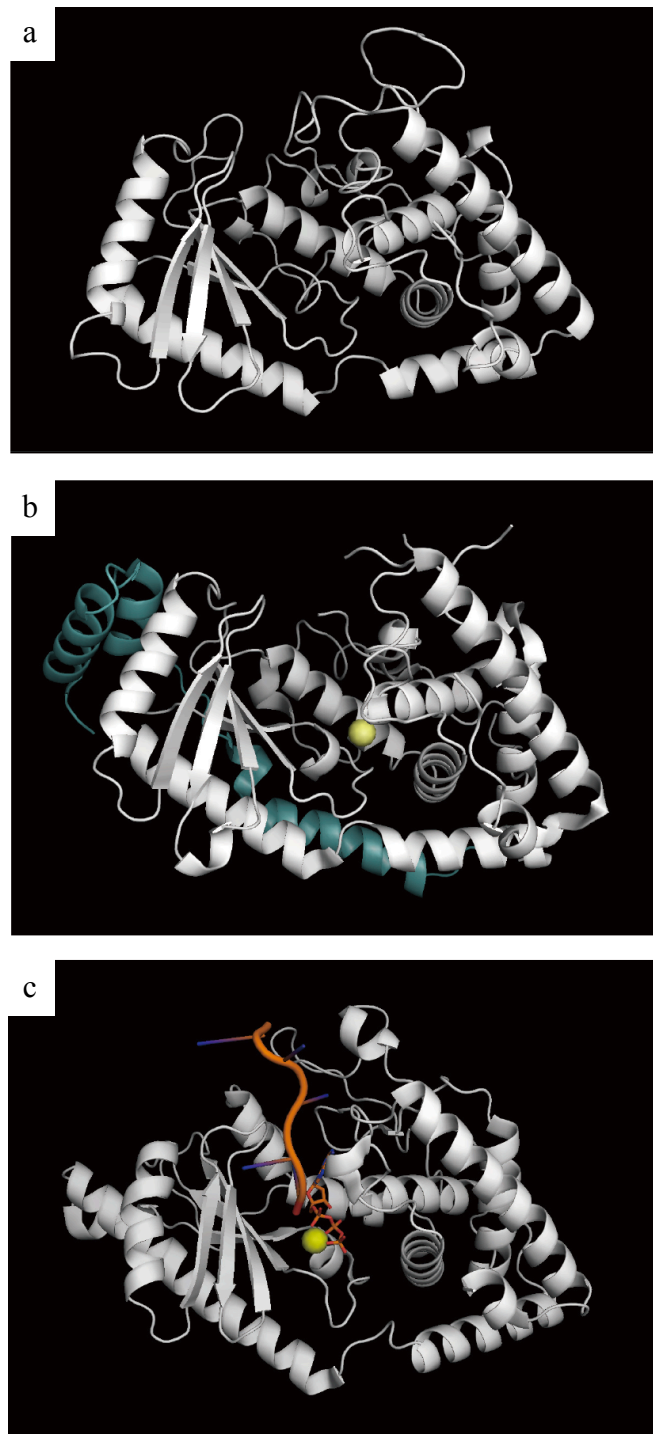


図4-11. *Homo sapiens* Gld2の予測立体構造

a. *Caenorhabditis elegans* Gld2の構造を基に予測した*Homo sapiens* Gld2の予測立体構造。b. *Caenorhabditis elegans*のGls2-Gld3複合体構造(PDBID:4ZRL)。Gld2を白、Gld3を薄緑で示している。c. Yeast poly(A) polymerase/PAP1 (PDBID:2Q66)の5残基から350残基の構造。オレンジはRNAを示す。a, b, cにおいて塩基性アミノ酸残基を薄い青、酸性アミノ酸残基をピンクで示し、保存性の高いものをそれぞれ濃い青、濃い赤で示した。b, cにおいて黄色はマグネシウムイオンを示す。

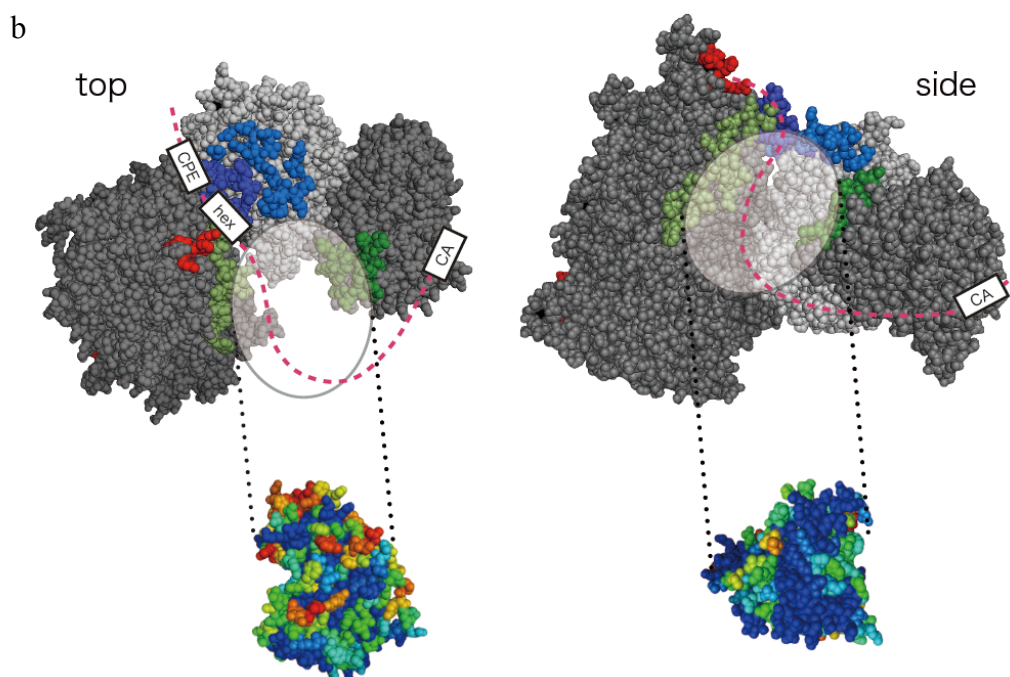
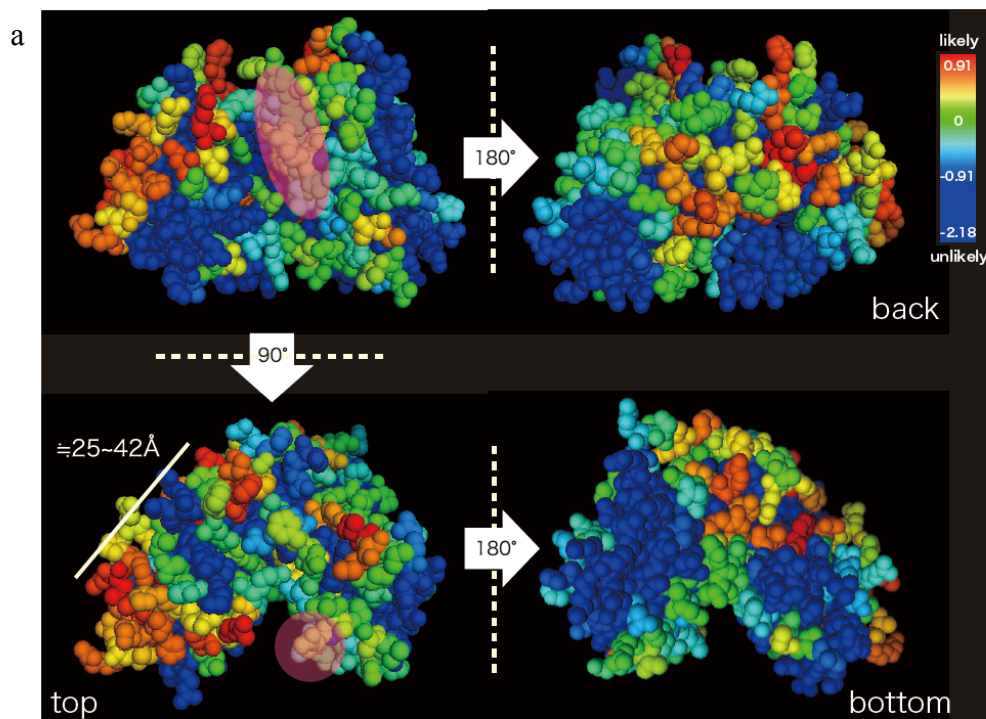


図4-12. GLd2とCPSF複合体の結合様式の予測

a. 予測した*Homo sapiens* Gld2上のタンパク質結合部位。タンパク質結合部位になりやすいアミノ酸残基を赤で示す。マゼンダ色の円で囲った部分はpap1の構造から類推されるmRNA結合部位である。*Homo sapiens* Gld2の立体構造のtop viewにおいて濃い赤で示されたアミノ酸残基を中心とした2つの予測タンパク質結合部位間の距離は最も近いところで約25Åであり、CPSF100とCPSF160間の距離と一致する。この領域はmRNA結合部位と予測される部位の逆側に位置している。b. Gld2の予測タンパク質結合部位から予想されるCPSF複合体との結合予想図。上段にはCPSF複合体上でGld2が結合する部位を楕円で示し、下段にはGld2分子の結合の向きを示している。マゼンダ色の破線はmRNAの予想位置を示しており、mRNA上のタンパク質認識部位をボックスで示した。「CA」と書いたボックスはポリA鎖付与部位となるmRNA上の配列である。

7. CPSF160上のSNV検索

dbSNPに対し、*Homo sapiens* CPSF160のSNV (missense) を検索したところ、変異位置の重複を含め、157個のSNVを得た。このうち、1000 genomes project(1000 Genomes Project Consortium et al. 2012)で確認された79のSNVにおいて、アレル頻度を得ることができた。得られたアレル頻度はすべて0.002以下であり、79個のSNVのうち、59個がアレル頻度は0.0002であり、複数の被験者に共通して検出されていたSNVは22個のみであった。このことから、得られたSNVはすべて集団に蓄積されておらず、進化医学の考え方に基けば、これらSNVはCPSF160の機能に何らかの影響を与えるものであると考えられる。しかし、コモンSNPや疾患に関与するアノテーションの得られたSNVはなかった。アミノ酸残基番号の重複を除いた150個全てのSNVについてアミノ酸残基位置を、CPSF160の予測複合体構造上に表示した(図4-13. a, b, c)。そのうち104個のSNVが立体構造上のタンパク質表面に存在した(図4-14. a)。104個のSNVのうち、予測立体構造を作成したCPSF160の配列のアミノ酸残基と変異後のアミノ酸残基を比較した場合、KYG propensityが上がる変異は49個、下がる変異は9個、有意な変化がみられなかった変異が46個あった(図4-14. b, 表4-5)。このことから、CPSF160の分子表面上に見られたSNVは、タンパク質結合性が高くなるような変異が多いことが分かった。

また、CPSF160の構造においては予測結合部位を含む側の分子半球と予測結合部位を含まない側の分子半球に大まかに分け、KYG propensityが上昇したSNV、およびKYG propensityが下降したSNVを計測した。CPSF160のタンパク質表面において、CPSF100、CPEB、Gld2のタンパク質結合部位が予測された半球では58個、タンパク質結合部位がまったく予測されなかった半球では46個のSNVが存在し、かつタンパク質結合性が高くなるような変異は、それぞれ27個、21個であった(表4-5)。タンパク質結合部位が予測された半

球において、SNVは予測タンパク質結合部位の近傍に多く存在し、また結合部位上に存在するSNVもあった(図4-13b)。特にCPSF100、CPEB、Gld2のタンパク質結合部位近傍は、KYG propensityの値が高く変化したSNVが多かった(表4-5)。

表4-5. KYG propensityの変化したSNVの局在と数

SNVの存在場所	SNV 数		
	総数	KYG propensity	
		Up	Down
CPSF160全体	150	67	21
タンパク質内部	46	19	12
タンパク質表面上	104	48	9
タンパク質結合部位が多く存在している タンパク質半球側	58	27	3
タンパク質結合部位のない半球側	46	21	6

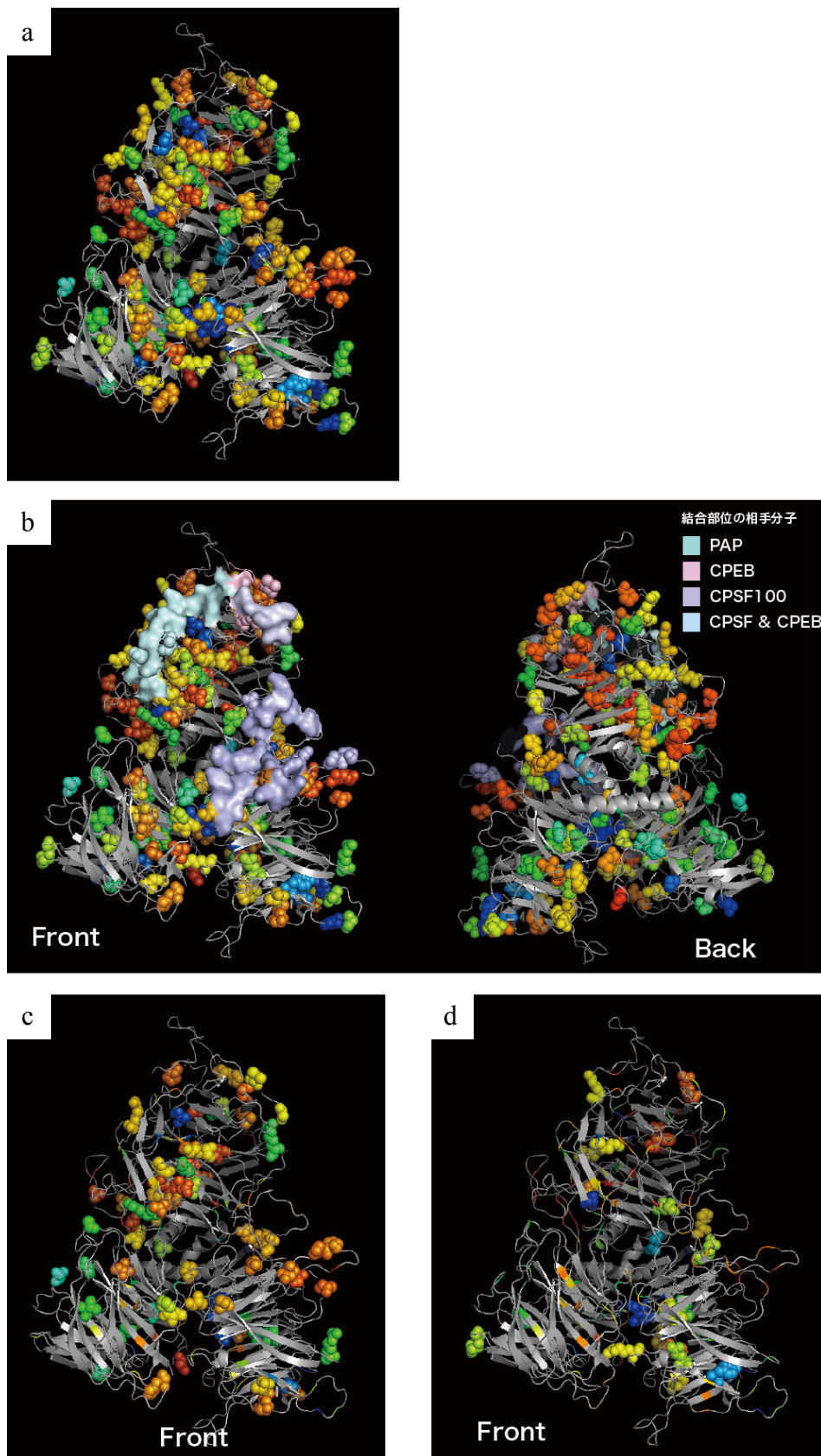


図13. CPSF160上のSNV

a.157個のSNVとその保存度。各SNVは原子を球で示している。b.CPSF160表面のタンパク質結合部位およびRNA結合部位とSNV。結合部位は面として表示した。結合部位と同色、且つ球で示されたアミノ酸残基は結合部以上のSNVを示す。c,d. 結合部位が集中しているCPSF160側からみたSNV。cではKYG-propensityが0.1以上上昇したSNVを、dでは0.1以上減少したSNVを示している。

