

# 統合データベース講習会：AJACS 蝦夷 3

## 「既存データベースを活用したタンパク質実験・構造データの探し方」

西方 公郎 理化学研究所 情報基盤センター 統合データベース特別ユニット・センター研究員

2013 年 11 月 6 日 (水) 16:50~17:50 北海道大学工学部 材料化学棟 1 階 計算機室 (MC105)

私たち生命の細胞は、DNA、RNA、タンパク質やアミノ酸、水やイオン、脂質などの生体分子からできています。これらの生体分子は、遺伝子の情報を基にして作られていて、それぞれが形(構造)を持っています。

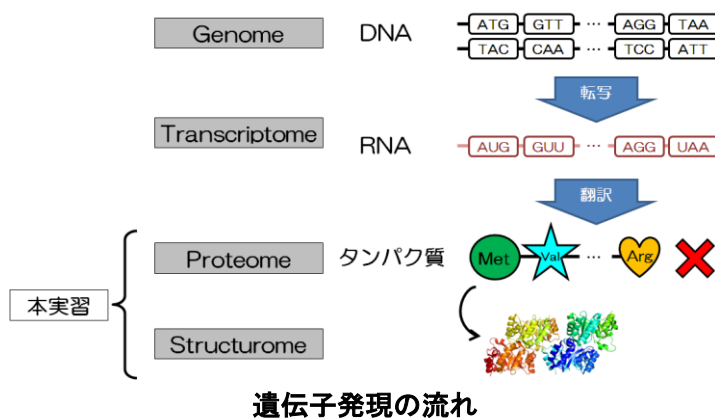
今世紀初めにヒトゲノムが解読されて以降、タンパク 3000 等の構造ゲノミクスプロジェクトによって、タンパク質を中心とした生体分子の立体構造が解析され、Protein Data Bank (PDB)等のデータベースに登録されてきました。これらの構造解析には X 線回折や核磁気共鳴 (NMR) 等の実験手法が用いられており、それらの実験データも豊富に存在します。

本実習では、PDB、UniProt、Pfam 等をはじめとした公共データベースに加えて、構造生物学実験のデータが登録されている理化学研究所のデータベースから、バイオ分野の若手研究者の皆さん(大学生・大学院生・PD)が求めるタンパク質データを探し出す方法を紹介したいと思います。

### 【はじめに】

右図は、高校の「生物」でも学習する、遺伝子の発現の流れを表しています。細胞内で、遺伝情報は主に、DNA→RNA→タンパク質という方向で合成が進みます。DNA のもつ遺伝情報が RNA に変換されることを転写、RNA からタンパク質がつけられることを翻訳と呼びます。

本実習では、様々なデータベースから、タンパク質の配列、機能、立体構造、実験の情報を取り出します。また、取得した立体構造をコンピュータを使って観察します。



### 【使用する道具】

- ・ パソコン

### 【使用する環境とソフトウェア】

- ・ インターネットに接続できる環境
- ・ 分子グラフィックスソフトウェア
  - RasMol (<http://rasmol.org/>)
  - PyMol (<http://www.pymol.org/>)

## 【本実習の流れ】

本日の実習では、以下の内容を行いたいと思います。

- [1. タンパク質関連のデータベースを知る] … 20 分
- [2. コンピュータでタンパク質を観察する] … 20 分
- [3. タンパク質立体構造解析の実験データを調べる] … 20 分

## 【実習手順】

### [1. タンパク質関連のデータベースを知る]

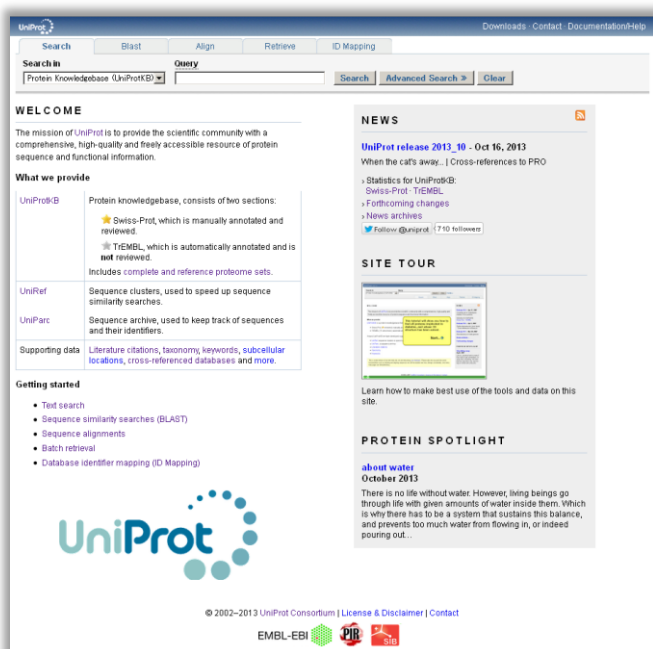
タンパク質関連のデータベースは様々なものが知られていますが、ここでは代表的な以下のものを紹介します。

