eF-site/eF-surf/eF-seekによる蛋白質分子表面の利用

木下賢吾

東大医科研

PDBj講習会2009

@東大駒場キャンパス

kinosita@hgc.jp

http://www.hgc.jp/~kinosita

Today's Topic++

- 背景
- 蛋白質の表面構造とは？
- 表面構造を
  - 見る：eF-site
  - 作る：eF-surf
  - 探す：eF-seek
- 関連サーバ（時間があれば）

どんどん増えるゲノム配列

Genome sequencing projects statistics

Organism	Complete	Draft assembly	In progress	total
Prokaryotes	931	1002	865	2798
Archaea	64	14	30	108
Bacteria	867	988	835	2690
Eukaryotes	22	178	178	378
Animals	4	74	61	139
Mammals	2	28	19	49
Birds		2	1	3
Fishes		3	6	9
Insects	1	21	7	29
Flatworms		2	2	4
Roundworms	1	9	11	21
Amphibians			1	1
Reptiles			1	1
Other animals		11	15	26
Plants	2	11	46	59
Land plants	2	8	41	51
Green Algae		3	5	8
Fungi	10	69	41	120
Ascomycetes	8	55	28	91
Basidiomycetes	1	10	8	19
Other fungi	1	4	5	10
Protists	6	22	26	54
Apicomplexans	1	11	7	19
Kinetoplasts	1	2	5	8
Other protists	4	9	13	26
total:	**953**	**1180**	**1043**	**3176**

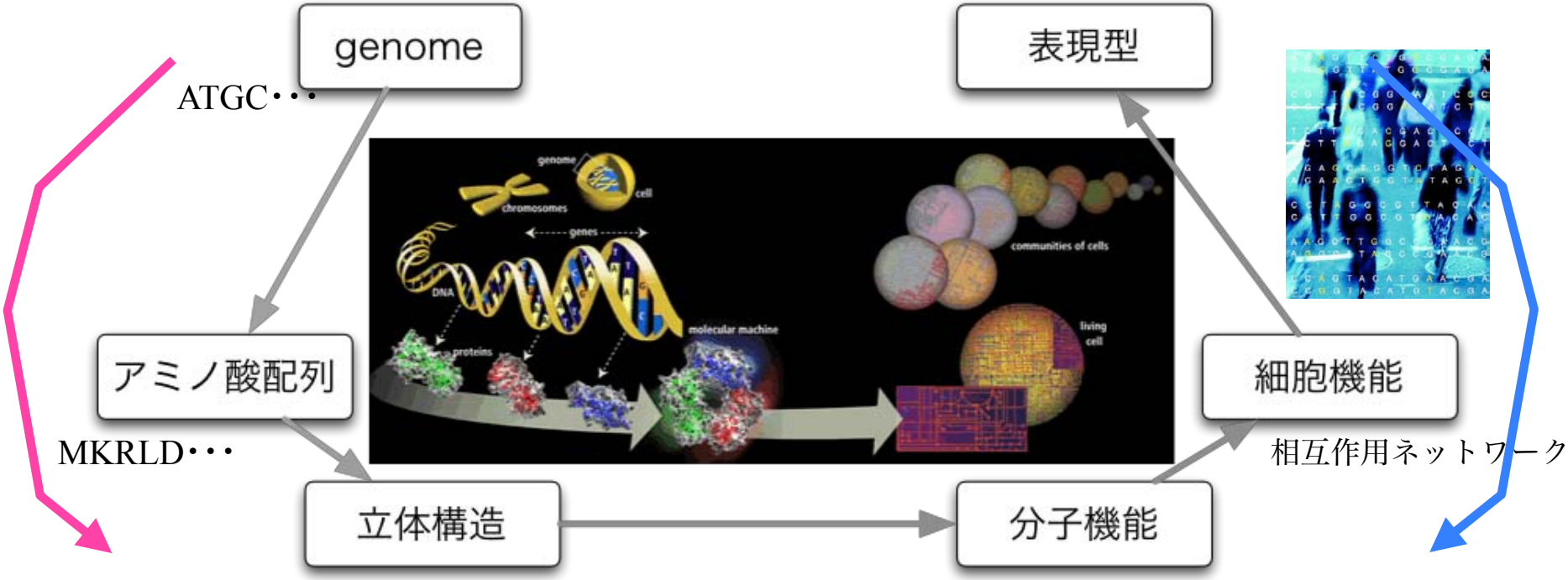
@2009/07/20:3ヶ月で187種増えた

でも約半分の遺伝子産物（蛋白質）の機能が分からない

Kengo Kinoshita
IMS Tokyo University

http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

生物における情報の流れに沿った機能推定

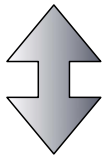


- 遺伝子（or 遺伝子産物の蛋白質）の機能推定
  - すべての遺伝子の機能を実験的に決めるのは不可能
    →計算による機能推定法が必要である
- 2つのアプローチ
  - bottom up approach：遺伝子との対応付けが容易
  - top down approach：高次の機能に対応できる