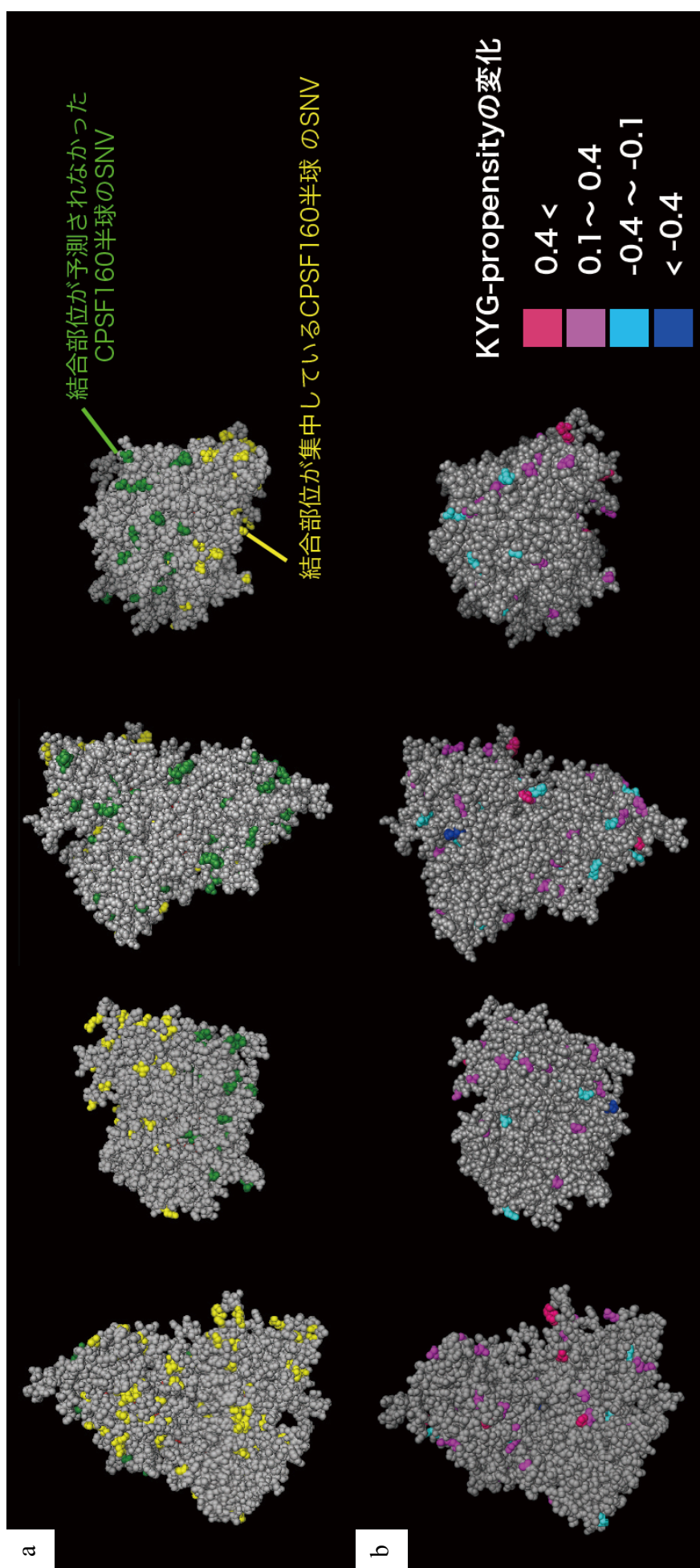


図4-14. KYG propensity の変化したSNVの局在と数

a, b. 1 つのCPSF160を奥に向けて90度ずつ回転させた。 グレーで示された領域はSNVが見つらなかったアミノ酸残基を示す。 a.CPSF160の立体構造を球と仮定した時、結合部位のある半球に存在するSNVを黄色、反対側の半球に存在するSNVを緑で示している。 b.KYG propensity の変化の傾向と大きさによってSNVを色分けした。 プラスに転じたものをマゼンダ系、 マイナスに転じたものを青系で示し、 より変化の著しかったSNVは濃い色で示した。

8. CPSF100の天然変性領域の機能予測

CPSF100のC末端領域を同体積の球に近似したところ、その半径は約18Åであった。

また、予測CPSF複合体立体構造上のCPEBの結合部位は、CPSF100立体構造のC末端側残基近くに予測されており、CPSF100立体構造のC末端側残基から、CPEBの結合部位間の距離は約15Åであった。

CPSF100のC末端領域はその分子量から、半径18Åの球と同等の体積を占める。この球をCPSF100予測立体構造のC末端残基を中心として配置すると、CPEBの結合部位と重なることが分かった。このことから、CPSF100のC末端領域はCPEBと結合していることが予測された。本研究で予測したCPEB結合部位は結合残基数が平均よりやや少ないので、CPSF100のC末端領域もCPEBの結合部位として機能している可能性が高い。

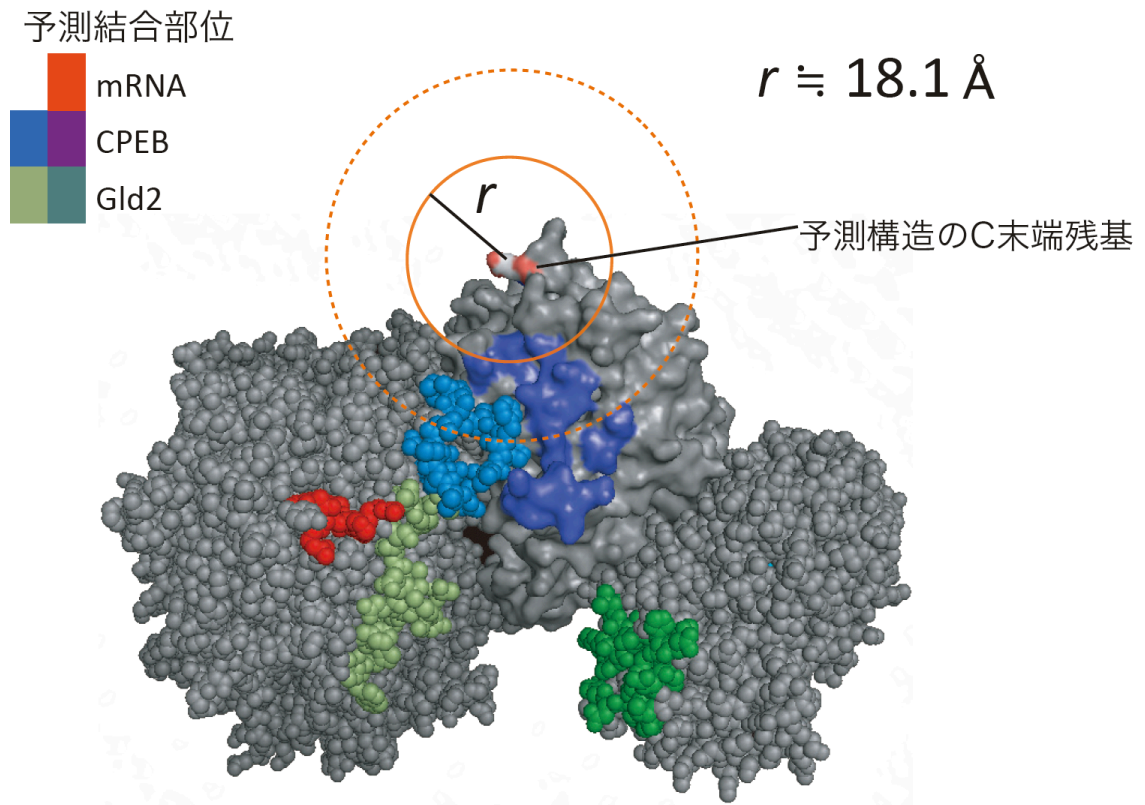


図4-15. CPSF複合体上でCPSF100のC末端領域が取りうる空間

実線はCPSF複合体予測立体構造上でCPSF100の予測構造のC末端残基を中心として球として予測したCPSF100のC末端領域が占める位置を、破線はCPSF100のC末端領域が存在しうる範囲を示す。

考察

CPSFは、CPAF160、CPSF100、CPSF73、およびCPSF30の少なくとも4つのサブユニットを有する複合体であるが、この複合体中でどのサブユニットどうしが直接結合しているかは実験的に明らかになっていない。そこで、複合体立体構造予測に先立って、本章では、直接結合しているCPSFサブユニットの同定を分子系統樹の解析により行った。このときCPSF30は、配列長があまりに短くアライメントするに耐えうるホモログ配列が特定できなかったため、系統樹を作成することができなかった。

CPSF160の系統解析の結果、CPSF160には、CFT1、SF3B3、DDB1の3つのパラログがあることが分かった。

SF3B3はSplicing factor 3B(SF3B)複合体の構成分子であり、SF3BはSF3A、U2 anRNAと複合体を組んだsnRNP複合体として、スプライソソームの会合に必須とされる (Das et al. 1999)。SF3Bには49 kD、130 kD、145 kD、155 kDのSAPsサブユニットがあり、このうちSAPs 130kDがSB3B3のもう一つの呼称である。

DDB1はDNA damage-binding complex (UV-DDB) の構成分子である。DDB1(p127)は主に紫外線によるDNA損傷時にはたらく分子の1つとして知られ、DDB2(p48)とヘテロダイマーを構成し、ヌクレオチド除去によるDNA修復を行う(Keeney et al. 1993, Abramić et al. 1991)。

表4-6. AnimaliaとYeastにおけるCleavage factorサブユニットとその複合体 (Shi et al. 2009)

Animalia	Yeast	構成する複合体
CPSF160	CFT1/Yhh1	CFII, CPF
CPSF100	CFT2/Ydh1	CFII, CPF
CPSF73	Ysh1	CFII, CPF
CPSF30	YTH1	CPF

CFT1はYhh1の別称である。Yhh1はCPSF160のホモログであることが知られており、mRNAと結合し、CA部位を認識する(Dichtl et al. 2002)。さらに CPSF73のホモログである Ysh1、CPSF100のホモログである Ydh1、tRNA前駆体のプロセッシングを行う Pta1 と Cleavage Factor II 複合体を構成する(表4-6, Millevoi and Vagner 2010)。Cleavage Factor IIは CPSF複合体のホモログをサブユニットとしてもつが、開裂部位の5'側に結合するCPSFとは異なり、開裂部位の3'側に結合する (Millevoi and Vagner 2010)。

これらの分子はすべて1000kD以上の分子であり、DNAもしくはRNAと相互作用する。CPSF160、CFT1、SF3B3、DDB1の4つの分子は、共通してMono-functional DNA-alkylating methyl methanesulfonate N-term(MMS1_N) ドメイン(pfam10433)とCPSF-A ドメイン(pfam: 03178)の2つのドメインをもっている。CPSF-AドメインはRNAもしくはDNAに結合する部分として知られており、CPSF160ではAAUAAAを認識する部位とされている (Li et al. 2001)。MMS1_Nドメインは *Saccharomyces cerevisiae*で見つかったE3 ユビキチンリガーゼ複合体の一分子MMS1のN末端側に存在するドメインであり(Hryciw et al. 2002)、保存性が高く、MMS1分子のタンパク質結合部位であることが示唆されている(Zaidi et al. 2008)。CPSF160でも、このドメインはタンパク質間結合に関与している可能性が示唆された。またCPSF160の系統樹にはArchaea、Bacteriaの配列が含まれなかった。このことからこれらの分子は比較的新しい分子であることを示唆できた。

CPSF73とCPSF100の分子系統解析から、この2つの分子は、互いにパラログな分子であることが分かった。また、この分子系統樹には、CPSF73、CPSF100以外にYsh1、IntS11、IntS9、Archaeal CPSFのグループが見られた。系統樹全体にBacteriaの配列も散見され、広い生物種で保存されている分子であることが分かった。IntS11、IntS9は、

Integrator complexと呼ばれる、RNAポリメラーゼII (RNAP II/Pol II)によって転写されたウリジンリッチな核内低分子RNA (snRNA)の3'末端のプロセッシングを行う複合体のサブユニットである。Integrator complexのサブユニットとして、IntS1からIntS12の12種の分子が同定されている(Baillat et al. 2005)。このうち、IntS11、IntS9は互いに結合することが実験的に確かめられており(Albrecht and Wagner 2012)、このことからCPSF73とCPSF100も直接結合すると考えられた。Archeal CPSFはIntS11-IntS9、CPSF73-CPSF100の様なパラログな分子が系統樹内には見られず、ホモダイマーもしくはパラログではない分子と複合体を構成する可能性がある。IntS11、IntS9のグループ内には菌界の配列は無く、動物界が分岐した後、CPSF73とCPSF100の遺伝子重複により出現した分子だと考えられる。

植物界も動物界同様にIntS11の枝に配列が見られたが、IntS11のグループ内に見られた植物界の配列のうち、アノテーションが判明しているものはCPSF73-IIのみであり、IntS11はなかった。また、IntS9に見られた植物界に由来する配列は一つのみであり(NCBI Reference Sequence: XP_002882534.1)、この配列にはアノテーションがついていなかった。さらに、CPSF73-IIはCPSF100と結合し、胚芽の形成に必須であることが示唆されている(Xu et al. 2006)。これらのことから、植物ではIntS11/IntS9のようなsnRNAの修飾に特化した複合体は存在しないと考えられる。

IntS9とIntS11はsnRNAの3'末端のプロセッシングを行なうことが知られており(Chen and Wagner 2010)、ポリA鎖伸長を伴うmRNA前駆体、およびヒストンmRNA前駆体の3'末端のプロセッシングは、CPSF73およびCPSF100が行なうことが分かっている。しかし、ポリA鎖伸長を伴うmRNA前駆体のプロセッシングを行なう複合体にはCPSF160が存在しているが、

ヒストンmRNA前駆体の3'末端のプロセッシングを行なう複合体にはCPSF160は存在していないことが示唆されている(Sullivan et al. 2009)。

このことからsnRNAのプロセッシングでは3'末端の切断を行なう複合体がポリA鎖伸長を伴うmRNA前駆体のプロセッシングのときと異なることで、修飾すべきsnRNAを認識しているのに対し、ヒストンmRNA前駆体のプロセッシングにおいてはmRNAを認識し、結合するCPSF160に変わる別の分子が存在することで、ポリA鎖伸長を伴うmRNAとヒストンmRNAの前駆体を区別していると考えられる。

母系mRNAの3'末端の切断およびポリA鎖の付与は細胞質に移行後行われるが、snRNAはそれ自身の3'末端修飾後もヒストンmRNAの修飾を行うU7 snRNPsを構成する(Yang et al. 2013)。そのため、snRNAとヒストンmRNAはその3'末端開裂を担う分子を分ける必要があったと考えられる。*in vitro*の実験系においては、CPSFはU2 snRNPsと結合し、mRNAの3'末端の開裂だけではなく、スプライシングを含めたプロセッシングに関与している可能性が示唆されている(Kyburz et al. 2006)。つまり、CPSFはU2RNPs中を構成するsnRNAとmRNAが双方とも近傍にいる環境下ではたらく必要があり、2つのRNAを認識しないよう、snRNAのプロセッシングをIntS9とIntS11が担っていると考えられる。動物のmRNAではCPSFの結合部位はmRNAの3'末端開裂部位の5'末端側に存在し、3'末端開裂後もCPSFはmRNAに結合する必要があると考えられる。しかし、Yeastでは開裂部位がCPSFホモログの結合部位の5'末端側に存在しており、mRNAの3'末端の開裂とともにCPSFホモログの複合体はmRNAから切り離される。このことから、YeastではCPSFホモログはsnRNAの近傍ではたらくことはないか、全く別のタンパク質がsnRNAのプロセッシングを行っているため、IntS11やIntS9のような分子は必要なかったと考えられる。今回作成した系統樹においてIntS11、IntS9の枝には菌界の配列がほとんど見られなかった。これらの点を総合すると