分類	データベース名	URL
配列データベース	UniProt	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>
タンパク質ドメインデータベース	Pfam SMART InterPro	<a href="http://pfam.sanger.ac.uk/">http://pfam.sanger.ac.uk/</a> <a href="http://smart.embl-heidelberg.de/">http://smart.embl-heidelberg.de/</a> <a href="http://www.ebi.ac.uk/interpro/">http://www.ebi.ac.uk/interpro/</a>
立体構造データベース	Protein Data Bank	<a href="http://pdbj.org/">http://pdbj.org/</a> <a href="http://www.rcsb.org/">http://www.rcsb.org/</a> <a href="http://www.ebi.ac.uk/pdbe/">http://www.ebi.ac.uk/pdbe/</a>
タンパク質実験データベース	IDPES (RIKEN Integrated Databases of Protein Structures and Experiments)	<a href="http://database.riken.jp/db/protein">http://database.riken.jp/db/protein</a>

## 配列データベース

UniProt (<http://www.uniprot.org/>)

UniProt (Universal Protein Resource) は有名なタンパク質配列データベースの一つです。タンパク質の配列情報がアノテーション (注釈付け) 情報とともに収録されています。



例えば、「TFIIB」で検索すると、以下の結果が表示されます。

UniProtKB Search results for TFIIB. The search query 'TFIIB' has returned 4,134 results. The results table shows various entries from different organisms, including Drosophila, Homo sapiens, and Arabidopsis. The entry Q00403 (TF2B\_HUMAN) is highlighted.

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
Q9NHP7	TF2B_DROVI	★	Transcription initiation factor IIB	TFIIB	Drosophila virilis (Fruit fly)	293
Q00403	TF2B_HUMAN	★	Transcription initiation factor IIB	GTF2B TF2B TFIIB	Homo sapiens (Human)	316
P29052	TF2B_DROME	★	Transcription initiation factor IIB	TFIIB CG5193	Drosophila melanogaster (Fruit fly)	315
P48513	TF2B_SOYBN	★	Transcription initiation factor IIB	TFIIB1	Glycine max (Soybean) (Glycine hispida)	313
Q54FD6	TF2B_DICDI	★	Transcription initiation factor IIB	gtf2b tfib DDB_G0290929	Dictyostelium discoideum (Slime mold)	325
P48512	TF2B1_ARATH	★	Transcription initiation factor IIB-1	TFIIB1 At2g41630 T3G6.15	Arabidopsis thaliana (Mouse-ear cress)	312
Q9SS44	TF2B2_ARATH	★	Transcription initiation factor IIB-2	TFIIB2 At3g10330 F14P13.7	Arabidopsis thaliana (Mouse-ear cress)	312
Q8W0V3	TF2B_ORYSJ	★	Transcription initiation factor IIB	TFIIB Os09g0534800 LOC_Os09g36440 QJ1112_E07.31 P0569E11.1	Oryza sativa subsp. japonica (Rice)	312
B4LVD1	B4LVD1_DROVI	★	TFIIB	TFIIB DvirTFIIB Dvir_GJ13764 GJ13764	Drosophila virilis (Fruit fly)	315
Q9AVY2	Q9AVY2_GUITH	★	TFIIB related factor hBRF	tfib-brf	Guillardia theta (Cryptomonas phi)	394

<http://www.uniprot.org/uniprot/?query=TFIIB&sort=score>

その内の一つ「TF2B\_HUMAN」を選ぶと、以下のページが表示されます。

ここには、ヒト由来の基本転写因子 TFIIB についての様々な情報(配列長、機能、構造情報、オントロロジーアノテーション、配列の特徴、参考文献)がアミノ酸配列と共に書かれています。

UniProtKB entry page for Q00403 (TF2B\_HUMAN). The page displays detailed information about the protein, including its name, origin, and function. The 'Names and origin' section shows the protein name 'Transcription initiation factor IIB' and the gene name 'GTF2B'. The 'Protein attributes' section shows the sequence length as 316 AA. The 'General annotation (Comments)' section provides a detailed description of the protein's function and its role in the transcription process.

**Names and origin**

Protein names	Recommended name: Transcription initiation factor IIB Alternative name(s): General transcription factor TFIIB S300-II
Gene names	Name: GTF2B Synonyms: TF2B, TFIIB
Organism	Homo sapiens (Human) [Reference proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorhini › Catarrhini › Hominoidea › Homo

**Protein attributes**

Sequence length	316 AA.
Sequence status	Complete.
Protein existence	Evidence at protein level

**General annotation (Comments)**

Function	General factor that plays a major role in the activation of eukaryotic genes transcribed by RNA polymerase II.
Cofactor	Binds 1 zinc ion per subunit.
Subunit structure	Associates with TFIID-III (DA complex) to form TFIID-III-IB (DAB-complex) which is then recognized by polymerase II. Interacts with the transcription elongation factor TCEA2. Interacts with HIV-1 Vpr. Interacts with GPBP1 (By similarity). Interacts with Epstein-Barr virus EBNA2. [Ref.5] [Ref.9] [Ref.10]
Subcellular location	Nucleus.
Sequence similarities	Belongs to the TFIIB family. Contains 1 TFIIB-type zinc finger.