Kengo Kinoshita
IMS Tokyo University

Bottom up approachの現時点での実際

- 原理的には単体のゲノムを見れば機能がわかるべき
  （←ゲノムは生命情報のすべてだから）



- 実際には分からないので、<u>似たものを探して機能を推定</u>
  - <u>配列</u>：文字列の類似性検索
  - <u>蛋白質の立体構造</u>：3次元の画像認識（today's topic）
  - <u>遺伝子の発現量</u>：遺伝子の共発現

機能未知遺伝子 → 機能既知遺伝子群 → 似た遺伝子から機能推定
or
共通の特徴を抽出

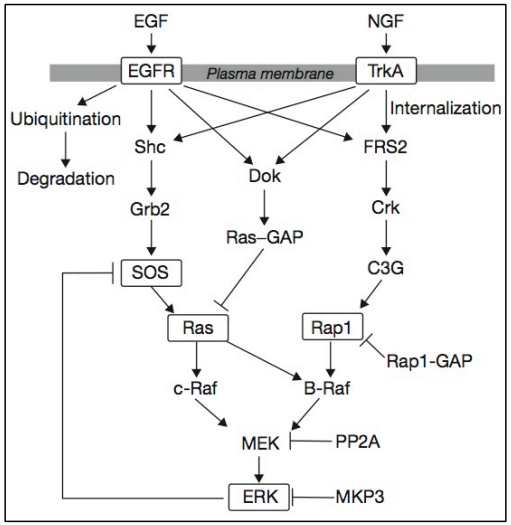
類似性検索

<u>できるだけ弱い類似性も検出したい</u>

構造からの機能予測の対象となる「機能」

- 対象となる機能:　分子機能
  - タンパク質１つで決まる機能
  - 文脈によらない機能

    構造を見ると分かるはず
- 分子機能
  - 基質認識
    - 低分子
    - 高分子→タンパク質間相互作用
  - 酵素反応



文脈による機能の例

シグナル伝達系でのクロストーク

立体構造の比較

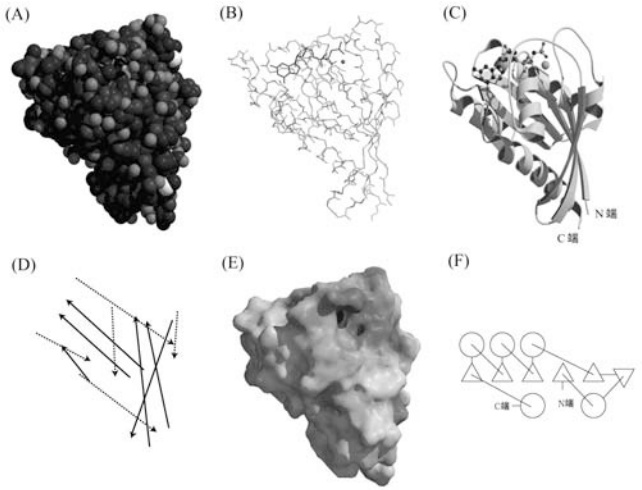
- 問題点
  - 立体構造には様々な表現が存在する
  - 何が似ていれば機能が似ているかが不明

⇩

- アプローチ
  - 様々な見方での比較方を開発し
    機能との相関を調べる研究

⇩

- 分子表面の類似性が良さそう
  - ＋静電ポテンシャル
  - ＋形状として表面の曲率も考慮



配列比較の論理

遺伝子配列が似ている

⇩

(進化的類縁関係)