ターゲットとなるRNAを誤認しないために、動物ではCPSF73-CPSF100とIntS11-IntS9の機能を分離する必要があったことが考えられた。

CPSF73には活性部位を構成する亜鉛イオンを配位するHXHXDHモチーフが非常によく保存されていた。一方、アライメントの解析の結果、このモチーフはCPSF100では保存されていなかった。このモチーフはMBLスーパーファミリーに特徴的なエンドヌクレアーゼ活性の活性部位に結合する金属イオンを配位する配列である (Melino et al. 1998)。CPSF100はエンドヌクレアーゼ活性をもっておらず、CPSF100は祖先配列からCPSF73と分岐するにあたって、エンドヌクレアーゼ活性とは別の機能を獲得したと考えられる。CPSF100の機能にはこの活性は不必要であり、CPSF73とは全く異なる機能を有していると考えられる。

CPSF73およびCPSF100の系統樹におけるBacteria由来の配列

BacteriaにおけるmRNAのポリA鎖の存在は、1960年代に示唆されていた (August et al. 1962) が、その後Bacteriaのpolyadenylationについてはほとんど研究がなされてこなかった。近年になり、BacteriaのポリA鎖は菌界、植物界、動物界のポリA鎖に比べ、機能的に未発達であることが分かってきた (Mohanty and Kushner 2011)。BacteriaのポリA鎖は15~30塩基程度であり、これは哺乳類などのポリA鎖の一般的な長さとしてされている100~150塩基に比べてかなり短い (Mohanty and Kushner 2011)。また、Bacteria全体では、アデノシンではなく、ウラシルによるポリU鎖や4種のヌクレオチドすべてを含むポリヌクレオチド鎖を有する(50%程度はアデノシンであるとされている)種もあり、安定性も低いことが分かっている (Mohanty and Kushner 2011)。また、グラム陽性菌である*B. subtilis*では全RNAの15-25%がポリA化されるのに対し、*Escherichia coli*では2%しかポリA化されない (Mohanty

and Kushner 2011)。このことから、Bacteriaでは、mRNAのポリA化は機能として確立された生命現象ではないと考えられている(Mohanty and Kushner 2011)。作成したCPSFサブユニットの分子系統樹では、複数のBacteriaのCPSF73のホモログと考えられる配列が様々な分岐位置に散逸して見られた。このことから、BacteriaではCPSF73と類似した配列をもつ分子が存在するが、配列保存度が低く、その機能は重要ではないと考えられる。

分子系統解析の結果、CPSF複合体の各サブユニットはそのパラログとは全く異なった分子と結合していることが分かった。つまりCPSFサブユニットはそのパラログとは異なる結合部位を有していると考えられる。

CPSF複合体の立体構造予測

アミノ酸配列が似通ったタンパク質は、類似の立体構造をとることが経験的に分かっている(Chothia and Lesk 1985)。このことから、共通祖先由来のタンパク質間では、その立体構造は類似であるため、相互作用部位や活性部位の位置も似通っている場合が多く、例えば、共通祖先由来の2つのタンパク質が相互作用する場合、その相互作用面は、相同なドメインで形成されることができると考えることができる。CPSF73とCPSF100は、この2つのタンパク質が共通してもつドメインで相互作用しており、その予測二量体構造には対象軸が存在した。また、CPSF73とCPSF100の予測結合部位は、FireDockで予測した結合残基とほぼ同じ位置の結合部位で結合していた。予測構造全体はU字に似た構造をとっており、CPSF100がmRNAの認識を行うCPSF160とエンドヌクレアーゼ活性をもつCPSF73をつなぐ足場としてはたらいっているような構造であった。CPSF73は、CPSF100と結合することで、エンドヌクレアーゼ活性が高まることが示唆されている(Kolev et al. 2008)。また、CPSF160のCPSF100との結合残基は、多くがCPSF160パラログにはないアミノ酸残基で構成されてい

た。CPSF160パラログはCPSF73パラログやCPSF100パラログとは結合していないため、CPSF160は挿入によってCPSF160特有の結合部位を獲得したと考えられる。以上のことから実験的事象と反しないCPSF複合体の立体構造を、生命情報学的なアプローチによって得ることができた。

CPSF複合体上のタンパク質結合部位予測

予測したCPSF複合体の立体構造を用いて、CPEB、Gld2の結合部位予測を行った。

先行研究により、CPEBはCPSF160とCPSF100、Gld2はCPSF160とCPSF73と結合することが示唆されていた(Papin et al. 2008, Laishram and Anderson 2010)。CPEBとGld2が結合する2つのCPSFサブユニット上のアミノ酸残基の距離とCPEBおよびGld2の直径により、結合部位となりうるアミノ酸残基を絞り込むことができた。立体的な位置関係によって絞り込んだアミノ酸残基から、KYG-proteinのスコアが高く、タンパク質結合部位の特性であるアミノ酸残基の凹凸が少ない表面を形成しているアミノ酸残基が36~59個程度集合している領域を特定することで、CPEB、Gld2の予測結合部位を得た。

CPSF複合体構造上で、CPEBおよびGld2の予測結合部位は互いに近接していた。二つの分子はCPSF複合体がU字型の複合体構造を形成していることで近接して存在できていることが分かった。CPEBの結合部位はCPSF160上のRNA結合部位に近く、Gld2はCPSF160上のRNAおよびCPEB結合部位からCPSF73の活性部位の間に予測された。母系mRNA上では、5'末端よりCPE配列(CPEB結合部位: UUUUAU, UUUAUなど)、hex配列(CPSF認識部位: AAUAAA)、CA配列(CPSF73によるmRNA3'末端開裂部位)があり、それぞれの距離はCPEとhexはオーバーラップしている場合を含めて0~30 nt、hexからCA部位までは10~30 ntであることが確認されている (Meijer et al. 2007)。 *Xenopus leavis*の卵母細胞では、母系

mRNPs中のPAPとして、GLD2 poly(A) polymeraseが同定されている(Rouhana et al. 2005)。GLD2はCPEBに結合し、CPEBにpoly(A) specific ribonucleaseであるPARNが結合することで、GLD2のポリA鎖伸長作用を拮抗阻害していることが示唆されている(Kim and Richter 2006)。Kim(2006)によると、PARNはGLD2ともCPEBとも直接結合すること、CPEBのリン酸化によりPARNが結合しなくなることが示唆されており、CPEB、GLD2、PARNの3分子は互いに結合しているか、かなり近接して存在していることが考えられる。

本研究の解析から、CPSF複合体上にCPEBの結合部位と考えられる十分な表面積のある部分は、1カ所しかなかった。しかし、翻訳時期の遅い母系mRNAの中には、2つ以上のCPE配列を持つものがある(Belloc et al. 2008a)。このような母系 mRNPsでは少なくとも1つのCPEBはリン酸化されることで分解され、その後に翻訳がなされることが分かっている(Belloc et al. 2008a, Piqué et al. 2008)。今回予測したCPSF複合体構造上のCPEB結合部位は、CPSF160上のRNA認識部位と考えられる領域と近接していた(図4-10. a,b)。このことはCPSFのHex配列への親和性がCPEBとの結合親和性より弱い場合、2つCPEBが互いにCPSFとの結合を拮抗阻害することにより、CPSFのmRNAへの結合も阻害されることを意味する。CPSFはGld2のリクルートを行うはたらきがあるので、結果としてCPE配列を2つもつmRNAのポリA鎖伸長が遅くなり、翻訳時期の遅延に寄与していると考えられる。

Gld2については、*Caenorhabditis elegans*において立体構造が明らかとなっていることから、*Caenorhabditis elegans* Gld2の立体構造を元に*Homo sapiens* Gld2の立体構造予測を行ない、Gld2がどのような向きでCPSF複合体に結合しているか予測した(図4-12. a)。また、類似構造を持つYeast PAP1の立体構造(PDBID:2Q66)からGld2の活性部位の位置が予想された。Gld2はmRNAとの結合部位および活性部位を外側に向けて、CPSF複合体に埋まり込

むように結合していることが分かった(図12. b)。この結果から、本研究で予測したCPSF複合体立体構造上のmRNA結合部位およびCPEB結合部位の位置を元にGld2のmRNA結合部位は、mRNAのポリA鎖付与部位より4~8塩基上流の位置に結合していると予測された。このことから、貯蔵mRNP複合体中でGld2は、CPSF100、CPSF160、CPEBの少なくとも3分子と結合することでmRNAのポリA鎖付与部位を認識しないように固定されており、結果的にmRNAのポリA鎖伸長阻害を維持していると考えられる。

貯蔵mRNPにおいてポリA鎖伸長が起こらない理由については、mRNP複合体中にポリA鎖特異的リボヌクレアーゼであるPARNの活性が確認されており、Gld2よりやや活性の強いPARNがGld2と拮抗することによって、mRNAのポリA鎖を短い状態に保っているためと考えられている(Kim and Richter 2006)。しかし、母系mRNAが細胞質で貯蔵されている間、2つの分子がリン酸化反応、脱リン酸化反応を交互に行ない続けるのは非常に効率が悪い。安定した転写抑制を維持するためには、Gld2が貯蔵mRNP中でポリA鎖を付与できない位置にいる期間があることは十分考えられる。

また、母系mRNPsの翻訳抑制解除の引き金と考えられているCPEBのリン酸化、および母系mRNAの翻訳活性に必須とされるポリA鎖の伸長の2つの現象をつなぐ分子機構は未知であった。しかし、本研究で行なった貯蔵mRNPsのCPSFを中心とした三次元的な相互作用の予測結果を元に、CPEBのリン酸化とポリA鎖の伸長の2つのイベントのつながりを考えると、CPEBのリン酸化によってGld2がCPEB、CPSF複合体との結合が困難になり、mRNPs中で比較的自由に動き回ることができるようになったGld2がmRNAのポリA鎖付与部位を認識できるようになるのではないかという予測ができ、ポリA鎖の伸長はCPEBのリン酸によって引き起こされる一連の構造変化の結果としてとらえることができる。

*Xenopus leavis*の減数分裂前期の未成熟卵を用いた実験において、*Xenopus leavis*

Gld2(XGld2)のモノクローナル抗体を用いてXGld2またはCPEBを阻害すると、プロゲステロン処理後の核膜崩壊(Germinal vesicle break down: GVBD)の進行が早まること、Mos mRNAのポリA鎖伸長が阻害されることが示唆されている(Papin et al. 2008)。これらの2つの事象は一見相反するような結果であり、これを説明するモデルは提唱されていない。しかし、本研究で示したGld2のCPEB、CPSFとの相互作用が翻訳抑制機構の要だと考えると、この2つの事象は説明することができる。すなわち、貯蔵型mRNPs中のXGld2、CPEB、CPSFの相互作用を阻害したことで翻訳抑制の維持ができないmRNPs構造となり、細胞質に存在する別のPAP、もしくは新規合成されたXGld2によりポリA鎖伸長が行われ、結果的にMosの翻訳とGVBDの進行が行われたと考えられる。

CPSF160立体構造上のSNV

今回最も多くのタンパク質の結合部位を予測したCPSF160について、そのSNVをdbSNPで検索したところ、アミノ酸配列長に対して1割以上にあたる157個のnon synonymous SNVを得た。

このうち、タンパク質表面のアミノ酸残基変異を引き起こしているnon synonymous SNVは104個であった。これらのSNVをKYG propensityの値を用いてタンパク質間相互作用への影響を調べたところ、104個のうち48個のSNVは変異によって有意に他のタンパク質分子との結合親和性があがっていた。このことから、CPSF160のタンパク質表面は他のタンパク質との結合親和性が下がるような変異に比べ、結合親和性があがるような変異に対しては寛容であることが分かった。しかし、これらのSNVはすべてアレル頻度が低く、集団に変異が蓄積されていなかった。このことから、CPSF160の機能は強固なタンパク質間結合

が起こると、それが阻害されること、CPSF160のパラログに比べてタンパク質親和性の弱い暫時的な足場タンパク質として機能していることが考えられる。これらのことから、CPSF160はあるタンパク質との特異的な結合を行わないと考えられ、このような性質は複数の転写時期によって構成因子を変化させる母系mRNPsの構造維持に役立っていると考えられる。

本章では、CPSF100のC末端領域がCPSF100特有の分子機能に関与している可能性についても調べた。天然変性領域は立体構造予測が行えないが、CPSF100のC末端領域は水溶性領域であるため、球に近似することが可能である。CPSF100のC末端領域と同体積の球がCPSF複合体予測立体構造上に占める位置を予測したところ、予測CPSF複合体立体構造上のCPEBの結合部位に重なることが分かった。一般的なタンパク質結合部位は18残基から30残基のアミノ酸残基で構成されているが(Chakrabarti and Janin 2002)、本研究で予測したCPSF複合体上のCPEBの予測結合部位は14残基となり、平均より少ないアミノ酸残基数で構成されていた。CPSF100の立体構造のC末端残基からCPEBの結合部位間での距離は15Åであり、CPEBの半径は26Å、CPSF100のC末端非構造領域の半径は18Åだった。これらのこと、およびCPEB、CPSF100の立体構造のC末端残基とも球で近似していることを踏まえると、CPSF100のC末端領域はCPEBと結合し、CPEBの結合補助をしていると考えられる。非構造領域は多角的相互作用を行う領域として知られるため、CPSF100のC末端領域はCPSF複合体が核内ではたらく際は、CPEBとは異なる分子との相互作用部位として機能していると考えられる。

CPEBは母系mRNPsの転写が活性されるとリン酸化される、または、過リン酸化により破壊されることが知られている(Mendez et al. 2000a)。CPEBの複数箇所がリン酸化される

ため(Mendez et al. 2002)、分子表面の酸性化が進むか、CPEBが分解され不在となる。このCPEBの変化がCPEBの周囲の貯蔵mRNPs構成分子の相互作用を変化させ、PARNなどの翻訳抑制にはたらく分子と分離することで(Kim and Richter 2006)、貯蔵mRNPsから翻訳活性型mRNPsへmRNPsの全体構造を変化させる。本研究で予測したCPSF複合体構造によって、CPSF100のC末端領域がCPEBの結合に関与している可能性が示唆された。CPSF100のC末端領域は非構造領域であり、その性質よりCPEBのリン酸化によるタンパク質表面の変化を受け、翻訳活性化の刺激として速やかに貯蔵mRNPs全体に伝えるしくみを有していると考えられる。

終章

初期発生機構研究における本研究の位置づけ

特に動物において、その子孫を残すために重要な初期発生は、受精に適切なタイミングで精子を待つためと考えられる卵細胞の細胞周期休止や、多精拒否機構のひとつと考えられる受精膜の形成など、顕微鏡で観察されるそれらの現象の下に、複雑な分子制御機構の存在が伺える興味深い研究対象である。初期発生では、種の存続に関わる様々な重要な現象が、個体の寿命からみれば非常に短い時間に起こり、非常に保存された機構が存在するように思えるが、生物種間で比較すると卵細胞の細胞周期休止期をとっても異なっており、多様性の高い側面をもつ。このように多様性の高い生命現象の解明では、多くの生物が共通してもつ、より重要な機構をあぶり出すためにも、様々な生物種で研究を行うことが非常に重要である。本研究で研究対象とした母系mRNPsによる翻訳抑制制御も、多くの動物の卵細胞において、すでに観察されている。