**Ontologies**

<http://www.uniprot.org/uniprot/Q00403>

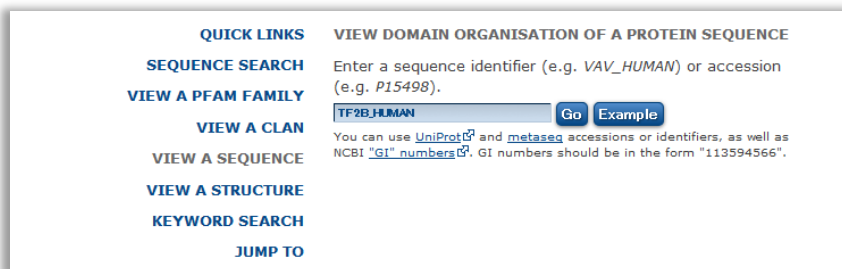
## タンパク質ドメインデータベース

Pfam (<http://pfam.sanger.ac.uk/>)

Pfam(Protein families database of alignments and HMMs)とはタンパク質ドメインのデータベースです。HMM(Hidden Markov Model)という確率モデルを使って、タンパク質のファミリーを収集し、整理したものが収録されています。

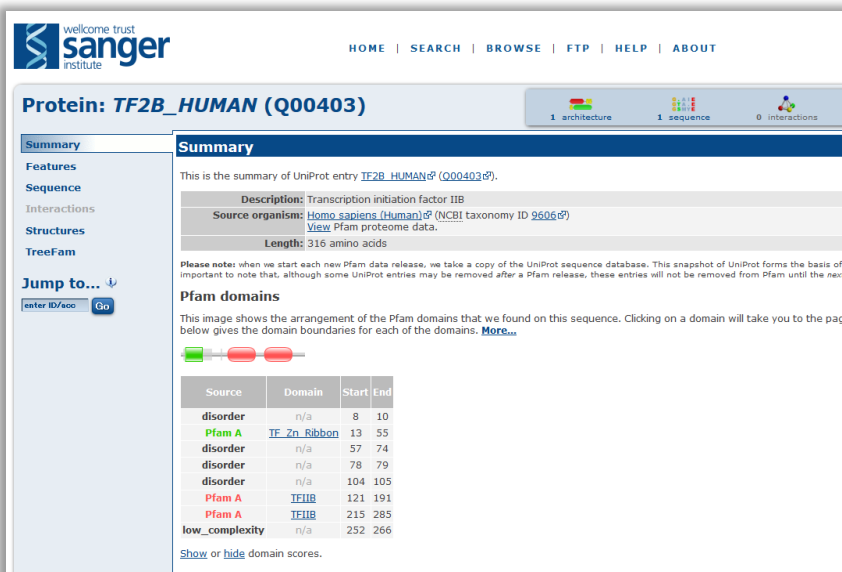


例えば、UniProt のタンパク質配列の ID 「TF2B\_HUMAN」を入力すると、



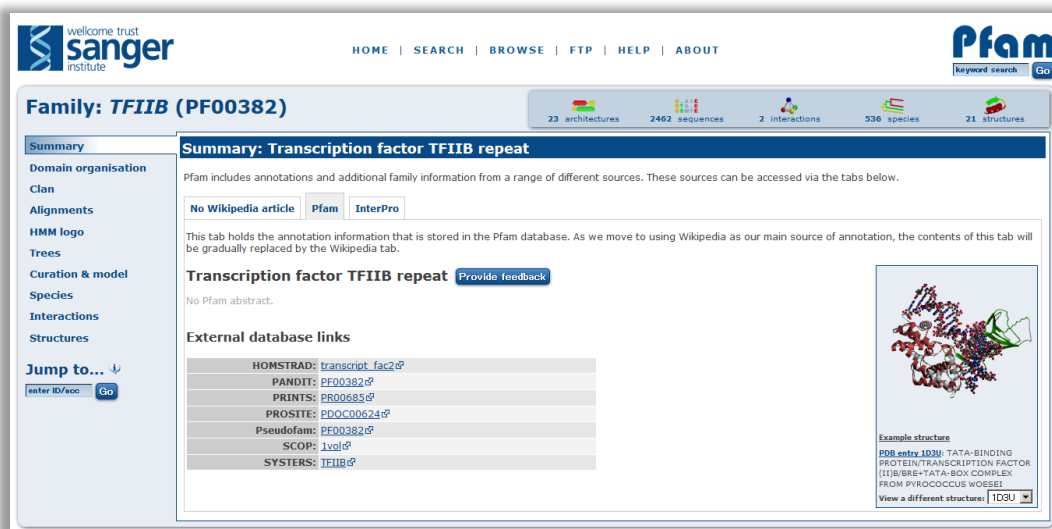
以下のページが表示されます。

このタンパク質配列が TF\_Zn\_Ribbon ドメインと TFIIB ドメインから構成されることが分かります。