⇩

機能が似ている

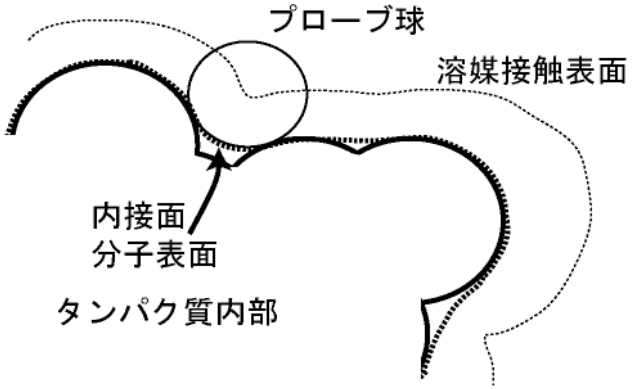
分子表面とは？

- 蛋白質とプローブ球の内接面(半径1.4Å)
- 分子表面の”良い”点
  - 分子表面は配列が似て無くても似てる
  - 原子配置が違っても似ている

⇕

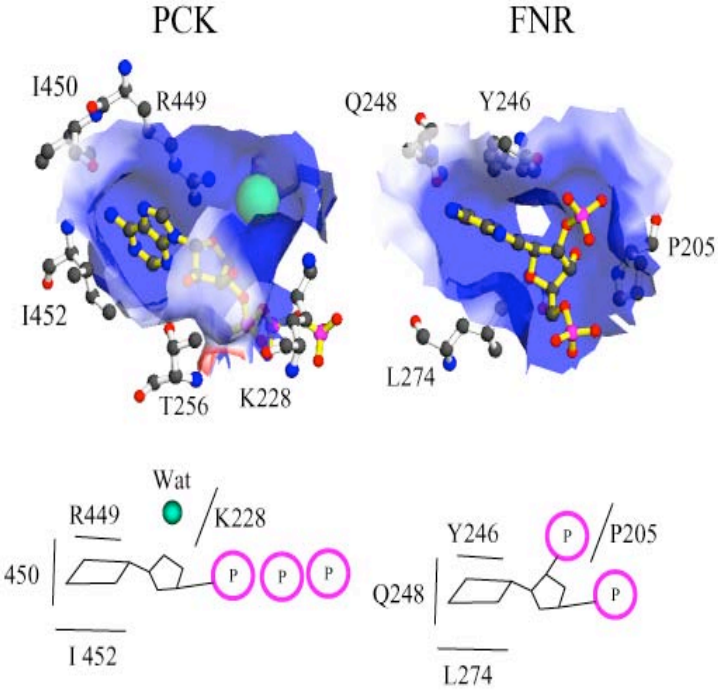
  - 原子の空間配置が似てる

    ≒ 配列が似ている

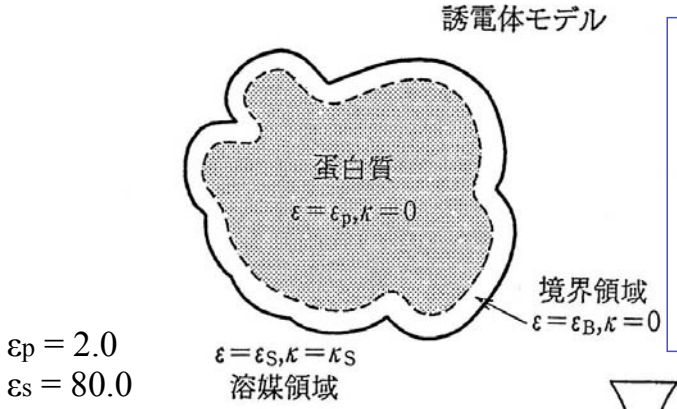


- 計算時間
  - 配列＜＜＜立体構造

    → 配列比較では検出できない類似性を検出したい！



蛋白質内外の電位の計算：連続体モデル



電磁気学の方程式

蛋白質領域　　*Poisson* 方程式

$$\nabla(\varepsilon(r)\nabla\varphi) = -4\pi\sum_i q_i\delta(r-r_i),$$

溶媒領域

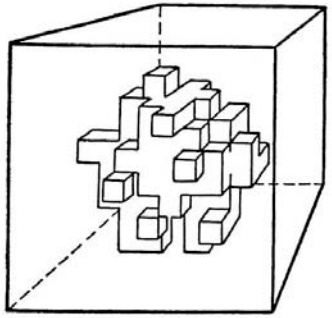
$$\nabla(\varepsilon(r)\nabla\varphi) = \varepsilon_S\kappa^2\varphi,$$ *Poisson-Boltzmann* 方程式

境界領域

$$\nabla(\varepsilon(r)\nabla\varphi) = 0.$$ *Laplace* 方程式

差分方程式に変換して数値計算を行う。

$$\frac{1}{h^2}\sum_{a'b'c'}[\bar{\varepsilon}_{a'b'c'}(\varphi_{a'b'c'}-\varphi_{abc})] = -\rho_{abc} + \varepsilon_S\bar{\kappa}^2_{abc}\varphi_{abc}$$



ブリックモデル（六面体要素）

図 3.7　蛋白質の誘電体モデルに対するポアソン・ボルツマン方程式を，ブリックモデルによって差分化し，数値的に解く手法の説明

上部のもとの連続体モデルでは，蛋白質・溶媒・両者の境界の各領域に対する電磁気学の静的微分方程式が成り立っている．蛋白質とその周囲の溶媒領域とを小立方体の集合で表すブリックモデルによって，これら微分方程式は，(a, b, c)で番号づけられた，それぞれの小立方体における差分方程式に変形される．

Kengo Kinoshita
IMS Tokyo University

eF-site database: http://ef-site.hgc.jp

PDBj Protein Data Bank Japan

eF-site
electrostatic surface of Functional-site

About eF-site | Tools | References | Links | Acknowledgements | Feedback

319647 Entries, Last Update: 26-Mar-2009

Keyword Search

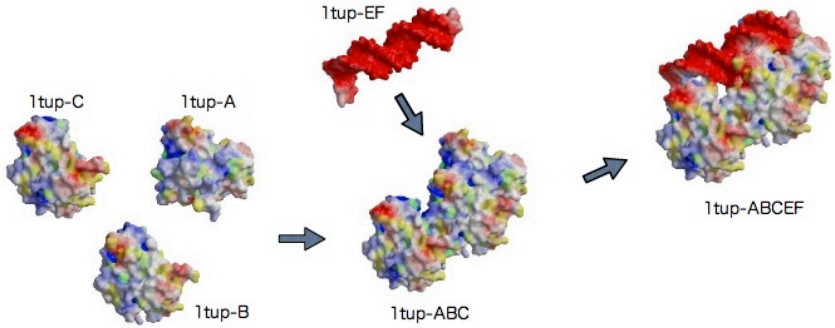
PDB code only (•) and ( ) or

Search Reset

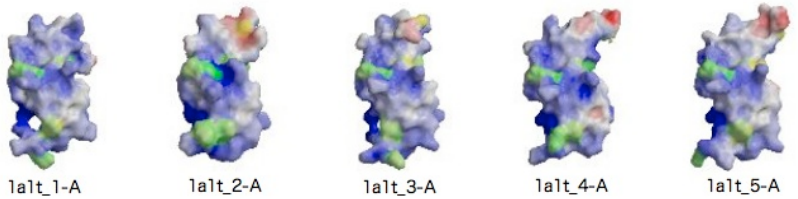
Category Search

- Antibody
- Prosite
- Active Site
- Membrane
- Binding Site

Examples of molecular surface



For all pdb entries, molecular surfaces are generated for individual subunits of proteins and the complex of proteins. In addition, when double strands DNA is included in the entry, a molecular surface for each dsDNA is also stored.