現在の母系mRNPsの翻訳制御における研究では、「どのような分子が複合体に含まれるか」、「複合体に含まれる各サブユニットはどのような機能をもつか」などが重視されて研究されており、母系mRNPsが多機能な複合体であることは明らかになっているにも関わらず、複合体内のサブユニット、もしくはその機能がどのように相互作用し、母系mRNPsの機能を発揮しているかについての原子レベルでの解明は、行われていない。これには、母系mRNPsを構成するような分子が、立体構造を決定することが困難であるため、また、一般的に相互作用を理解するための網羅的手法が限られていることが大きな原因だと考えられる(von Mering et al. 2002)。また、母系mRNPsは、そのコードする遺伝子の翻訳時期によって、各構成因子が異なること、また生物によっても構成因子が異なることが明らかとなっている(Radford et al. 2008)。

母系mRNPsにおいても、多くの生物が共通してもつ特性を明らかにすることは、その翻

訳抑制機構の解明に重要であると考え。また、母系mRNPs複合体は、貯蔵型-翻訳型の母系mRNAの翻訳ステータス、母系mRNAのコードする遺伝子の翻訳時期、そして生物種によって構成因子が少しずつ異なることから、母系mRNPs複合体の特性として、複数の構成因子の増減が可能であるような構造をもつと考えられた。そこで、本研究では、母系mRNPs複合体全体の構造が、多くの生物に共通する母系mRNPsのもつ特性のひとつであるという仮定のもと、実験的に明らかにすることが困難な母系mRNPs複合体の立体構造予測に取り組んだ。母系mRNPs複合体の構成因子は数十個程度あるため、本研究では、翻訳時期によらず、母系mRNPs中に存在すると考えられている分子のみにしぼり、最終的に *Homo sapiens* のCPSF複合体を予測し、CPSF複合体-Gld2-CPEBがどのような相互作用様式を行っているか考察した。

本研究によって得られた成果

本研究では母系mRNPsの三次元構造予測の足がかりとして、基本構成分子の1つと考えられるCPSF複合体に着目した。CPSF複合体は、貯蔵型mRNAに結合し、転写可能な成熟mRNAに変化する際に行われるポリA鎖の伸長を行う複合体である。また、CPEB、polyA polymeraseであるGld2といった貯蔵型mRNPsに特徴的な分子と結合し、貯蔵型mRNPsを構成する上でも重要な働きを担う複合体である。

第一章では、CPSF複合体を構成するサブユニットの立体構造モデルを作成した。その際、モデルの作成が困難な天然変性領域の予測を各サブユニットに対して行った。その結果、CPSF160のpropeller-Bドメイン領域、およびCPSF100のC末端ドメインに、比較的大きな天然変性領域が存在することが分かった。第四章の結果より、CPSF160の2つ目のプロペラドメイン上に近接してCPSF100、CPEBおよびGld2結合部位が予測された。また、

CPSF100のC末端ドメインはCPSF100に特徴的な配列であり、配列類似性の非常に高いIntS9もこのドメインはもたない。天然変性領域は一般的にタンパク質どうしの結合が重要な分子によく見られ、その柔軟性がタンパク質どうしの結合をしやすくしていると考えられており、CPSF160とCPSF100のこの天然変性領域は、CPSFのはたらきに重要なタンパク質間結合に寄与している領域であると考えられた。

第二章では、CPSFのような多数の複合体を想定した結合部位の予測手法の開発を行った。一般的に複合体構造予測ソフトウェアでは、何らかの物理化学的な複合体構造の特徴をもとに複合体の候補構造にスコアを付け、1つ1つの候補構造をランキングする。この手法で予測を行うと、ランキング10位以内の候補構造でも様々な結合様式の複合体が含まれており、尤もらしい複合体構造を選択するのは困難である。そこで本研究では、各候補構造をその複合体立体構造の形状に基づいてクラスタリングし、各クラスターのスコアを算出することで、尤もらしい結合様式の候補構造を絞り込む手法の開発を行った。この結果、クラスタリングは比較的平らで広い結合面を形成する複合体立体構造予測に有効にはたらき、含まれる候補構造数が多いクラスターを選出するだけで、尤もらしい結合様式の候補構造を絞り込むことができると分かった。一方、自由度の低い結合部位を形成する複合体構造の予測には、スコアによる予測が重要であることが分かった。そこで本研究では、結合部位を構成しやすいアミノ酸残基の頻度によって結合部位を予測するKYG-proteinのスコアを用いたクラスタースコアを算出することとした。この結果、平らで広い結合部位、および自由度の低い結合部位の両方を想定した複合体構造予測プログラムを開発できた。

CPSFサブユニットはそのパラログとは全く異なるサブユニット構成の複合体を形成している。特にCPSF160のパラログであるDDB1は、CPSFサブユニットとは全く異なる立体構

造の分子と結合することがX線結晶構造解析により明らかとなっている。これらのことから、CPSFサブユニットはパラログとは異なる結合部位を有すると考えられた。

そこで第三章では、パラログ間で異なるサブユニット構成をもつ複合体どうし(パラログセット)の結合部位を6つに分類し、そのような結合部位にみられる特徴の抽出を行った。その結果、6タイプ中の5つに分類されるパラログセットを得た。また、ホモ多量体、ヘテロ多量体ともに、パラログが単量体である複合体をサブユニット構成が変化した最も多いパラログセットとして得た(HoIタイプ、HeIタイプ)。しかし、この2つのパラログセットは、大きく異なる傾向をしめすことが分かった。ホモ多量体(HeIタイプ)では単量体のタンパク質の界面の残基に比べ、多量体のサブユニット界面の残基は一般的に結合部位の特徴である疎水度と保存度が高い傾向を示していた。一方、ホモ多量体(HoIタイプ)は、結合残基全般に関しては疎水性度、保存度も単量体と同程度であったが、保存度が際立って高い残基がいくつかみられることがわかった。この残基は多量体構造を安定させる“hot-spot”残基であると考えられる。

この2タイプを含めた全5タイプのパラログセットにおいて、パラログセット内で見いだされた異なる結合部位の特徴をもとに、CPSFサブユニットがこのような結合部位をもつかを調べた。その結果、CPSF160ではC末端側のプロペラドメインに多くの結合残基が予測された。このドメインは、DDB1においてDDB2の結合部位となっているドメインである。CPSF160ではDDB2とは異なる立体構造をもつタンパク質と結合するために、この部分にCPSF160特有の結合残基を得る必要があったと予測した。CPSF73においてもいくつかの結合残基が予測されたが、それらの結合残基は結合部位周辺には存在しなかった。一方、CPSF100ではCPSF73では活性部位にあたる部分に多くの結合残基が予測された。こ

のことから、CPSF100はCPSF73のエンドヌクレアーゼ活性とは異なる新たな機能を有している可能性を挙げた。

第四章では、CPSFサブユニットの組み上げを行った。まず、複合体中でどのサブユニットペアが直接結合しているか、分子系統解析によるアプローチにより予測を行った。その結果、CPSF73およびCPSF100は互いにパラログであり、Archaeaでは1つの分子として有していた機能を2つに分けた可能性が示唆されていることから、2つが直接結合してはたらく可能性が高いこと、CPSF160は動物界、植物界、菌界にのみみられる比較的新しい分子であることが分かった。

この結果を受けて、CPSF73とCPSF100の複合体構造を、次にCPSF73-CPSF100ヘテロ2量体とCPSF160の複合体構造を、第二章で開発した手法を用いて予測を行った。CPSF30は予測できたモデルが配列長の30%程度しかなく、サブユニット表面を十分予測できていないと判断したため、分子界面のアミノ酸残基の特性に依存する結合予測を行わなかったが、CPSF複合体の比較的大きなサブユニットであるCPSF100-CPSF73-CPSF160複合体の三次元立体構造を得ることができた。

さらに、ドッキングにより予測したCPSF複合体の結合残基と第三章でパラログセットのサブユニット界面の特徴から予測したCPSFサブユニットの予測残基の比較を行った。CPSF73、CPSF100の結合部位では予測結合残基位置が50%以上一致しており、特にCPSF100においては、予測CPSF複合体におけるCPSF160の結合部位となる領域がパラログ間で共有しない結合部位として予測された部分と合致していた。このことから、CPSF100はエンドヌクレアーゼ活性を失うかわりに結合部位を新たに獲得することで、CPSF73と機能を分離するとともに、mRNAとの高いタンパク質親和性をもつCPSF160をCPSF73と繋ぐ足場タンパク質としてのはたらきを持つようになったと予想される。

三次元立体構造を得ることの利点は、複合体周辺にどのような空間が形成されているかの理解が容易にできることである。また、原子レベルの解像度で三次元立体構造を理解することにより、サブユニットどうしの相互作用だけではなく、各サブユニットの機能がどのように発揮されているかが明らかになる場合がある。

本研究で得られたCPSFの予測三次元立体構造を用いて、母系mRNAの結合部位、CPEB、Gld2の結合部位、またはそれらの分子が複合体上に占める空間を予測した。さらに、CPSF複合体の三次元立体構造をもとに、CPSFサブユニット、CPEB、Gld2の機能の実現様式について考察し、脊椎動物の初期発生における遺伝子発現制御機構の1つである母系mRNAの翻訳抑制機構の維持を実現するモデルを提示することができた。

以下に、CPSF複合体の立体構造により考察することができたことをまとめる。

① CPSF160のタンパク質界面の特性

CPSF160はCPSFサブユニットの中で最も大きなサブユニットであり、mRNA認識を行うと考えられている。本研究によって、CPSF160はCPSF100、CPEBおよびGld2タンパク質結合部位を持つことが予測された。また、そのパラログであるDDB1、SF3B3、YHH1にくらべ、タンパク質表面のアミノ酸配列保存性が低いことが分かった。CPSF160のような遺伝子発現に関与するタンパク質は、その機能が生存に重要であるため、保存性が高くなると考えられる。このことから、CPSF160のタンパク質表面の保存性の低さは、CPSF160特有な機能に関与している可能性が高いと考えられた。CPSF160界面の特性を調べるため、CPSF160の遺伝子上に見つかったnon-synonymous SNVについて、その特徴を調べた。その結果、CPSF160のnon-synonymous SNVは、タンパク質との親和性が上昇するようなアミノ酸残基置換を起こす変異が多く、これは親和性が減少する方へアミノ酸残基置換したSNV

数の5倍以上であった。特に、CPSF100、CPEB、Gld2の結合部位を予測した側面には、タンパク質結合親和性を減少させるようなSNVは少ない傾向にあった。しかし、*Homo sapiens* CPSF160のアミノ酸配列に見つかったSNVは、そのすべてが集団内でのアレル頻度が低かった。

これらのことから、CPSF160のタンパク質表面はタンパク質結合親和性の高いSNVが入りにくい性質をもち、タンパク質分子親和性を高く保つことがCPSF160の機能に重要であるが、同時にタンパク質結合親和性が高くないこともCPSF160の機能に重要と考えられる。こういった特性から、CPSF160は母系mRNPsの中で足場タンパク質のような働きを担っており、結合しているタンパク質を翻訳活性型mRNPsと比較してコンパクトな構造である貯蔵mRNPs内にとどめておき、翻訳活性型のmRNPsに変化した際にそれらのタンパク質との結合を解放するような働きをもつと考えられる。

② CPSF100のC末端側領域の機能

CPSF100はその機能が明らかにされていない分子である。本研究により、CPSF100はC末端側にCPSF73にはない天然変性領域をもつことが分かった。単体では天然変性領域は立体構造を予測することはできないが、その分子量から複合体上で占める位置や空間を計算することができる。この結果より、CPSF100のC末端領域はCPEBとの結合を行っている可能性が高いと考えられた。CPSF73とCPSF100の複合体は、ヒストンmRNAでは、CPSF160の介在なしにSLBPやLsm11と結合し、その翻訳制御を行っていることが示唆されている(Sullivan et al. 2009)。CPSF100のC末端側領域は、複数のmRNAのプロセッシングにおいて、CPSF73の活性を維持するためにCPEBや他の翻訳制御を担う分子と結合できるような多角的な結合部位となる必要があったと考えられる。

③ CPEBの翻訳時期調節

母系mRNA 3'末端上のCPEB結合部位であるCPE部位は、細胞質でポリA鎖伸長が観察されるmRNAに、その転写時期に関わらず存在する配列である。しかし、翻訳時期の遅い遺伝子をコードしているmRNAではCPEが2つあることが分かっており、人工的にCPEを複数付けたmRNAは発現が遅くなることが分かっている(Piqué et al. 2008)。本研究で予測したCPSF複合体上のCPEB結合部位とmRNA結合部位は近接していたが、CPSF複合体上のこれらの結合部位は、CPSF複合体にCPEBとmRNAが同時に結合できる位置であった。しかし、CPEBがmRNAより先にCPEBに結合すると、CPEBはCPSF複合体上のmRNAの結合するための経路を占領してしまう可能性が考えられた。これらのことから、mRNAの結合部位において、CPSFのmRNA親和性がCPEBとの結合親和性より弱かった場合、2つのCPEBが互いにCPSFとの結合を拮抗阻害することにより、CPSFのmRNAへの結合も阻害されることを意味する。CPSFはGld2のリクルートを行うはたらきがあるので、結果としてCPE配列を2つもつmRNAのポリA鎖伸長が遅くなり、翻訳時期の遅延に寄与していると考えられる。

④ CPSF-Gld2-CPEBの相互作用による翻訳抑制維持

貯蔵mRNPにおいてポリA鎖伸長が抑制されているメカニズムについては、ポリAリボヌクレアーゼであるPARNによるGld2の拮抗阻害が示唆されている(Kim and Richter 2006)。本研究でGld2とCPSF複合体の予測立体構造との結合予測を行ったところ、Gld2はmRNAとの結合部位および活性部位を外側に向けて、CPSF複合体に埋まり込むように結合しており、さらにCPEBがGld2をCPSF複合体に押しえ込むように結合していることが分かった。

この予測構造では、Gld2はmRNAのポリA鎖付与部位より4~8塩基上流の位置に結合しており、ポリA鎖伸長が行えないことが分かった。母系mRNAが細胞質で貯蔵されている間、2つの分子がリン酸化反応、脱リン酸化反応を交互に行ない続けるのは非常に効率が悪く、安定した転写抑制を維持するためにGld2が貯蔵mRNP中でポリA鎖を付与できない位置にいる期間があることは十分考えられる。

⑤ 翻訳活性化に伴うGld2のリモデリング

母系mRNPsの翻訳抑制解除の引き金と考えられているCPEBのリン酸化と母系mRNAの翻訳活性に必須とされるポリA鎖の伸長の2つの現象をつなぐ分子機構は未知であった。また、*Xenopus leavis*の未成熟卵を用いた実験において、抗体によりXGld2またはCPEBを阻害すると、核膜崩壊の進行が早まるが、Mos mRNAのポリA鎖伸長が阻害されるといふ、相反する現象が報告されている(Papin et al. 2008)。この2つは本研究で示したCPSF-Gld2-CPEBの複合体構造による翻訳抑制維持のモデルにより説明することができる。

本モデルでは、CPEBのリン酸化はその界面の組成の変化、もしくはCPEBの分解によって、Gld2とCPEB、およびCPSF複合体との結合を困難とし、結果としてmRNPs中で比較的自由に動き回ることができるようになったGld2がmRNAのポリA鎖付与部位を認識できるような位置に再配置されると考えられる。このような構造変化が母系mRNPsに起こっていると考えると、ポリA鎖の伸長はCPEBのリン酸によって引き起こされる一連の構造変化の結果であると考えられる。また、Gld2の阻害によりMos mRNAのポリA鎖伸長に阻害がみられるにもかかわらず、Mosの必要な核膜崩壊の進行が早まるという事象についても、本モデルではCPSF-Gld2-CPEBの複合体構造が一度解除されることで、翻訳抑制の維持ができないmRNPs構造となり、細胞質に存在する別のPAP、もしくは新規合成されたXGld2に