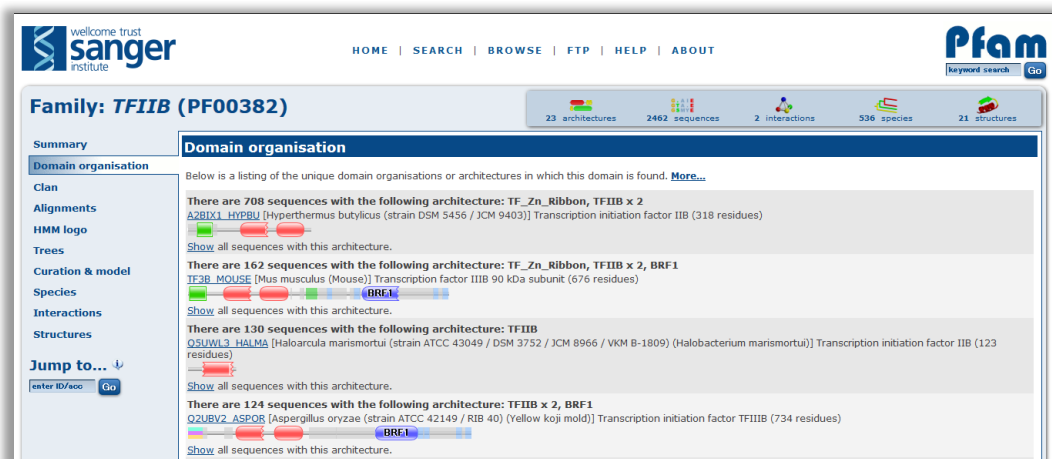
[http://pfam.sanger.ac.uk/protein/TF2B\\_HUMAN](http://pfam.sanger.ac.uk/protein/TF2B_HUMAN)

さらに、「TFIIB」をクリックすると、以下のページが表示され、このドメインについての様々な情報を見ることができます。



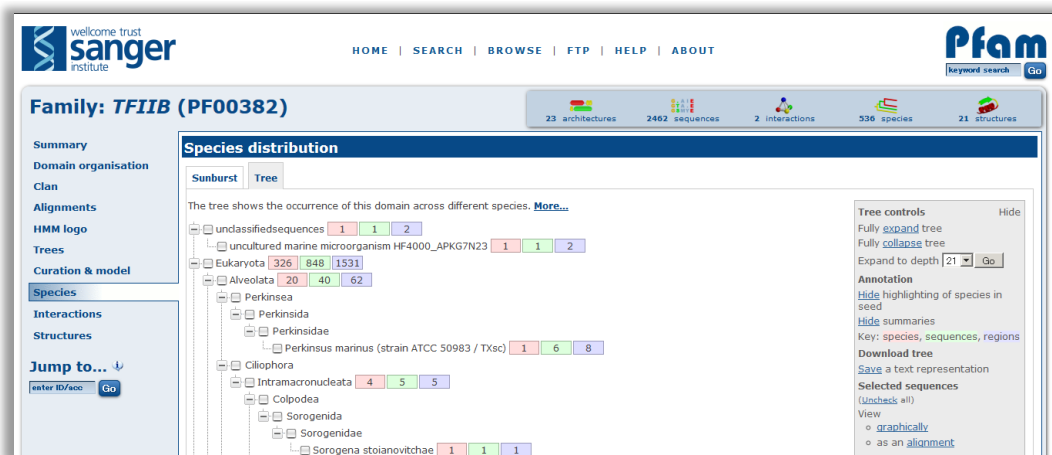
<http://pfam.sanger.ac.uk/family/TFIIB>

Domain Organization ... TFIIB ドメインを含むタンパク質の一覧



<http://pfam.sanger.ac.uk/family/TFIIB#tabview=tab1>

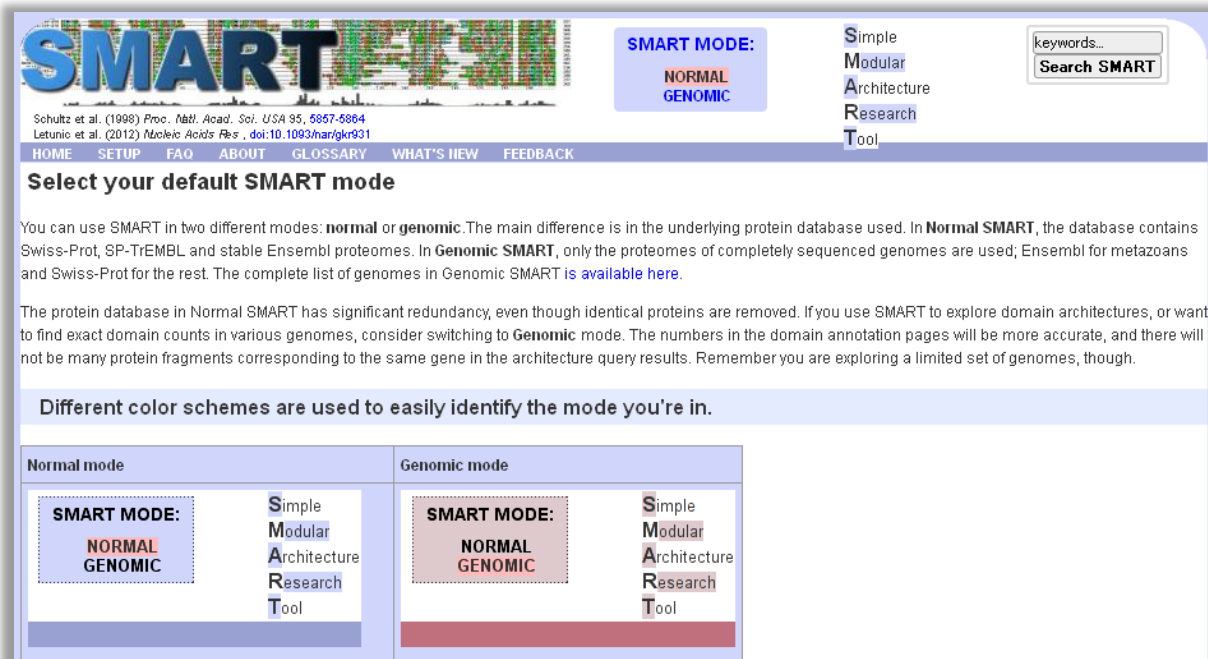
Species ... 各生物種で TFIIB というファミリーが見つかった数が表示されます。