When multiple models are stored in PDB, molecular surfaces for all the models are generated.

ほぼすべてのPDBエントリーを計算

サブユニット毎に計算

NMRはすべてのモデルを計算

eF-site ID

検索の仕方

PDBj Protein Data Bank Japan

eF-site

electrostatic surface of Functional-site

About eF-site | Tools | References | Links | Acknowledgements | Feedback

281730 Entries, Last Update: 19-Apr-2008

Keyword Search

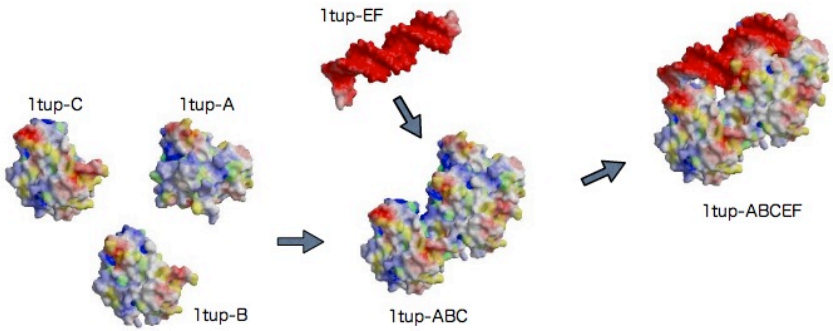
PDB code only ◉ and ○ or

Search Reset

Category Search

- Antibody
- Prosite
- Active Site
- Membrane
- Binding Site

Examples of molecular surface



For all pdb entries, molecular surfaces are generated for individual subunits of proteins and the complex of proteins. In addition, when double strands DNA is included in the entry, a molecular surface for each dsDNA is also stored.

Top pageから検索

OR

PDBjからリンクをたどる

xPSSS xml based Protein Structure Search Service

Summary [1C7K]

About xPSSS
Update Information
wwPDB Snapshot Mirror

Summary | Structural Details | Experimental Details | Functional Details | Sequence Neighbor | Download/Display | Link

search ◉ PDB ID ○ keyword

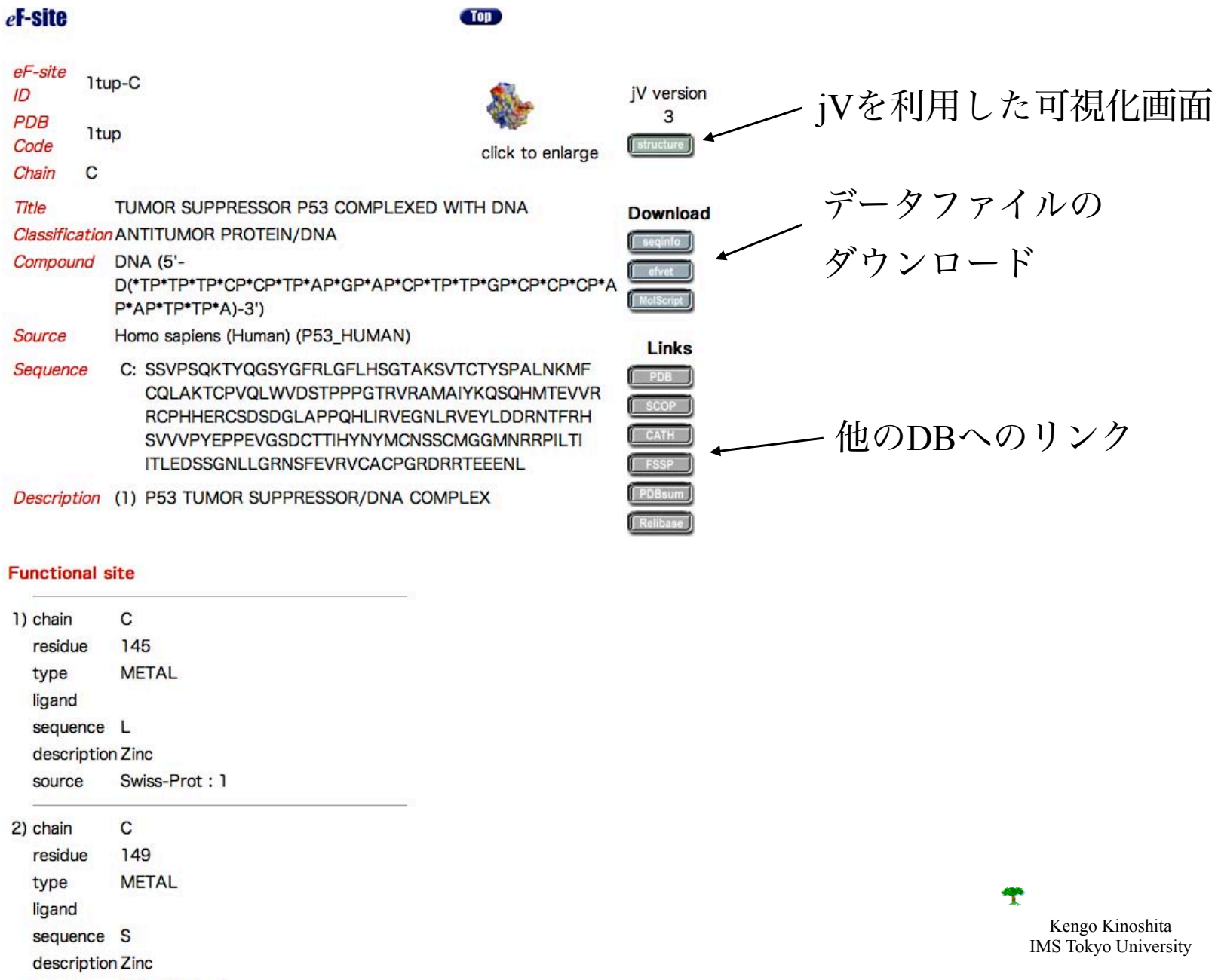
PDB ID	1C7K sequence information (FASTA format) download PDB format file
Descriptor	ZINC ENDOPROTEASE (E.C.3.4.24.-)
Title	CRYSTAL STRUCTURE OF THE ZINC PROTEASE
Functional Keywords	alpha and beta protein, METALLOPROTEINASE HYDROLASE
Biological source	Streptomyces caespitosus (BACTERIA)
Cellular location	[UNP - SNPA_STRCS] Secreted
Total number of polymer chains	1
Total molecular weight	14492.9 (the details in Structural Details Page)
Authors	Kurisu, G. , Harada, S. , Kai, Y. (*deposition date* : 2000-02-19, *release date* : 2001-04-25)
Primary citation	Kurisu, G. , Kai, Y. , Harada, S. Structure of the zinc-binding site in the crystal structure of a zinc endoprotease from Streptomyces caespitosus at 1 Angstroms resolution. *J.Inorg.Biochem.*, 82:225 - 228, 2000.(PubMed : 11132632)
Other Database Information	CATH , CE , FSSP , SCOP , VAST , UniProt (UNP - P56406) , eF-site , KEGG (EC 3.4.24.-)