よりポリA鎖伸長が行われ、結果的にMosの翻訳とGVBDの進行が行われたと考えることができる。

初期発生機構研究における本研究の展望

本研究では、十分なCPSFサブユニットのアミノ酸配列が得られなかったため、*Homo sapiens*の母系mRNPsの部分構造しか予測できなかった。また、CPSF複合体構造を始め、本研究で得られた母系mRNPsの予測結果は、生命情報学的手法による予測のみで得られたものであり、これらの実証には実験による確認作業が必要となる。しかし、母系mRNPsの複合体立体構造を、異なる生物種のものどうしで比較することで、母系mRNPsによる翻訳抑制機構のもつ構成因子の変化を受け入れながらも、特定の機能を発揮する多様性を明らかにすることができると考える。複合体の構造によって多様性を受け入れる機構の解明は、常に生物種に固有な現象と生物種に共通して現れる現象が複雑に絡み合う発生生物学領域に、新たな尺度を与えると考える。

複合体機能解析分野における本研究の位置づけと展望

ゲノム塩基配列の読み取りが進むにつれ、生命現象の解析にはRNAおよびタンパク質分子のはたらきがDNAの発現に深く関わり、相互に制御し合うことで複雑な発現制御を行っていることが分かってきた。遺伝子発現の制御に関与するタンパク質群はリボソームに代表されるように複数のタンパク質分子の複合体であるものが多い。また単体として働いていると考えられていたシグナル伝達における酵素でも、一過的には複合体構造を形成することが明らかになっている(El-Habr et al. 2014)。

近年、Cryo-EMのタンパク質の構造研究に用いられるようになったことで、原子レベル

での機能理解を目指す複合体研究がみられるようになってきた(Nguyen et al. 2015, Webb et al. 2015, Lin et al. 2015)。一方、タンパク質の立体構造を用いる研究では、対象分子の実験のしやすさに大きく影響を受ける。本研究で対象とした、CPSFサブユニット、そして複合体としてのCPSFも立体構造を特定するのが困難な分子のひとつである。本研究で開発した複合体構造予測手法の手続きや、得られたサブユニット界面の特徴は、他の複合体の構造予測、相互作用部位予測にも適用することが可能であり、今後広がって行く複合体研究に有用と考えられる。

用語説明

本研究で触れたmRNA上の配列、およびmRNPsサブユニット分子について、その詳細な機能、および構造を示す。

Hex(hexanucleotide)配列

ポリA鎖修飾が行われるmRNAが有する、AAUAAAの6残基の配列(Bilger et al.1994)。この6塩基は3'側非翻訳領域に存在しており、高度に保存されている。

CA配列 (cleavage site) / PA配列 (polyadenylation site)

Hex配列から15~30塩基下流存在する(Zhao et al. 1999)。ポリA鎖の付与に先んじて開裂を受ける2塩基の部位。哺乳類においてはシトシンとアデニンの塩基であることが多い。mRNAのポリA鎖の付与はpre-mRNAの3'末端の一部切り離し、その代わりポリA鎖を付与する。その際このCA配列の間のリン酸基が切断される。またこの部位にポリA鎖が付与されるため、PA配列とも呼ばれる。

CPE配列 (cytoplasmic polyadenylation element)

cytoplasmic polyadenylationを行うmRNAに特異的に見られる配列であり (Hake and Richter 1994)、Hex配列の上流に存在する。UUUUAUまたはUUUUAUの塩基配列である場合が多いが (Weill et al. 2012)、CPE配列には約6~11塩基のヴァリエーションがあり (Fox et al. 1989)、Hex配列とオーバーラップしている場合もある (Piqué et al. 2008)。

CPSF / CPSF複合体 (Cytoplasmic polyadenylation specificity factor)

CPSFは、160kD sbunit(CPSF160)、100kD sbunit (CPSF100)、73kD sbunit (CPSF73)、30kD sbunit (CPSF30) の少なくとも4つ以上のサブユニットからなる複合体である (Bienroth et al. 1991, Murthy and Manley 1992)。

CPSFは母系mRNA前駆体の3'末端に存在するCA部位を開裂し、ポリA鎖付与部位を露出させるはたらきを持つが、この分子機能はもともと核内でポリA鎖伸長を行うmRNAの3'末端の開裂反応として明らかにされ (Rüegsegger et al. 1996)、各サブユニット単体のはたらきについても多くの研究がなされている (Murthy and Manley 1995, Kolev et al. 2008, Mandel et al. 2006, Murthy and Manley 1992)。

CPSF160

CPSF160は160kDのCPSF複合体の中で最も分子量の大きいサブユニットであり、後述するCPSF30およびhFip1とともにmRNAに直接結合している可能性が高いサブユニットである (Kaufmann et al. 2004)。UV-CrosslinkingやリコンビナントCPSF160を用いた研究によって、CPSF160はmRNAのHex配列に特異的に結合することが示されており (Kaufmann et al. 2004, Murthy and Manley 1995)、この結合親和性はCPSF160単体よりCPSF複合体時の方が

高いことが明らかになっている(Murthy and Manley 1995)。またCPSF160はpolyA polymeraseとも結合していることが示されている(Murthy and Manley 1995)。

CPSF160はβプロペラドメインまたはWD40ドメインと呼ばれる、つぶれたような円筒状の構造を3つもつ。この構造は羽根のような扇形の構造を同心円上に複数構成して形作られる。一枚の羽根はWD40 repeatと呼ばれるW(トリプトファン)とD(アスパラギン酸)で終わる約40残基としてアミノ酸配列上にみられる。WD40 repeatは複数のβストランドを形成しており、これらのβストランドが折り畳まれて一枚の羽根型の構造が形成され、各羽根は相互作用し合うことで円盤状の構造を形作っている(図用-1. a)。βプロペラ構造は多くの分子に散見される構造であり、CPSF160の3つのβプロペラドメインはすべて7枚羽根であるが、4から10枚の羽根をもつβプロペラタンパク質が存在する(由良 2013, Faber et al. 1995, Beisel et al. 1999, Fath et al. 2007, Xia et al. 1996, Quistgaard et al. 2009)。βプロペラ構造はシグナル伝達や翻訳制御などのタンパク質どうしの相互作用が重要となる場所ではたらくタンパク質分子によく見られ、一般的に足場タンパク質としての機能があると考えられている(Stirnemann et al. 2010)。Yeast ホモログであるYhh1では2つ目のβ-プロペラ構造でRNAP IIのC末端ドメインに結合することがわかっている(Dichtl et al. 2002)。

Homo sapiens CPSF160はRNA認識モチーフとして保存されているRRM1およびRRM2を有し、RRM1はN末側から379–386残基、RRM2は344–349残基の位置に存在する(Murthy and Manley 1995)。またYhh1のドメイン欠損実験では、500残基から750残基の領域がRNAとの結合に必須であることが示されている(Dichtl et al. 2002)。このYhh1の500残基から750残基の領域はWD40 repeat配列が5つ並んでいる領域にあたり、spliceosomal U2 snRNP (*Homo sapiens* SAP130/SF3B3)と高いアミノ酸配列類似性を示す(Neuwald and Poleksic 2000, Das et al. 1999)。

CPSF100

CPSF100は100kDのCPSFサブユニットであり、metallo-β-lactamase (MBL) スーパーファミリーに属しており、後述するCPSF73と高い類似性があることが以前より示唆されていた(Callebaut et al. 2002, Jenny et al. 1996)。このスーパーファミリーに属するタンパク質の多くはエンドヌクレアーゼ活性をもち、その活性部位に2つの金属イオンを配位している。金属イオンの配位子はHXHXDH(Xは任意のアミノ酸残基)というモチーフとして確認することができる(Dominski et al. 2005)。しかし、CPSF100にはHXHXDHモチーフが欠如しているため、エンドヌクレアーゼ活性を保持しているとは考えにくい。いくつかの研究から、CPSF100の存在がmRNA3'末端の開裂に必要であることは示されてきたが(Kolev et al. 2008, Yang and Doublé 2011)、CPSF100自身の機能は明らかになっていない。哺乳類ではCPSF100はCPSF73と相互作用していることが示されている(Sullivan et al. 2009)。またβ-CASPファミリーのタンパク質群はホモダイマーを形成する(Dominski et al. 2013)ことから、CPSF100がそのホモログであるCPSF73と結合することがCPSFの機能の発揮に参与している可能性が挙げられている(Kolev et al. 2008)。

CPSF100はN末端側にMBLスーパーファミリーに共通した配列である、 β -CASPドメイン配列が挿入されたようなMBLドメイン配列を有し、C末端側のドメインはCPSF100固有の配列をもつ。前述した通りCPSF100はCPSF73と共通祖先由来であり、このCPSF100固有の配列はCPSF73のC末端配列と類似の構造をとるが(Mandel et al. 2006)、アミノ酸配列の全長はCPSF100の方が100残基ほど長い。CPSF73、CPSF100に特徴的なC末端領域の立体構造は決定されておらず、類似の配列を持つ立体構造もないため、この領域がどのような構造を作るのかは分かっていない。*Methanosarcina mazei*のCPSF73-CPSF100ホモログであるArchaeal CPSFでは、CPSF73やCPSF100が有する様なC末端領域はなく、N末端側にCPSF73およびCPSF100にはない約180残基の配列で構成されるKH-domainを有している(Silva et al. 2011)。ArchaeaではMBL superfamilyに属すCPSFホモログはこのArchaeal CPSFしか見つかっておらず、CPSF73とCPSF100の両方の機能を備えていると考えられており(Albrecht and Wagner. 2012)、KH-domainはmRNAの認識部位であるという考察が行われている(Xiang et al. 2014)。

CPSF73

CPSF73は73kDのCPSFサブユニットであり、CPSF100同様MBLスーパーファミリーの β -CASPファミリーに属す。CPSF73のyeast homologであるYsh1をもちいた実験において亜鉛イオン配位残基の変異が致死的事であることや、Hela細胞の実験系において、キレート材を使用するとmRNA前駆体のポリA鎖伸長に先駆けて起こる3'末端の開裂が阻害される (Ryan et al. 2004)ことから、CPSF複合体の中でCPSF73が唯一mRNAの3'末端の切断を行っている分子であると考えられている(Callebaut et al. 2002)。CPSF73はN末端側1-460残基に β -CASPドメインを挿入した形でMBLドメインをもち、C末側は他のMBLスーパーファミリーにはない、CPSF73に特有なドメインをもつ。endonuclease活性を有する部位はMBLドメイン中にあり、活性部位には2つの亜鉛イオンが7つのアミノ酸残基に配位しており、求核置換反応によりリン酸結合の開裂を触媒する(Mandel et al. 2006)。

Mandelらによって決定されたCPSF73の結晶構造では活性部位はタンパク質分子内部に埋まっておらず、mRNAの開裂反応が抑制されていることが示唆された。CPSF73のMBLドメインを用いたエンドヌクレアーゼ活性実験では、想定されていたよりRNA切断活性が低く、CaCl₂で処理することで活性が高くなることが示されている(Mandel et al. 2006)。カルシウムイオンがCPSF73に何らかの構造変化を引き起こし、活性部位が露出するなどの変化が起きることが示唆されているが、この活性化機構については明らかにされていない(Xiang et al. 2014)。

CPSF30

CPSF30は30kDのCPSFの中で最も小さいサブユニットであるが、様々な機能が示唆されている。mRNA前駆体の3'末端開裂とポリA鎖付加に必須なアミノ酸配列とされる5個のCCCH Zinc fingerモチーフが、菌類および動植物類で高度に保存されており(Barabino et al.

1997)、動物門のCPSF30では5個のCCCH Zinc fingerモチーフに加え、C末端側にZinc knuckleモチーフが保存されている。

Zinc fingerおよびZinc knuckleはDNA結合モチーフとして知られるが、これらの構造によってRNAに結合する分子が知られている(Barabino et al. 1997, Xiang et al. 2014)。特にこれらの構造はU-rich配列に結合しやすく、核内polyadenylationではHex配列の5'側に存在するU-rich配列であるUSE(upstream element)にCPSF30が結合していると考えられている(Chan et al. 2011)。CPSF30のUSEへの結合はCPSF160の選択的なmRNA認識を補助している可能性があげられている(Kaufmann et al. 2004)。

Yth1の4つ目のzinc fingerモチーフは後述するFip1との結合を担う(Barabino et al. 1997, Takahashi et al. 2003)。またCPSFはCPSF30を介して真核生物において転写の実行を行うRNAP IIとの結合していることが示されている(Nag et al. 2007)。このようにpolyadenylationに関わる複合体中でCPSF30は他分子との結合に重要な役割を担っており、その立体構造を決定することはポリA鎖伸長反応機構の解明に有用であると考えられるが、2つ目と3つ目のzinc fingerモチーフを含む限られた領域しか立体構造が決定されていない(Montelione et al. 2008, PDBID: 2RHK)。

Fip1

本研究ではCPSF複合体のサブユニットをCPSF160、CPSF100、CPSF73、CPSF30の4つとしたが、Fip1をCPSF複合体の5つ目のサブユニットをとみなす文献も多い(Nag et al. 2007, Chan et al. 2011, Laishram and Anderson 2010)。Fip1が初めて同定されたのはYeastであり、Yeast Fip1はGld2と結合する(Meinke et al. 2008)。しかしYeast Fip1はCPSFサブユニットのホモログとは複合体を形成していない(Millevoi and Vagner 2010)。また*Homo sapiens* Fip1(hFip1)はYeast Fip1に比べ長く2つのドメインがC末端側に追加されているため、その構造の違いからhFip1をhFip1として区別して記載する場合もある(Kaufmann et al. 2004)。

Fip1は核内polyadenylationの場ではN末端領域でCPSF160およびCPSF30と結合し(Kaufmann et al. 2004)、C末端領域でRNAのU-rich配列に結合することが知られている(Chan et al. 2011)。しかし細胞質でpolyadenylationを行う母系mRNPsの構成要素としてFip1は未だ同定されていないため、本研究ではFip1はCPSFのサブユニットとして考えないこととした。

WDR33

WDR33は146kDのタンパク質であり、N末側にCPSF160がもつようなWD repeat領域をもつタンパク質である(Ito et al. 2001)。WDR33はCPSF73の免疫沈降実験によって得られ、WDR33を欠乏させるとmRNA前駆体の3'末端の開裂とポリA鎖伸長が阻害される(Shi et al. 2009)。またWDR33のYeastホモログであるPft2がCPSF73のホモログであるysh1と強く結合することが分かっている(Ohnacker et al. 2000)。しかしWDR33はCPSF複合体のサブユニットとはしない文献(Millevoi et al. 2010)とCPSF複合体のサブユニットとする文献(Xiang et al.