<http://pfam.sanger.ac.uk/family/TFIIB#tabview=tab7>

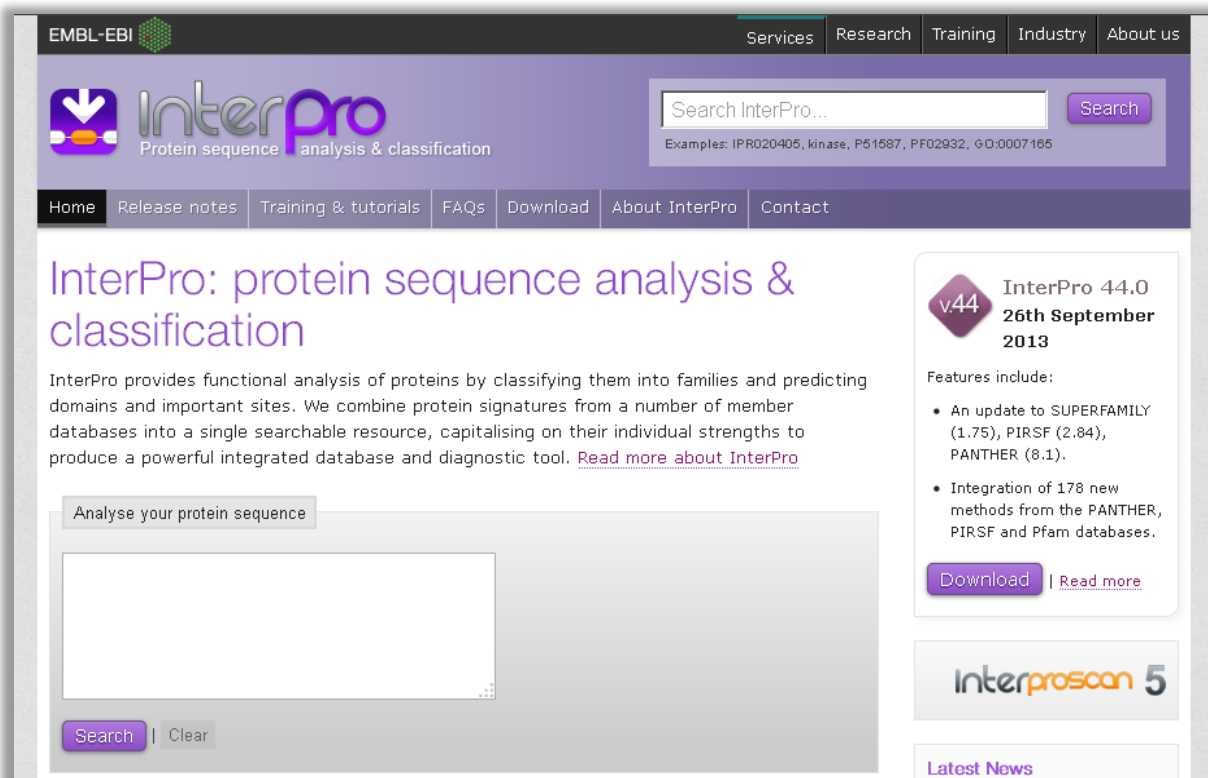
SMART (<http://smart.embl-heidelberg.de/>)

類似のデータベースは他にも様々なものが知られています。SMART(Simple Modular Architecture Research Tool)は、細胞外ドメインやシグナル伝達ドメインに特化したデータベースです。



InterPro (<http://www.ebi.ac.uk/interpro/>)

InterPro(Integrated documentation resource for Protein families, domains and functional sites)はEBI(European Bioinformatics Institute)から提供されている蛋白質ファミリー、ドメイン、機能部位のデータベースです。



UniProt の TF2B\_HUMAN の FASTA 形式の配列 (<http://www.uniprot.org/uniprot/Q00403.fasta> から取得できます) をテキストボックスにペーストして、Search ボタンをクリックすると、以下の結果が表示されます。

The screenshot shows the InterPro website interface. At the top, there's a navigation bar with links like Home, Release notes, Training & tutorials, FAQs, Download, About InterPro, and Contact. A search bar is also present. The main content area displays the protein details for Transcription initiation factor IIB (Q00403). It includes a sidebar with filters for Entry type (Family, Domains, Repeats, Site) and Status (Unintegrated). The main section shows the protein's accession (Q00403), species (Homo sapiens), and length (316 amino acids). Below this, there's a section for Protein family membership, showing the Transcription factor TFIIB (IPR000812). The Domains and repeats section displays a domain architecture diagram. The Detailed signature matches section lists various domain signatures and their matches, including Transcription factor TFIIB, Zinc finger, TFIIB-type, Cyclin-like, Transcription factor TFIIB, cyclin-like domain, and Transcription factor TFIIB, conserved site. The GO term prediction section lists Biological Process, Molecular Function, and Cellular Component terms.