Structure Images

normal position 250X250 500X500 | rotated about x by 90° 250X250 500X500 | rotated about y by 90° 250X250 500X500 | viewer

jV version3 is used with Java(TM)Plug-in 1.5 or later.

各エントリーのページ：Summary Page



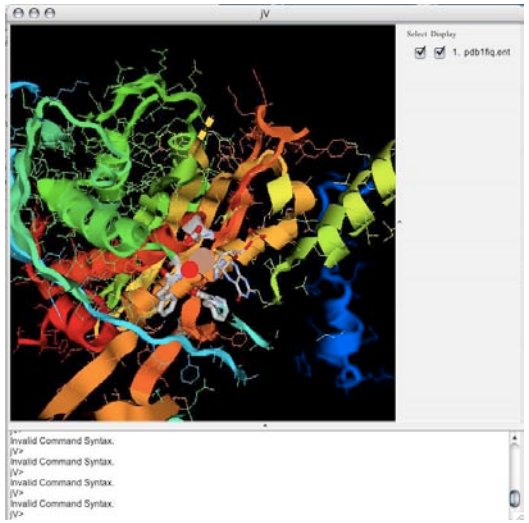
jVを利用した可視化画面

データファイルの
ダウンロード

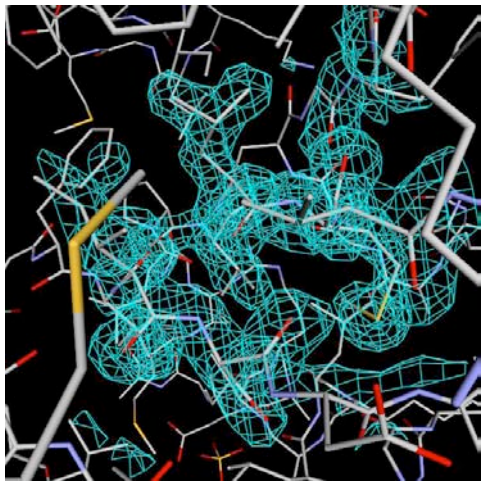
他のDBへのリンク

jV formerly known as PDBjViewer

- Java + JOGL
  - 見た目は Rasmolと似ている、Rasmol-likeなスクリプトも使える
- stand aloneでもappletでも動く
- PDB-ML, PDB, polygon (分子表面などの”オブジェクト”)フォーマットに対応
- 複数分子を扱える
- xpsssの機能情報の表示
- windows XP (sp2), Mac OS-X (10.3), vine3.1やfedora core 3などの主なLinux
- OpenGL supported graphic card and latest graphic driver
- http://ef-site.hgc.jp/wiki/jV/