2014)が混在するため、WDR33も本研究ではCPSFのサブユニットとして考えないこととした。

CPSF73(II) / CPSF73-II

*Arabidopsis thaliana*ではCPSF160、CPSF100、CPSF73、CPSF30のCPSFサブユニット以外に、CPSF73に類似のCPSF73(II)もCPSF複合体としてのサブユニットとして存在する(Xu et al. 2004, Millevoi and Vagner 2010)。

CPEB (Cytoplasmic polyadenylation element binding protein)

CPEBはRNA認識モチーフ(RRM)を2つもち(Fernández-Miranda and Méndez. 2012)、mRNA前駆体上にあるCPE配列(UUUUUAU, UUUUAAU)と結合する(Hake and Richter 1994)。CPEBは核内、特に紡錘体上に多数存在し、母系mRNAが合成されるとともに、そのCPE配列に結合すると考えられている(Eliscovich et al. 2008, Groisman et al. 2000)。核外に輸送された母系mRNAはCPEBを中心に貯蔵型の母系mRNPsを形成する。CPEBはCPSFおよびpoly(A) polymerase Gld2とも結合しており(Mendez et al. 2000b)、細胞質内polyadenylationの活性化はCPEBのリン酸化によって起こる(Mendez et al. 2000a)。このリン酸化はマウスではAurora Kinaseによって行われるが、*Xenopus leavis*の卵ではCDK1が最も強力な候補とされている(Charlesworth et al. 2013, Keady et al. 2007)。

CPEBは過リン酸化により細胞中に蓄積されていた全体量の70~90%が第二減数分裂への導入期に分解される(de Moor et al. 1997)。これは比較的遅い時期に発現するmRNAのpolyadenylationにCPEBの分解が関与する結果だと考えられている (Mendez et al. 2002, Setoyama et al. 2007)。

Poly(A) polyadenylase (PAP: Gld2)

最もよく研究されているPAPは核内polyadenylationを担うPAP α であり、核に局在する(Xiang et al. 2014)。卵細胞質内でpolyadenylationを担うのはGld2/PAPD4というpoly(A) polymeraseである(Rouhana et al. 2005)。Gld2は*Caenorhabditis Elegans*の減数分裂や有糸分裂をはじめとする生殖細胞でおこる発生現象の調節因子の1つとして見つかった(Kadyk and Kimble. 1998)。Gld2はそれ自体のpoly(A) polymerase活性は低いことが分かっており、Gld3と結合することでpolymerase活性を有するようになる(Wang et al. 2002, Kwak et al. 2004)。

Poly(A) ribonuclease (PARN)

PARNはポリA鎖特異的なribonucleaseでありCPEBおよびGld2と相互作用する (図序-2, Kim and Richter. 2006)。CPEBおよびCPSFは母系 mRNAが転写されるとすぐにその3' UTR領域に結合する(Eliscovich et al. 2008)が、PARNとGld2は母系mRNAが細胞質に出てから母系 mRNPsの構成タンパク質となる(2006, Lin et al. 2010)。

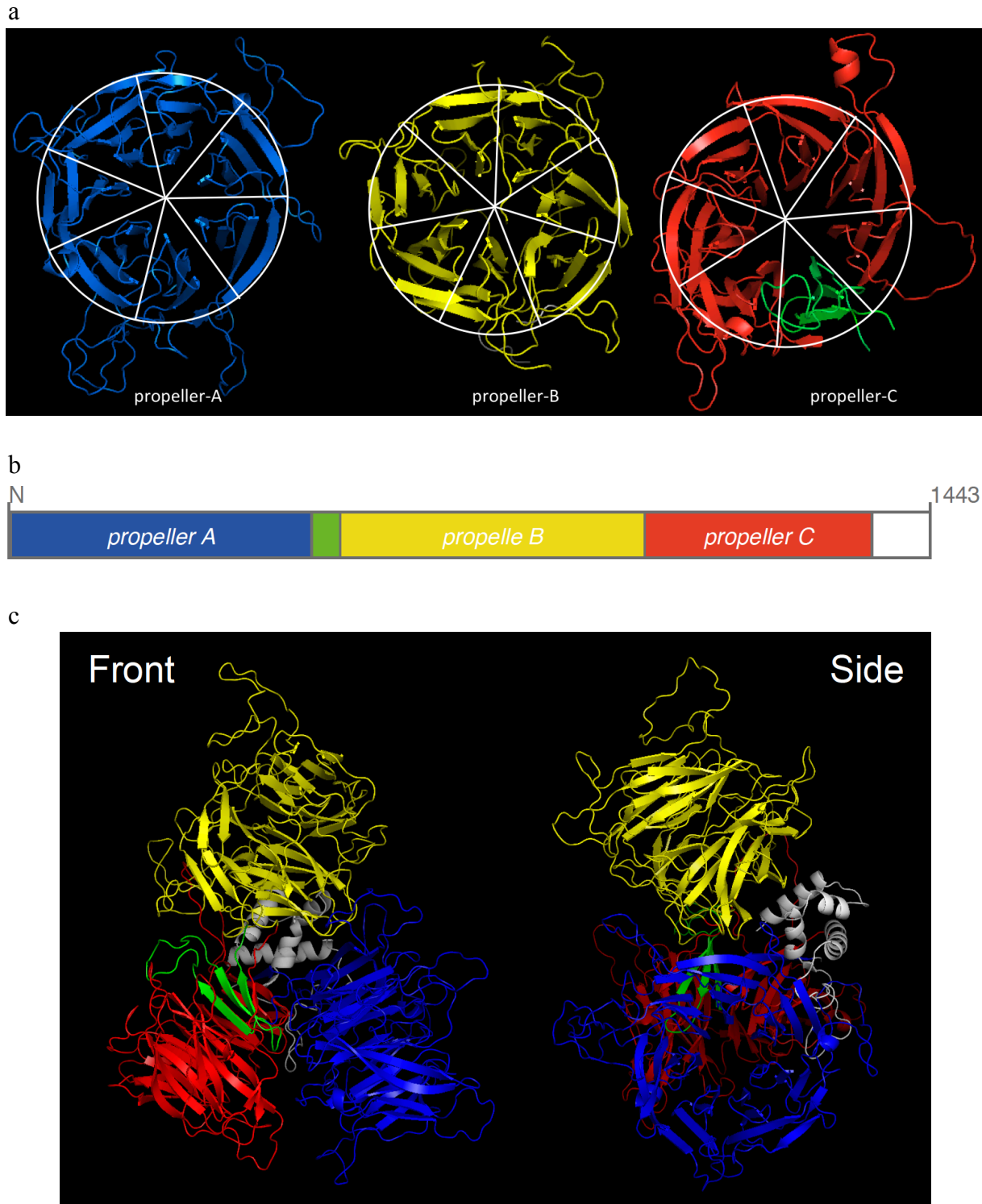
母系mRNPsではGld2とPARNは同時に活性をもっているが、互いの働きが拮抗することで母系mRNAのポリA鎖の伸長が抑制され、ひいては翻訳が抑制される (Kim and Richter.

2006, Lin et al. 2010)。CPEBがリン酸化されることで貯蔵型mRNPからPARNが遊離し、母系mRNPsは翻訳活性型になる(Kim and Richter 2006)。

embryonic poly(A) binding protein (ePAB)

ePABは*Xenopus leavis*の卵で発見された (Voeltz et al. 2001)。ePABの機能はポリA鎖に結合することでPARNなどのヌクレアーゼからポリA鎖を保護し(Kim and Richter 2008, Kim and Richter 2007)、mRNAのキャップ部分に結合しているeIF4E-eIF4G複合体と結合することでポリA鎖の伸長を促進する (Kim and Richter 2008, Kim and Richter 2007, Voeltz et al. 2001)。ePABは貯蔵型RNPではCPEBと結合しているが、ポリA鎖の伸長が行われるとポリA鎖に結合するようになる(Kim and Richter 2007)。

ePABは卵内では高い発現量が認められるが、初期胚時には検出されなくなり、それ以降はPABP1がポリA鎖を保護する(Voeltz et al. 2001)。



図用-1. 3つのプロペラドメインで構成されるCPSF160の立体構造

N末端側からpropeller-A(青)、propeller-B(黄)、propeller-C(赤)を示す。CPSF160はその全体構造が3つのプロペラドメインにより構成されており、全ての7枚羽根のプロペラ構造である。propeller-A、propeller-Bは連続する配列により構成されているが、propeller-Cは緑で示す1枚の羽根部分がpropeller-Aとpropeller-Bの間に位置する配列で形成されており、連続した配列で形成されていない(a)。CPSF160のアミノ酸配列における各プロペラドメインの位置をボックスで示す(b)。各プロペラドメイン三次元的なCPSF160立体構造における位置関係。

参考文献

- 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56-65.
- Abramić M, Levine AS, Protić M. 1991. Purification of an ultraviolet-inducible, damage-specific DNA-binding protein from primate cells. *J Biol Chem* 266:22493-22500.
- Adams J. 2003. The proteasome: structure, function, and role in the cell. *Cancer Treat Rev* 29 Suppl 1:3-9.
- Aderem A, Ulevitch RJ. 2000. Toll-like receptors in the induction of the innate immune response. *Nature* 406:782-787.
- Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprpto A, Karni-Schmidt O, Williams R, Chait BT, Sali A, Rout MP. 2007. The molecular architecture of the nuclear pore complex. *Nature* 450:695-701.
- Albrecht TR, Wagner EJ. 2012. snRNA 3' end formation requires heterodimeric association of integrator subunits. *Mol Cell Biol* 32:1112-1123.
- Aloy P, Russell RB. 2002. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci USA* 99:5896-5901.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403-410.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.
- Anfinsen CB, Haber E, Sela M, White FH Jr. 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci USA* 47:1309-1314.
- Andrusier N, Nussinov R, Wolfson HJ. 2007. FireDock: fast interaction refinement in molecular docking. *Proteins* 69:139-159.
- Aoto S, Yura K. 2015. Case study on the evolution of hetero-oligomer interfaces based on the differences in paralogous proteins. *Biophysico* 12:103-117.
- August JT, Ortiz PJ, Hurwitz J. 1962. Ribonucleic acid-dependent ribonucleotide incorporation. I. Purification and properties of the enzyme. *J Biol Chem* 237:3786-3793.
- Bai XC, McMullan G, Scheres SH. 2015. How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* 40:49-57.
- Baillat D, Hakimi MA, Naar AM, Shilatifard A, Cooch N, Shiekhatar R. 2005. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell* 123:265-276.

- Barabino SM, Hübner W, Jenny A, Minvielle-Sebastia L, Keller W. 1997. The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins. *Genes Dev* 11:1703-1716.
- Barkoff AF, Dickson KS, Gray NK, Wickens M. 2000. Translational control of cyclin B1 mRNA during meiotic maturation: coordinated repression and cytoplasmic polyadenylation. *Dev Biol* 220:97-109.
- Beisel HG, Kawabata S, Iwanaga S, Huber R, Bode W. 1999. Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab *Tachypleus tridentatus*. *EMBO J* 18:2313-2322.
- Belloc E, Méndez R. 2008a. A deadenylation negative feedback mechanism governs meiotic metaphase arrest. *Nature* 452:1017-1021.
- Belloc E, Piqué M, Méndez R. 2008b. Sequential waves of polyadenylation and deadenylation define a translation circuit that drives meiotic progression. *Biochem Soc Trans* 36:665-670.
- Ben-Tabou de-Leon S, Davidson EH. 2007. Gene regulation: gene control network in development. *Annu Rev Biophys Biomol Struct* 36:191. Review.
- Berggård T, Linse S, James P. 2007. Methods for the detection and analysis of protein-protein interactions. *Proteomics*. 7:2833-2842.
- Berman HM, Kleywegt GJ, Nakamura H, Markley JL. 2013. How community has shaped the Protein Data Bank. *Structure* 21:1485-1491.
- Berman HM, Kleywegt GJ, Nakamura H, Markley JL. 2014. The Protein Data Bank archive as an open data resource. *J Comput Aided Mol* 28:1009-1014
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* 28:235-242.
- Berman HM, Henrick K, Nakamura H. 2003. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10:980.
- Bienroth S, Keller W, Wahle E. 1993. Assembly of a processive messenger RNA polyadenylation complex. *EMBO J* 12:585-594.
- Bienroth S, Wahle E, Suter-Crazzolara C, Keller W. 1991. Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors. *J Biol Chem* 266:19768-19776.
- Bilger A1, Fox CA, Wahle E, Wickens M. 1994. Nuclear polyadenylation factors recognize cytoplasmic polyadenylation elements. *Genes Dev* 8:1106-1116.
- Bossi RT, Negri A, Tedeschi G, Mattevi A. 2002. Structure of FAD-bound L-aspartate oxidase: insight into substrate specificity and catalysis. *Biochemistry* 41:3018-3024.
- Bowen B, Steinberg J, Laemmli UK, and Weintraub H. 1980. The detection of DNA-binding proteins by protein blotting. *Nucleic Acids Res* 8:1-20.
- Brenner SE. 1999. Errors in genome annotation. *Trends Genet* 15:132-133.

- Brymora A, Valova VA, Robinson PJ. 2004. Protein-protein interactions identified by pull-down experiments and mass spectrometry. *Curr Protoc Cell Biol* Chapter 17:Unit 17.5.
- Callebaut I, Moshous D, Mornon JP, de Villartay JP. 2002. Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Res* 30:3592-3601.
- Camacho CJ, Weng Z, Vajda S, DeLisi C. 1999. Free energy landscapes of encounter complexes in protein-protein association. *Biophys J* 76:1166-1178.
- Chakrabarti P, Janin J. 2002. Dissecting protein-protein recognition sites. *Proteins* 47:334-343.
- Chan S, Choi EA, Shi Y. 2011. Pre-mRNA 3'-end processing complex assembly and function. *Wiley Interdiscip Rev RNA* 2:321-335.
- Charlesworth A, Meijer HA, de Moor CH. Specificity factors in cytoplasmic polyadenylation. 2013. *Wiley Interdiscip Rev RNA* 4:437-461.
- Chen J, Wagner EJ. 2010. snRNA 3' end formation: the dawn of the Integrator complex. *Biochem Soc Trans* 38:1082-1087.
- Chen R, Li L, Weng Z. 2003. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* 52:80-87.
- Chen R, Weng Z. 2002. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins* 47:281-294.
- Cheng J, Sweredoski M, Baldi P. 2005. Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and Knowledge Discovery* 11:213-222
- Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J* 5:823-826.
- Comeau SR, Gatchell DW, Vajda S, Camacho CJ. 2004. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res* 32:W96-99.
- Contreras-Moreira B, Fitzjohn PW, Bates PA. 2003. In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. *J Mol Biol* 328:593-608.
- Covarrubias AS, Högbom M, Bergfors T, Carroll, P, Mannerstedt K, Oscarson S, Parish T, Jones TA, Mowbray SL. 2008. Structural, biochemical, and in vivo investigation of the threonine synthase from *Mycobacterium tuberculosis*. *J Mol Biol* 381:622-633.
- David A, Razali R, Wass MN, Sternberg MJ. 2012. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum Mutat* 33:359-363.
- Das BK, Xia L, Palandjian L, Gozani O, Chyung Y, Reed R. 1999. Characterization of a protein complex containing spliceosomal proteins SAPs 49, 130, 145, and 155. *Mol Cell Biol* 19:6796-6802.
- Das K, Ma LC, Xiao R, Radvansky B, Aramini J, Zhao L, Marklund J, Kuo RL, Twu KY, Arnold E, Krug RM, Montelione GT. 2008. Structural basis for suppression of a host antiviral response by influenza A virus. *Proc Natl Acad Sci USA* 105:13093-13098.