InterPro は、上記で述べた Pfam, SMART に加えて、PRINTS (<http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>), PROSITE (<http://prosite.expasy.org/>), PRODOM (<http://prodom.prabi.fr/>) というタンパク質ドメインデータベースを統合しています。1回の問い合わせで、これらのデータベースを検索し、それらをまとめた結果を返してくれます。



## 立体構造データベース

### Protein Data Bank (PDB)

Protein Data Bank (PDB) は、タンパク質や核酸 (DNA, RNA)、または、それらの複合体の立体構造が登録されたデータベースです。現在、9 万件以上の立体構造データが登録され、日本、米国、欧州、それぞれでサイトが公開されており、いずれのサイトからも同じデータを取得することができます。

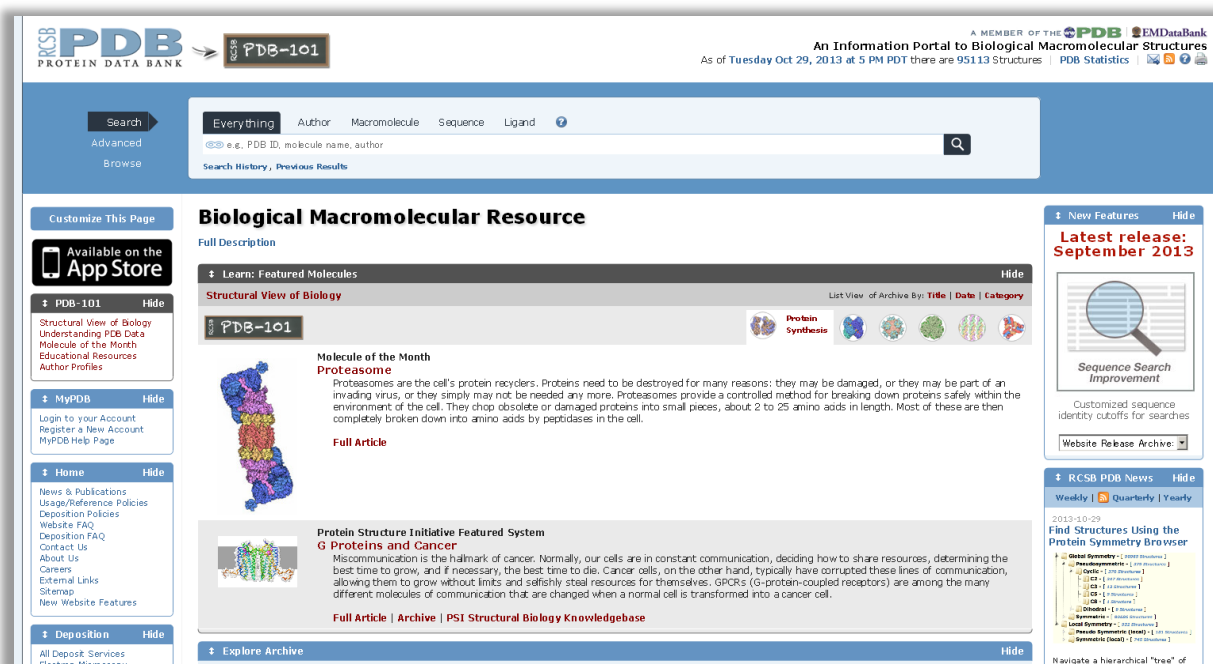
### PDBj (Protein Data Bank Japan) (<http://pdbj.org/>)

大阪大学の蛋白質研究所で運営されているタンパク質の立体構造の情報が登録されたデータベースです。タンパク質の立体構造は、X 線や NMR という装置を使った実験で決定されます。



### RCSB PDB (RCSB Protein Data Bank) (<http://www.rcsb.org/>)

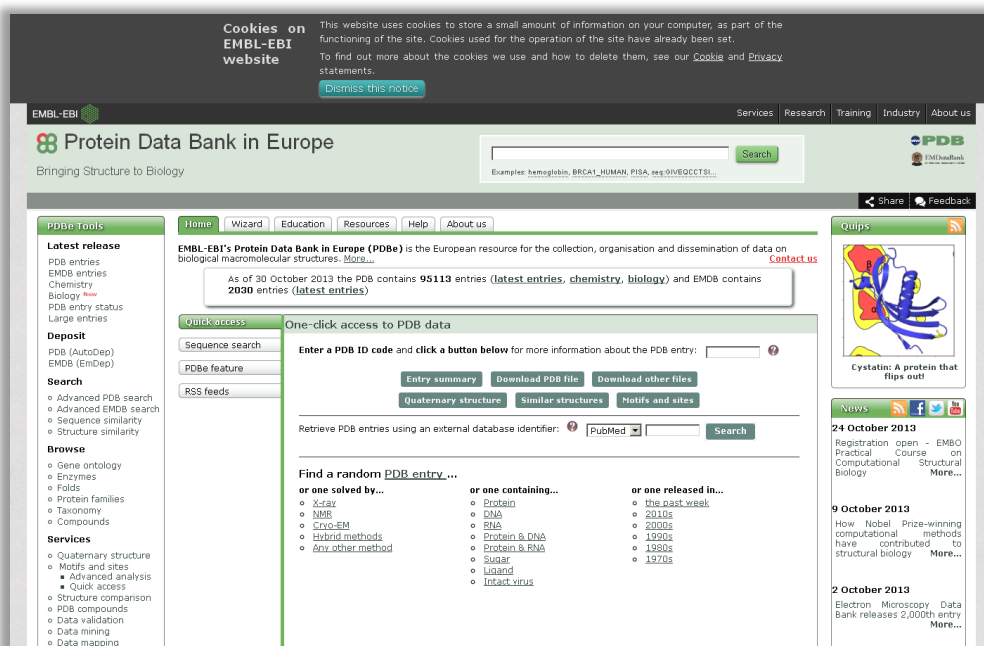
米国にある RCSB (Research Collaboratory for Structural Bioinformatics) によって運営されているデータベースです。