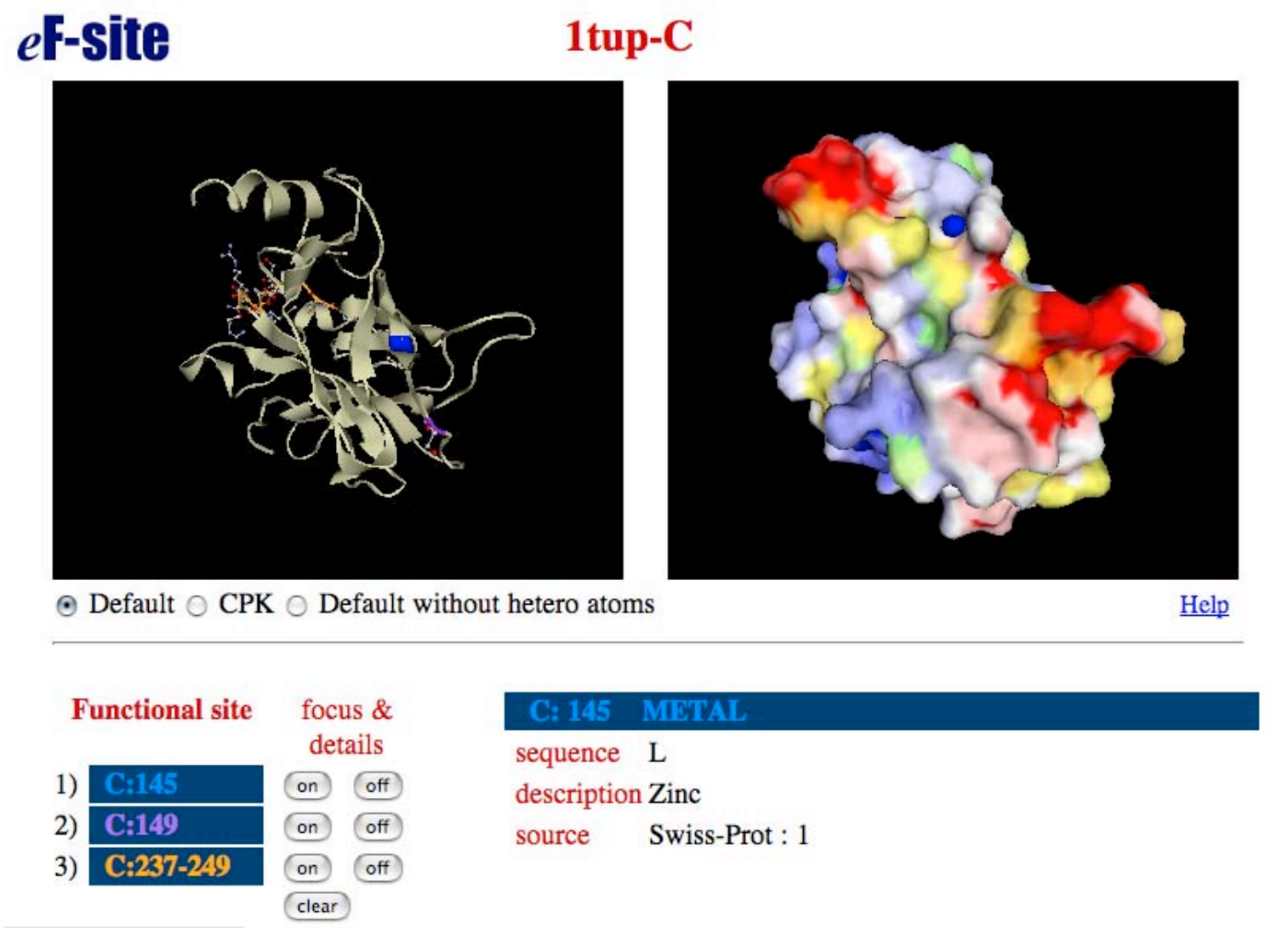


スタンドアローン版



オブジェクトの例
電子密度マップ

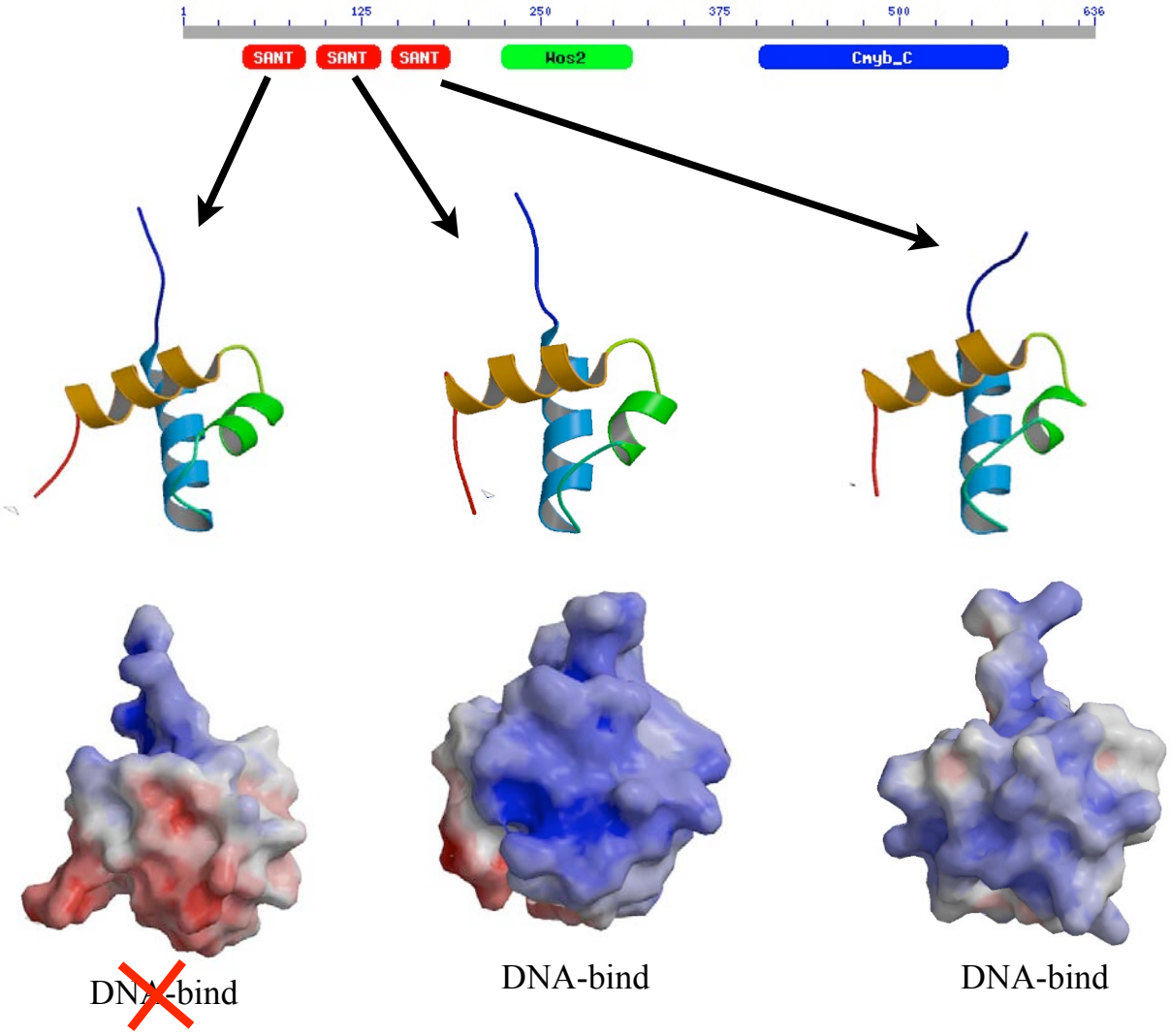
可視化画面：Structure Page



左右の画面が連動して回る

分子表面を眺めればそれで分かることもある

- （例）Myb proto-oncogene protein



DNA-bind　　DNA-bind　　DNA-bind

データファイルのダウンロード

- seqinfo
  - 機能部位情報を納めたXML形式のファイル
  - http://ef-site.hgc.jp/eF-site/schema/seqinfo20.xsd
- efvet-ml
  - 表面の形状と静電ポテンシャルの値を納めたXML形式のファイル → jVで表示可能
  - http://ef-site.hgc.jp/eF-site/schema/efvet30.xsd
- molscript file (次のスライド)
- efvet-flat
  - 基本的には内部利用
  - まとめてダウンロードする際に利用