- DeLano WL. 2008. The PyMOL Molecular Graphics System. Palo Alto CA USA: DeLano Scientific LLC.
- Dellis S, Strickland KC, McCrary WJ, Patel A, Stocum E, Wright CF. 2004. Protein interactions among the vaccinia virus late transcription factors. *Virology* 329:328-336.
- de Moor CH, Richter JD. 1997. The Mos pathway regulates cytoplasmic polyadenylation in *Xenopus* oocytes. *Mol Cell Biol* 17:6419-6426.
- Dichtl B, Blank D, Sadowski M, Hübner W, Weiser S, Keller W. 2002. Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination. *EMBO J* 21:4125-4135.
- Ding L, Becker AB, Suzuki A, Roth RA. 1992. Comparison of the enzymatic and biochemical properties of human insulin-degrading enzyme and *Escherichia coli* protease III. *J Biol Chem* 267:2414-2420.
- Dominski Z, Carpousis AJ, Clouet-d'Orval B. 2013. Emergence of the β -CASP ribonucleases: highly conserved and ubiquitous metallo-enzymes involved in messenger RNA maturation and degradation. *Biochim Biophys Acta*. 1829:532-551
- Dominski Z, Yang XC, Purdy M, Wagner EJ, Marzluff WF. 2005. A CPSF-73 homologue is required for cell cycle progression but not cell growth and interacts with a protein having features of CPSF-100. *Mol Cell Biol* 25:1489-1500.
- Dominski Z, Marzluff WF. 2007. Formation of the 3' end of histone mRNA: getting closer to the end. *Gene*. 396:373-390.
- Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21:3433-3434
- Douangamath A, Walker M, Beismann-Driemeyer S, Vega-Fernandez MC, Sterner R, Wilmanns M. 2002. Structural evidence for ammonia tunneling across the $(\beta\alpha)_8$ barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Structure* 10:185-193.
- Dworkin MB, Shrutkowski A, Dworkin-Rastl E. 1985. Mobilization of specific maternal RNA species into polysomes after fertilization in *Xenopus laevis*. *Proc Natl Acad Sci USA* 82:7636-7640.
- El-Habr EA, Levidou G, Trigka EA, Sakalidou J, Piperi C, Chatziandreu I, Spyropoulou A, Soldatos R, Tomara G, Petraki K, Samaras V, Zisakis A, Varsos V, Vrettakos G, Boviatsis E, Patsouris E, Saetta AA, Korkolopoulou P. 2014. Complex interactions between the components of the PI3K/AKT/mTOR pathway, and with components of MAPK, JAK/STAT and Notch-1 pathways, indicate their involvement in meningioma development. *Virchows Arch* 465:473-485.
- Eliscovich C, Peset I, Vernos I, Méndez R. 2008. Spindle-localized CPE-mediated translation controls meiotic chromosome segregation. *Nat Cell Biol* 10:858-865.
- Eriksson S, Cederlund E, Bergman T, Jörnvall H, Bohman C. 1991. Characterization of human deoxycytidine kinase. Correlation with cDNA sequences. *FEBS Lett* 280:363-366.

- Ettensohn CA. 2013. Encoding anatomy: developmental gene regulatory networks and morphogenesis. *Genesis* 51:383-409. Review.
- Faber HR, Groom CR, Baker HM, Morgan WT, Smith A, Baker EN. 1995. A crystal structure of the C-terminal domain of rabbit serum haemopexin. *Structure* 3:551-559.
- Fath S, Mancias JD, Bi X, Goldberg J. 2007. Structure and organization of coat proteins in the COPII cage. *Cell* 129:1325-1336.
- Fernández-Miranda G, Méndez R. 2012. The CPEB-family of proteins, translational control in senescence and cancer. *Ageing Res Rev* 11:460-472.
- Ferron F, Longhi S, Canard B, Karlin D. 2006. A practical overview of protein disorder prediction methods. *Proteins* 65:1-14. Review.
- Fields S, Song O. 1989. A novel genetic system to detect protein-protein interactions. *Nature* 340:245-246.
- Fischer ES, Scrima A, Böhm K, Matsumoto S, Lingaraju GM, Faty M, Yasuda T, Cavadini S, Wakasugi M, Hanaoka F, Iwai S, Gut H, Sugawara K, Thomä NH. 2011. The molecular basis of CRL4DDB2/CSA ubiquitin ligase architecture, targeting, and activation. *Cell* 147:1024-1039.
- Fox CA, Sheets MD, Wickens MP. 1989. Poly(A) addition during maturation of frog oocytes: distinct nuclear and cytoplasmic activities and regulation by the sequence UUUUUAU. *Genes Dev* 3:2151-2162.
- Fryxell KJ. 1996. The coevolution of gene family trees. *Trends Genet* 12:364-369.
- Gavin AC, Bösch M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Höfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edlmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141-147.
- Ginalski K, Elofsson A, Fischer D, Rychlewski L. 2003 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics* 19:1015-1018.
- Go M, Miyazawa S. 1980. Relationship between mutability, polarity and exteriority of amino acid residues in protein evolution. *Int J Pept Protein Res* 15:211-224.
- Goodsell DS, Olson AJ. 2000 Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct* 29:105-153.
- Groisman I, Huang YS, Mendez R, Cao Q, Theurkauf W, Richter JD. 2000. CPEB, maskin, and cyclin B1 mRNA at the mitotic apparatus: implications for local translational control of cell division. *Cell* 103:435-447.
- Hake LE, Richter JD. 1994 CPEB is a specificity factor that mediates cytoplasmic polyadenylation during *Xenopus* oocyte maturation. *Cell* 79:617-627.

- Han S, Xie W, Hammes SR, DeJong J. 2003. Expression of the germ cell-specific transcription factor ALF in *Xenopus* oocytes compensates for translational inactivation of the somatic factor TFIIA. *J Biol Chem* 278:45586-45593.
- Henn-Sax M, Thoma R, Schmidt S, Hennig M, Kirschner K, Sterner R. 2002. Two (β)₈-barrel enzymes of histidine and tryptophan biosynthesis have similar reaction mechanisms and common strategies for protecting their labile substrates. *Biochemistry* 41:12032-12042.
- Hijikata A, Yura K, Noguti T, Go M. 2011. Revisiting gap locations in amino acid sequence alignments and a proposal for a method to improve them by introducing solvent accessibility. *Proteins* 79:1868-1877
- Hirose S, Shimizu K, Kanai S, Kuroda Y, Noguchi T. 2007. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* 23:2046-2053.
- Hryciw T, Tang M, Fontanie T, Xiao W. 2002. MMS1 protects against replication-dependent DNA damage in *Saccharomyces cerevisiae*. *Mol Genet Genomics* 266:848-857.
- Huarte J, Stutz A, O'Connell ML, Gubler P, Belin D, Darrow AL, Strickland S, Vassalli JD. 1992. Transient translational silencing by reversible mRNA deadenylation. *Cell* 69:1021-1030.
- Huarte J, Belin D, Vassalli A, Strickland S, Vassalli JD. 1987. Meiotic maturation of mouse oocytes triggers the translation and polyadenylation of dormant tissue-type plasminogen activator mRNA. *Genes Dev* 1:1201-1211.
- Huynen MA, Bork P. 1998. Measuring genome evolution. *Proc Natl Acad Sci USA* 95:5849-5856.
- Ito S, Sakai A, Nomura T, Miki Y, Ouchida M, Sasaki J, Shimizu K. 2001. A novel WD40 repeat protein, WDC146, highly expressed during spermatogenesis in a stage-specific manner. *Biochem Biophys Res Commun*. 280:656-663.
- Janin J, Bahadur RP, Chakrabarti P. 2008. Protein-protein interaction and quaternary structure. *Quarterly Reviews of Biophysics* 41:133-180.
- Jones S, Thornton JM. 1996. Principles of protein-protein interactions. *Proc Natl Acad Sci USA* 93:13-20.
- Kadyk LC, Kimble J. 1998. Genetic regulation of entry into meiosis in *Caenorhabditis elegans*. *Development* 125:1803-1813.
- Kashiwagi K, Kashiwagi A, Kurabayashi A, Hanada H, Nakajima K, Okada M, Takase M, Yaoita Y. 2010. *Xenopus tropicalis*: an ideal experimental animal in amphibia. *Exp Anim* 59:395-405. Review.
- Kato YS, Yagi T, Harris SA, Ohki SY, Yura K, Shimizu Y, Honda S, Kamiya R, Burgess SA, Tanokura M. 2014. Structure of the microtubule-binding domain of flagellar dynein. *Structure* 22:1628-1638.
- Kaufmann I, Martin G, Friedlein A, Langen H, Keller W. 2004. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J* 23:616-626.

- Keady BT, Kuo P, Martínez SE, Yuan L, Hake LE. 2007. MAPK interacts with XGef and is required for CPEB activation during meiosis in *Xenopus* oocytes. *J Cell Sci* 120:1093-1103.
- Keeney S, Chang GJ, Linn S. 1993. Characterization of a human DNA damage binding protein implicated in xeroderma pigmentosum E. *J Biol Chem* 268:21293-21300.
- Kim H, Lee JH, Lee Y. 2003. Regulation of poly(A) polymerase by 14-3-3epsilon. *EMBO J* 22:5208-5219.
- Kim JH, Richter JD. 2006. Opposing polymerase-deadenylase activities regulate cytoplasmic polyadenylation. *Mol Cell* 24:173-183.
- Kim JH, Richter JD. 2007. RINGO/cdk1 and CPEB mediate poly(A) tail stabilization and translational regulation by ePAB. *Genes Dev* 21:2571-2579.
- Kim JH, Richter JD. 2008. Measuring CPEB-mediated cytoplasmic polyadenylation-deadenylation in *Xenopus laevis* oocytes and egg extracts. *Methods Enzymol.* 448:119-38.
- Kim OT, Yura K, Go N. 2006a. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res* 34:6450-6460.
- Kim EY, Schrader N, Smolinsky B, Bedet C, Vannier C, Schwarz G, Schindelin H. 2006b. Deciphering the structural framework of glycine receptor anchoring by gephyrin. *EMBO J* 25:1385-1395.
- Klem T, Davisson VJ. 1993. Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis. *Biochemistry* 32:5177-5186.
- Klupp BG, Böttcher S, Granzow H, Kopp M, Mettenleiter TC. 2005. Complex formation between the UL16 and UL21 tegument proteins of pseudorabies virus. *J Virol* 79:1510-1522.
- Kobayashi E, Yura K, Nagai Y. 2013. Distinct conformation of ATP molecule in solution and on protein. *Biophysics* 9:1-12.
- Kolev NG, Yario TA, Benson E, Steitz JA. 2008. Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3'-end maturation. *EMBO Rep* 9:1013-1018.
- Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM. 2003. A "Frankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins* 6:369-379.
- Koshland DE Jr, Hamadani K. 2002. Proteomics and models for enzyme cooperativity. *J Biol Chem* 277:46841-46844.
- Kumar S, Dudley JT, Filipinski A, Liu L. 2011. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet* 27:377-386.
- Kwak JE, Wang L, Ballantyne S, Kimble J, Wickens M. 2004. Mammalian GLD-2 homologs are poly(A) polymerases. *Proc Natl Acad Sci USA* 101:4407-4412.

- Kyburz A, Friedlein A, Langen H, Keller W. 2006. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. *Mol Cell* 23:195-205.
- Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157:105-132.
- Laber B, Maurer W, Hanke C, Gräfe S, Ehlert, Messerschmidt A, Causen T. 1999. Characterization of recombinant *Arabidopsis thaliana* threonine synthase. *Eur J Biochem*, 263:212-221.
- Laishram RS, Anderson RA. 2010. The poly A polymerase Star-PAP controls 3'-end cleavage by promoting CPSF interaction and specificity toward the pre-mRNA. *EMBO J* 29:4132-45.
- Lang D, Thoma R, Henn-Sax M, Sterner R, Wilmanns M. 2000. Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science* 289: 1546-1550.
- Larsen TA, Olson AJ, Goodsell DS. 1998. Morphology of protein-protein interfaces. *Structure* 6:421-427.
- Li E, Davidson EH. 2009. Building developmental gene regulatory networks. *Birth Defects Res C Embryo Today* 87:123-30.
- Li H, Xie B, Rahmeh A, Zhou Y, Lee MY. 2006. Direct interaction of p21 with p50, the small subunit of human DNA polymerase delta. *Cell Cycle* 5:428-36.
- Li L, Chen R, Weng Z. 2003. RDOCK: refinement of rigid-body protein docking predictions. *Proteins* 53:693-707.
- Lin CL, Evans V, Shen S, Xing Y, Richter JD. 2010. The nuclear experience of CPEB: implications for RNA processing and translational control. *RNA*. 16:338-48..
- Lin DY, Huang S, Chen J. 2015. Crystal structures of a polypeptide processing and secretion transporter. *Nature* 523:425-30.
- Lin R. 2010. Maternal mRNA and the PolyA Tail. *Nature Education* 3:47
- Lindert S, Stewart PL, Meiler J. 2009. Hybrid approaches: applying computational methods in cryo-electron microscopy. *Curr Opin Struct Biol*. 19:218-25.
- Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003a. Protein disorder prediction: implications for structural proteomics. *Structure* 11:1453-9.
- Linding R, Russell RB, Neduva V, Gibson TJ. 2003b. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 31:3701-8
- Livstone MS, Oughtred R, Heinicke S, Vernot B, Huttenhower C, Durand D, Dolinski K. 2011. Inferring protein function from homology using the Princeton Protein Orthology Database (P-POD). *Curr Protoc Bioinformatics Chapter 6:Unit 6.11*.
- Lo Conte L, Chothia C, Janin J. 1999. The atomic structure of protein-protein recognition sites. *J Mol Biol* 285:2177-98.
- Lorenzen S, Zhang Y. 2007. Identification of near-native structures by clustering protein docking conformations. *Proteins* 68:187-94.

- Lapasset L, Pradet-Balade B, Lozano JC, Peaucellier G, Picard A. 2005. Nuclear envelope breakdown may deliver an inhibitor of protein phosphatase 1 which triggers cyclin B translation in starfish oocytes. *Dev Biol* 285:200-10.
- Lyskov S, Gray JJ. 2008. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res* 36(Web Server issue):W233-8.
- Malumbres M, Mateos LM, Lumbreras MA, Guerrero C, Martin JF. 1994. Analysis and expression of the thrC gene of *Brevibacterium lactofermentum* and characterization of the encoded threonine synthase. *Appl Environ Microbiol* 60:2209-2219.
- Mandel CR, Bai Y, Tong L. 2008. Protein factors in pre-mRNA 3'-end processing. *Cell Mol Life Sci* 65:1099-122. Review.
- Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L. 2006. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* 444:953-6.
- Markley JL, Ulrich EL, Westler WM, Volkman BF. 2003. Macromolecular structure determination by NMR spectroscopy. *Methods Biochem Anal.* 44:89-113.
- Martin G, Gruber AR, Keller W, Zavolan M. 2012. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep.* 1:753-63.
- Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. 2000. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29:291-325.
- Meier M, Janosik M, Kery V, Kraus JP, Burkhard P. 2001. Structure of human cystathionine beta-synthase: a unique pyridoxal 5'-phosphate-dependent heme protein. *EMBO J* 20:3910-3916.
- Meijer HA, Radford HE, Wilson LS, Lissenden S, de Moor CH. 2007. Translational control of maskin mRNA by its 3' untranslated region. *Biol Cell* 99:239-50.
- Melino S, Capo C, Dragani B, Aceto A, Petruzzelli R. 1998. A zinc-binding motif conserved in glyoxalase II, beta-lactamase and arylsulfatases. *Trends Biochem Sci* 23:381-2.
- Mendez R, Barnard D, Richter JD. 2002. Differential mRNA translation and meiotic progression require Cdc2-mediated CPEB destruction. *EMBO J* 21:1833-44.
- Mendez R, Hake LE, Andresson T, Littlepage LE, Ruderman JV, Richter JD. 2000. Phosphorylation of CPE binding factor by Eg2 regulates translation of c-mos mRNA. *Nature* 404:302-7.
- Mendez R, Murthy KG, Ryan K, Manley JL, Richter JD. 2000. Phosphorylation of CPEB by Eg2 mediates the recruitment of CPSF into an active cytoplasmic polyadenylation complex. *Mol Cell* 6:1253-9.
- von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417:399-403.
- Mikkelsen NE, Munch-Petersen B, Eklund H. 2008. Structural studies of nucleoside analog and feedback inhibitor binding to *Drosophila melanogaster* multisubstrate deoxyribonucleoside kinase. *FEBS J* 275:2151-2160.