PDBe (Protein Data Bank in Europe) (<http://www.ebi.ac.uk/pdbe/>)

欧州の EBI (European Bioinformatics Institute) によって運営されているデータベースです。



## [2. コンピュータでタンパク質を観察する] (補足資料参照)

1. タンパク質の形の情報が登録されているデータベースにアクセスする。

例. PDBj (<http://pd bj. org/>)

2. 観察したいタンパク質を調べる。
3. タンパク質のファイル(PDB ファイル)をダウンロードする。
4. ソフトウェアを起動する。
5. 3. でダウンロードした PDB ファイルを読み込む。
6. タンパク質の形が表示されます！ マウスを使って動かしたり、色々な場所を観察しましょう。

## [3. タンパク質立体構造解析の実験データを調べる] (スライド参照)

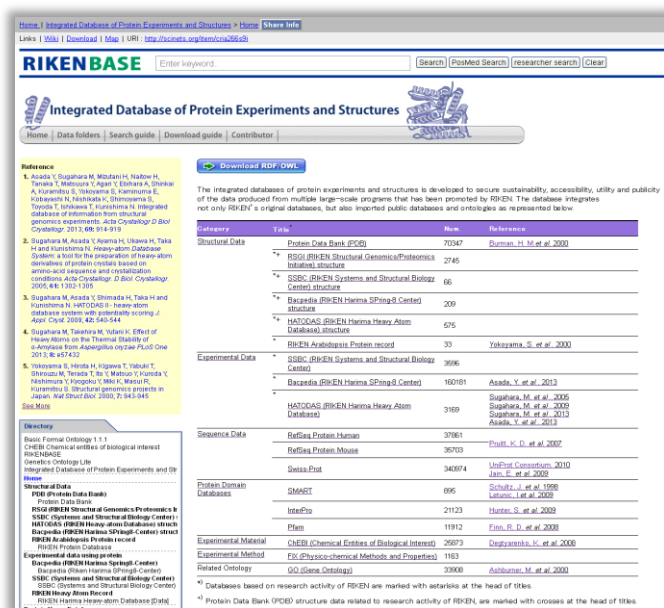
タンパク質の立体構造は、X 線回折や NMR という構造生物学実験によって解析されています。PDB のようなデータベースには、立体構造情報は登録されていますが、それに付随する様々な実験情報は必ずしも登録されていません。また、立体構造の決定に至らなかった大量のタンパク質実験の情報も存在しますが、通常は公開されていません。

理化学研究所では、構造生物学実験で生み出されたタンパク質実験の情報を PDB 等の公共データと共に統合したデータベースが公開されています(下記 URL 参照)。このデータベースの使い方を紹介します。

## 理研タンパク質実験・構造統合データベース

IDPES (RIKEN Integrated Databases of Protein Experiments and Structures)

URL: <http://database.riken.jp/db/protein>



このデータベース IDPES は、タンパク 3000 プロジェクトで RIKEN Structural Genomics/Proteomics Initiative (RSGI)によって構造解析された、ヒト、マウス、シロイヌナズナ、高度好熱菌由来のタンパク質の実験データを、PDB の構造データ等の公共データベースとともに統合したデータベースです。

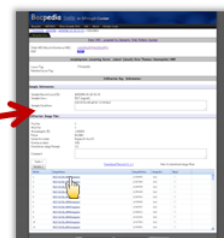
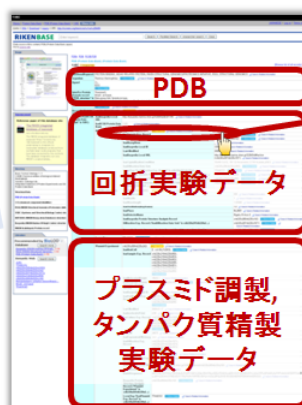
膨大な構造生物学関連データを統合するため、Semantic Web 技術を用いたプラットフォーム RIKENBASE (<http://database.riken.jp/>, Masuya H et al, 2010) 上に IDPES を構築しました。コンテンツはオントロジーの階層構造に基づいて、配列データ、タンパク質ドメインデータ、構造データ等に分類されています。ユーザは任意のキーワードで目的のデータを取得できます。

IDPES からは、理研播磨研究所の SPring8 や理研横浜研究所の SSBC (Systems and Structural Biology Center) から提供された X 線回折イメージや結晶観察写真等の実験データを取得することができます。これらのデータは、構造生物学研究やタンパク質デザインに有用であると期待されます。