    - http://ef-site.hgc.jp/eF-site/tools.html
    - 同じ所にXMLへのコンバーターも用意

Molscript fileの使い方

- Molscript (Kraulis, J. Appl. Cryst, v24, p946-950, 1991)
  - PDBファイルからタンパク質の絵を描く
  - 絵を描くことに特化していて綺麗
  - 柔軟なコマンドでこった表現も可能
    - Objectコマンドでポリゴンも表示可能
      - 分子表面はこのコマンドで使うファイルとして提供

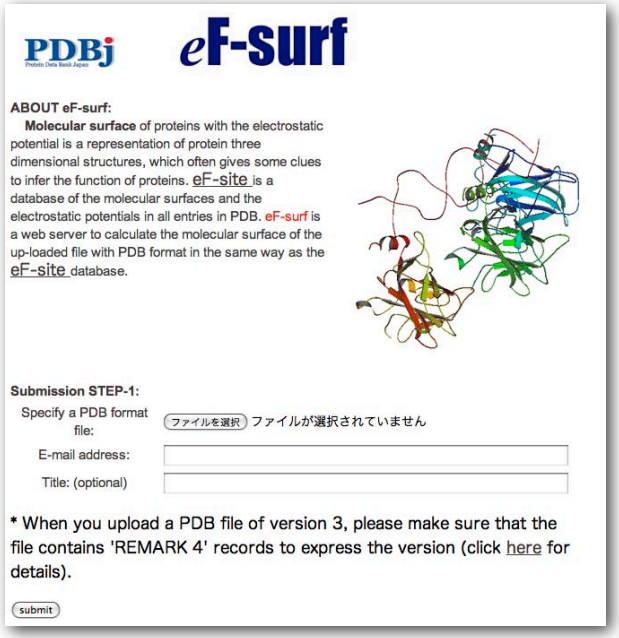
ex) molscript input file

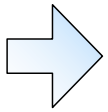
```
plot
  background white;
  read mol "1rop.pdb";
  transform atom * by centre position atom *;
  object 1rop.obj;
end_plot
```