- Millevoi S, Vagner S. 2010. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Res* 38:2757-74.
- Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z. 2007. Integrating statistical pair potentials into protein complex prediction. *Proteins* 69:511-20.
- Mir-Montazeri B, Ammelburg M, Forouzan D, Lupas AN, Hartmann MD. 2011. Crystal structure of a dimeric archaeal cleavage and polyadenylation specificity factor. *J Struct Biol* 173:191-5.
- Mohanty BK, Kushner SR. 2011. Bacterial/archaeal/organelle polyadenylation. *Wiley Interdiscip Rev RNA* 2:256-76. Review.
- Mortarino M, Negri A, Tedeschi G, Simonic T, Duga S, Gassen HG, Ronchi, S. 1996. L-aspartate oxidase from *Escherichia coli* I. Characterization of coenzyme binding and product inhibition. *Eur J Biochem* 239:418-426.
- Mullaney BP, Pallavicini MG. 2001. Protein-protein interactions in hematology and phage display. *Exp Hematol* 29:1136-46.
- Munch-Petersen B, Piskur J, Søndergaard L. 1998. Four deoxynucleoside kinase activities from *Drosophila melanogaster* are contained within a single monomeric enzyme, a new multifunctional deoxynucleoside kinase. *J Biol Chem* 273:3926-3931.
- Murthy KG, Manley JL. 1992. Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. *J Biol Chem* 267:14804-11.
- Murthy KG, Manley JL. 1995. The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. *Genes Dev* 9:2672-83.
- Nakel K, Bonneau F, Eckmann CR, Conti E. 2015. Structural basis for the activation of the *C. elegans* noncanonical cytoplasmic poly(A)-polymerase GLD-2 by GLD-3. *Proc Natl Acad Sci USA* 112:8614-9.
- Nag A, Narsinh K, Martinson HG. 2007. The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. *Nat Struct Mol Biol* 14:662-9.
- Neuwald AF, Poleksic A. 2000. PSI-BLAST searches using hidden markov models of structural repeats: prediction of an unusual sliding DNA clamp and of beta-propellers in UV-damaged DNA-binding protein. *Nucleic Acids Res* 28:3570-80.
- Nguyen TH, Galej WP, Bai XC, Savva CG, Newman AJ, Scheres SH, Nagai K. 2015. The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* 523:47-52.
- Nishizawa K. 2009. Natively unfolded proteins: An overview. *Biophysics* 4:53-58. Review.
- Nothias JY, Majumder S, Kaneko KJ, DePamphilis ML. 1995. Regulation of gene expression at the beginning of mammalian development. *J Biol Chem* 270:22077-80.
- Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. 2003. Predicting intrinsic disorder from amino acid sequence. *Proteins* 53 Suppl 6:566-72

- Ohnacker M, Barabino SM, Preker PJ, and Keller W. 2000. The WD- repeat protein pfs2p bridges two essential factors within the yeast pre- mRNA 30-end-processing complex. *EMBO J* 19:37-47.
- Olsen GJ, Woese CR. 1993. Ribosomal RNA: a key to phylogeny. *FASEB J* 7:113-23.
- Oshikane H, Sheppard K, Fukai S, Nakamura Y, Ishitani R, Numata T, Sherrer RL, Feng L, Schmitt E, Panvert M, Blanquet S, Mechulam Y, Söll D, Nureki O. 2006. Structural basis of RNA-dependent recruitment of glutamine to the genetic code. *Science* 312:1950-1954.
- Papin C, Rouget C, Mandart E. 2008. Xenopus Rbm9 is a novel interactor of XGld2 in the cytoplasmic polyadenylation complex. *FEBS J* 275:490-503.
- Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* 14:609-14.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* 96:4285-8.
- Pierce B, Phillips AT, Weng Z. 2007. Structure Prediction of Protein Complexes. In Xu Y, Xu D, Liang J. (Eds.), *Computational Methods for Protein Structure Prediction and Modeling Volume 2: Structure Prediction*. (pp. 109-134). New York: Springer.
- Pintar A, Carugo O, Pongor S. 2002. CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics* 18:980-4.
- Piqué M, López JM, Foissac S, Guigó R, Méndez R. 2008. A combinatorial code for CPE-mediated translational control. *Cell* 132:434-48.
- Piñol-Roma S, Dreyfuss G. 1992. Shuttling of pre-mRNA binding proteins between nucleus and cytoplasm. *Nature* 355:730-2.
- Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 21:3435-8
- Proudfoot NJ, Furger A, Dye MJ. 2002. Integrating mRNA processing with transcription. *Cell* 108:501-12. Review.
- Quistgaard EM, Madsen P, Grøftehaug MK, Nissen P, Petersen CM, Thirup SS. 2009. Ligands bind to Sortilin in the tunnel of a ten-bladed beta-propeller domain. *Nat Struct Mol Biol* 16:96-8.
- Radford HE, Meijer HA, de Moor CH. 2008. Translational control by cytoplasmic polyadenylation in *Xenopus* oocytes. *Biochim Biophys Acta* 1779:217-29.
- Ramakrishnan V. 2002. Ribosome structure and the mechanism of translation. *Cell* 108:557-72.
- Rao VS, Srinivas K, Sujini GN, Kumar GN. 2014. Protein-protein interaction detection: methods and analysis. *Int J Proteomics* 2014:147648.

- Rice S, Lin AW, Safer D, Hart CL, Naber N, Carragher BO, Cain SM, Pechatnikova E, Wilson-Kubalek EM, Whittaker M, Pate E, Cooke R, Taylor EW, Milligan RA, Vale RD. 1999. A structural change in the kinesin motor protein that drives motility. *Nature* 402:778-84.
- Rouhana L, Wang L, Buter N, Kwak JE, Schiltz CA, Gonzalez T, Kelley AE, Landry CF, Wickens M. 2005. Vertebrate GLD2 poly(A) polymerases in the germline and the brain. *RNA* 11:1117-30
- Ryan K, Calvo O, Manley JL. 2004. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. *RNA* 10:565-73.
- Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437:1173-8.
- Rüegsegger U, Beyer K, Keller W. 1996. Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *J Biol Chem* 271:6107-13.
- Sabath I, Skrajna A, Yang XC, Dadlez M, Marzluff WF, Dominski Z. 2013 3'-End processing of histone pre-mRNAs in *Drosophila*: U7 snRNP is associated with FLASH and polyadenylation factors. *RNA* 19:1726-44.
- Sabini E, Hazra S, Konrad M, Lavie A. 2007. Nonenantioselectivity property of human deoxycytidine kinase explained by structures of the enzyme in complex with L- and D-nucleosides. *J Med Chem* 50:3004-3014.
- Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406-25.
- Sali A, Overington JP. 1994. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci* 3:1582-96.
- Schneider TD, Stephens RM. 1990. Sequence logos: A new way to display consensus sequences. *Nucl. Acids Res* 18:6097-6100.
- Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. 2005. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 33(Web Server issue):W363-7.
- Schwarz G, Schrader N, Mendel RR, Hecht H, Schindelin H. 2001. Crystal structures of human gephyrin and plant cnx1 G domain: Comparative analysis and functional implications. *J Mol Biol* 312:405-418.
- Sadowski MI, Jones DT. 2007. Benchmarking template selection and model quality assessment for high-resolution comparative modeling. *Proteins* 69:476-485.
- Setoyama D, Yamashita M, Sagata N. 2007. Mechanism of degradation of CPEB during *Xenopus* oocyte maturation. *Proc Natl Acad Sci USA* 104:18001-18006.

- Shen Y, Joachimiak A, Rosner MR, Tang WJ. 2006. Structures of human insulin-degrading enzyme reveal a new substrate recognition mechanism. *Nature* 443:870-874.
- Sherry ST, Ward M, Sirotkin K. 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* 9:677-679.
- Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR 3rd, Frank J, Manley JL. 2009. Molecular architecture of the human pre-mRNA 3' processing complex. *Mol Cell* 33:365-376.
- Shi Y, Fernandez-Martinez J, Tjioe E, Pellarin R, Kim SJ, Williams R, Schneidman D, Sali A, Rout MP, Chait BT. 2014. Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol Cell Proteomics* 13:2927-2943
- Shindyalov IN, Bourne PE. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11:739-747.
- Shionyu M, Takahashi K, Go M. 2001. Variable subunit contact and cooperativity of hemoglobin. *J Mol Evol* 53:416-429.
- Shrake A, Rupley JA. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol* 79:351-364.
- Silva AP, Chechik M, Byrne RT, Waterman DG, Ng CL, Dodson EJ, Koonin EV, Antson AA, Smits C. 2011. Structure and activity of a novel archaeal β -CASP protein with N-terminal KH domains. *Structure* 19:622-632.
- Sola M, Bavro, VN, Timmins J, Franz T, Richard-Blum S, Schoehn G, Ruigrok RWH, Paarmann I, Saiyed T, O'Sullivan GA, Schmitt B, Betz H, Weissenhorn W. 2004. Structural basis of dynamic glycine receptor clustering by gephyrin. *EMBO J* 22:250-259.
- Stebbins-Boaz B, Cao Q, de Moor CH, Mendez R, Richter JD. 1999. Maskin is a CPEB-associated factor that transiently interacts with eIF-4E. *Mol Cell* 4:1017-1027.
- Stenberg F, Chovanec P, Maslen SL, Robinson CV, Ilag LL, von Heijne G, Daley DO. 2005. Protein complexes of the Escherichia coli cell envelope. *J Biol Chem* 280: 34409-34419.
- Stirnemann CU1, Petsalaki E, Russell RB, Müller CW. 2010. WD40 proteins propel cellular networks. *Trends Biochem Sci* 35:565-574.
- Sullivan KD, Steiniger M, Marzluff WF. 2009. A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. *Mol Cell* 34:322-332.
- Summa CM, Levitt M. 2007. Near-native structure refinement using in vacuo energy minimization. *Proc Natl Acad Sci USA* 104:3177-3182.
- Takahashi Y, Helmling S, Moore CL. 2003. Functional dissection of the zinc finger and flanking domains of the Yth1 cleavage/polyadenylation factor. *Nucleic Acids Res* 31:1744-1752.
- Taoka S, Lepore BW, Kabil O, Ojha S, Ringe D, Banerjee R. 2002. Human cystathionine beta-synthase is a heme sensor protein. Evidence that the redox sensor is heme and not the vicinal

cysteines in the CXXC motif seen in the crystal structure of the truncated enzyme. *Biochemistry* 41:10454-10461.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631-637.

The UniProt Consortium. 2011. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* 39:D214-D219.

Tomii K, Akiyama Y. 2004. FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics* 20:594-595.

Tompa P. 2005. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* 579:3346-3354.

Tovchigrechko A, Vakser IA. 2006. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res* 34:W310-4.

Tsuchiya Y, Kinoshita K, Nakamura H. 2006. Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity. *Protein Eng Des Sel* 19:421-9.

Tumbula DL, Becker HD, Chang W, Söll D. 2004. Domain-specific recruitment of amide amino acids for protein synthesis. *Nature* 407:106-110.

Uversky VN. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 11:739-756.

Valdar WS, Thornton JM. 2001. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* 4:108-124.

Voeltz GK, Ongkasuwan J, Standart N, Steitz JA. 2001. A novel embryonic poly(A) binding protein, ePAB, regulates mRNA deadenylation in *Xenopus* egg extracts. *Genes Dev* 15:774-788.

de Vries SJ, van Dijk M, Bonvin AM. 2010. The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc* 5:883-897.

Vullo A, Bortolami O, Pollastri G, Tosatto S. 2006. Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res* 34:W164-W168

Walsh I, Martin AJ, Di Domenico T, Tosatto SC. 2012. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* 28:503-509.

Wang L, Eckmann CR, Kadyk LC, Wickens M, Kimble J. 2002. A regulatory cytoplasmic poly(A) polymerase in *Caenorhabditis elegans*. *Nature* 419:312-316.

Wang L, Xu K, Lin J, Shao S, Zhang T, Xu H, Wei Z, Zhang Z. 2013. A novel genetic system based on zinc finger nucleases for the identification of interactions between proteins in vivo. *PLoS One* 8:e85650.

- Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337:635-645.
- Webb B, Sali A. 2014. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* 47:5.6.1-5.6.32.
- Webb BA, Forouhar F, Szu FE, Seetharaman J, Tong L, Barber DL. 2015. Structures of human phosphofructokinase-1 and atomic basis of cancer-associated mutations. *Nature* 523:111-114.
- Weill L, Belloc E, Bava FA, Méndez R. 2012. Translational control by changes in poly(A) tail length: recycling mRNAs. *Nat Struct Mol Biol* 19:577-585.
- Wiederstein M, Sippl MJ. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* 35:W407-W410.
- Wilt FH. 1973. Polyadenylation of maternal RNA of sea urchin eggs after fertilization. *Proc Natl Acad Sci USA* 70:2345-2349.
- Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293:321-331.
- Xia Z, Dai W, Zhang Y, White SA, Boyd GD, Mathews FS. 1996. Determination of the gene sequence and the three-dimensional structure at 2.4 angstroms resolution of methanol dehydrogenase from *Methylophilus W3A1*. *J Mol Biol* 259:480-501.
- Xiang K, Tong L, Manley JL. 2014. Delineating the structural blueprint of the pre-mRNA 3' end processing machinery. *Mol Cell Biol* 34:1894-1910
- Xiang S, Nichols J, Rajagopalan KV, Schindelin H. 2001. The crystal structure of *Escherichia coli* MoeA and its relationship to the multifunctional protein gephyrin. *Structure* 9: 299-310.
- Xu R, Ye X, Quinn Li Q. 2004. AtCPSF73-II gene encoding an Arabidopsis homolog of CPSF 73 kDa subunit is critical for early embryo development. *Gene* 324:35-45.
- Xu R, Zhao H, Dinkins RD, Cheng X, Carberry G, Li QQ. 2006. The 73 kD subunit of the cleavage and polyadenylation specificity factor (CPSF) complex affects reproductive development in Arabidopsis. *Plant Mol Biol* 61:799-815.
- Yang Q, Doublé S. 2011. Structural biology of poly(A) site definition. *Wiley Interdiscip Rev RNA* 2:732-747.
- Yang XC, Sabath I, Dębski J, Kaus-Drobek M, Dadlez M, Marzluff WF, Dominski Z. 2013. A complex containing the CPSF73 endonuclease and other polyadenylation factors associates with U7 snRNP and is recruited to histone pre-mRNA for 3'-end processing. *Mol Cell Biol* 33:28-37.
- Yang ZR, Thomson R, McNeil P, Esnouf RM. 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21:3369-3376
- Yankovskaya V, Horsefield R, Tornroth S, Luna-Chavez C, Miyoshi H, Leger C, Byrne B, Cecchini G, Iwata S. 2003. Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science* 299:700-704.

- Yun MK, Nourse A, White SW, Rock CO, Heath RJ. 2007. Crystal structure and allosteric regulation of the cytoplasmic *Escherichia coli* L-asparaginase I. *J Mol Biol* 369:794-811.
- Yura K, Hayward S. 2009. The interwinding nature of protein-protein interfaces and its implication for protein complex formation. *Bioinformatics* 25:3108-3113.
- Yura K, Yamaguchi A, Go M. 2006. Coverage of whole proteome by structural genomics observed through protein homology modeling database. *J Struct Funct Genomics* 7:65-76.
- 由良敬, 高井美智留. 2013. 進化分子工学: 高速分子進化によるタンパク質・核酸の開発. NTS出版. 61-70
- Zhao J, Hyman L, Moore C. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* 63:405-445. Review.
- Zhang Y, Stec B, Godzik A. 2007. Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. *Structure* 15:1141-1147.
- Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M, Snyder M. 2001. Global analysis of protein activities using proteome chips. *Science* 293:2101-2105.