PDBオリジナルサイト  
RIKENオリジナルの  
実験データを  
取得できない

統合化後のPDB  
(on RIKENBASE)  
PDBの構造データとともに  
実験データを閲覧できる



<http://bacpedia.harima.riken.jp/bacteria/View/Frm/Contents/topview.aspx>

Bacpediaオリジナルサイト  
RIKENオリジナルの実験  
データを取得できる



<http://database.riken.jp/db/ssbc>

### SSBCデータベース

高等動物(ヒト, マウス)由来のタンパク質を用いて行われた構造生物学実験について、結晶化プレート写真や回折実験データを取得できる

IDPES には、以下のデータベースが統合されています。詳細は参考文献などもご覧下さい。

分類	データベース名	RIKENBASE URL	参考文献
構造データ	Protein Data Bank	<a href="http://database.riken.jp/sw/item/riali">http://database.riken.jp/sw/item/riali</a>	Burman, H. M. <i>et al.</i> 2000
	RIKEN Arabidopsis Protein record*	<a href="http://database.riken.jp/sw/item/rib46i">http://database.riken.jp/sw/item/rib46i</a>	Yokoyama, S. <i>et al.</i> 2000
実験データ	SSBC (RIKEN Systems and Structural Biology Center)*	<a href="http://database.riken.jp/db/ssbc">http://database.riken.jp/db/ssbc</a>	
	Bacpedia (RIKEN Harima SPring-8 Center)*	<a href="http://database.riken.jp/db/bacpedia">http://database.riken.jp/db/bacpedia</a>	Asada, Y. <i>et al.</i> 2013
	HATODAS (RIKEN Harima Heavy Atom Database)*	<a href="http://database.riken.jp/db/hatodas">http://database.riken.jp/db/hatodas</a>	Sugahara, M. <i>et al.</i> 2005 Sugahara, M. <i>et al.</i> 2009 Sugahara, M. <i>et al.</i> 2013 Asada, Y. <i>et al.</i> 2013
配列データ	RefSeq Protein Human	<a href="https://database.riken.jp/item/ria141i">https://database.riken.jp/item/ria141i</a>	Pruitt, K. D. <i>et al.</i> 2007
	RefSeq Protein Human	<a href="https://database.riken.jp/item/ria118i">https://database.riken.jp/item/ria118i</a>	
	Swiss-Prot	<a href="https://database.riken.jp/item/ria214i">https://database.riken.jp/item/ria214i</a>	UniProt Consortium. 2010 Jain, E. <i>et al.</i> 2009
タンパク質 ドメイン データ	Pfam	<a href="http://database.riken.jp/item/rib125i">http://database.riken.jp/item/rib125i</a>	Finn, R. D. <i>et al.</i> 2008
	SMART	<a href="http://database.riken.jp/item/ria95i">http://database.riken.jp/item/ria95i</a>	Schultz, J. <i>et al.</i> 1998 Letunic, I. <i>et al.</i> 2009
	InterPro	<a href="http://database.riken.jp/item/rib1241i">http://database.riken.jp/item/rib1241i</a>	Hunter, S. <i>et al.</i> 2009
実験材料 オントロジー	ChEBI (Chemical Entities of Biological Interest)	<a href="http://database.riken.jp/item/ria244i">http://database.riken.jp/item/ria244i</a>	Degtyarenko, K. <i>et al.</i> 2008
実験方法 オントロジー	FIX (Physico-chemical Methods and Properties)	<a href="http://database.riken.jp/item/ria265i">http://database.riken.jp/item/ria265i</a>	

\*理研オリジナルのデータベース

## 【謝辞】

### 理化学研究所 情報基盤センター 統合データベース特別ユニット

豊田 哲郎 ユニットリーダー	吉田 有子
小林 紀郎	望月 芳樹
蒔田 由布子	松嶋 明宏
土井 考爾	石井 学
下山 紗代子	高橋 聡史

### 理化学研究所 バイオリソースセンター

榎屋 啓志 ユニットリーダー

### 理化学研究所 放射光科学研究センター

国島 直樹 チームリーダー

### 理化学研究所 生命システム基盤研究領域

横山 茂之 領域長  
明 恒次郎

(敬称略)

RIKENBASE is supported by RIKEN and the National Bioscience Database Centre (NBDC) of the Japan Science and Technology Agency (JST).

## 【参考図書】

### [初心者向け]

- ・タンパク質の構造入門 第2版 (<http://www.amazon.co.jp/dp/4315515604>)
- ・実践バイオインフォマティクス (<http://www.amazon.co.jp/dp/4873110688/>)
- ・バイオインフォマティクスのためのPerl入門 (<http://www.amazon.co.jp/dp/487311103X/>)

### [より学びたい人向け]

- ・Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis  
(<http://www.amazon.co.jp/dp/0470848391>)
- ・バイオインフォマティクス ゲノム配列から機能解析へ 第2版  
(<http://www.amazon.co.jp/dp/4895924262/>)

## 講師プロフィール

愛知県名古屋市出身。埼玉大学理学部分子生物学科 卒業、横浜市立大学大学院 国際総合科学研究科 生体超分子科学専攻。博士（理学）。理化学研究所 生命情報基盤研究部門の学生アルバイト・リサーチアソシエイトを経て、現在は理化学研究所 情報基盤センター センター研究員。専門は生化学、分子生物学、物理化学、構造バイオインフォマティクス。学部時代は生化学・分子生物学の実験、大学院時代は膜タンパク質の分子シミュレーションを行い、現職では統合データベースや合成生物学関連のバイオインフォマティクス研究に携わっている。 E-mail: [koro.nishikata@riken.jp](mailto:koro.nishikata@riken.jp)