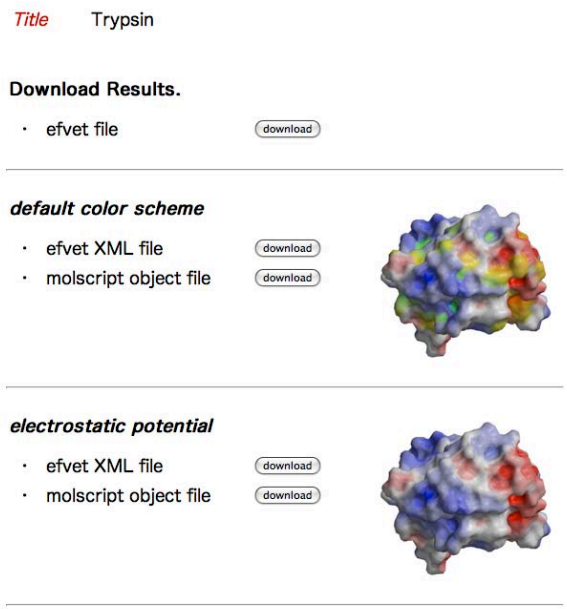
この例では1rop.objがobject file

分子表面の計算
eF-surf@http://ef-site.hgc.jp/eF-surf

- eF-siteと同じ条件で分子表面を計算
- 結果はjV用 XMLファイルとmolscript形式でダウンロード出来る



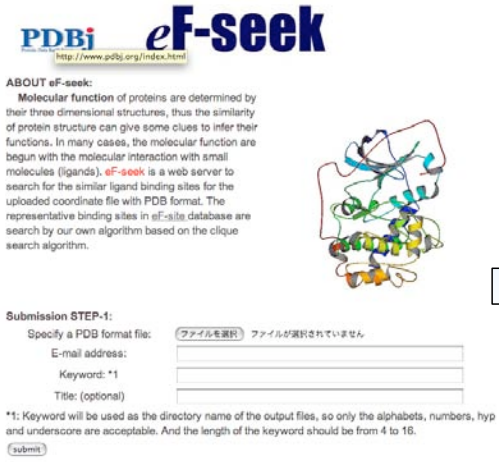


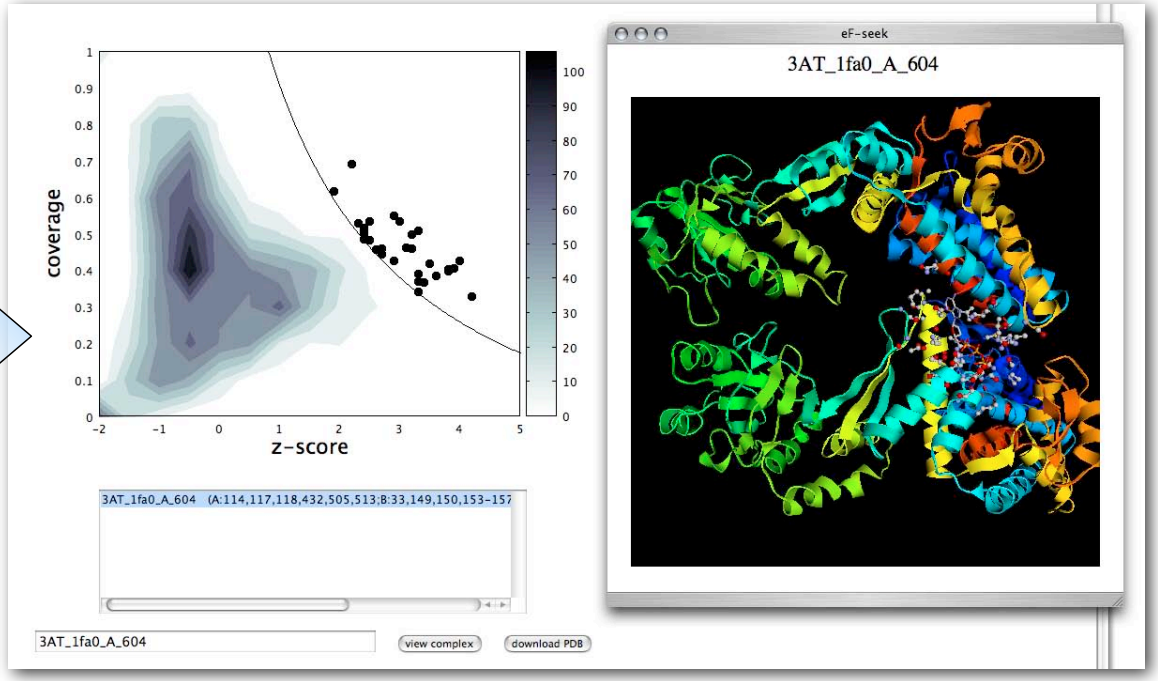


1) PDBファイルをupload
2) メールアドレスを入力
3）確認メールにあるURLにアクセスしてジョブを開始
4) 15min-60min程度で終わる（予定

低分子結合部位の予測
eF-seek@http://ef-site.hgc.jp/eF-seek

- eF-siteに対する類似性検索による機能部位の予測
- 代表結合部位に対して検索
- 2006年12月から運用開始





アップロードされたPDBファイルに対して機能部位を予測し、複合体の構造を返す

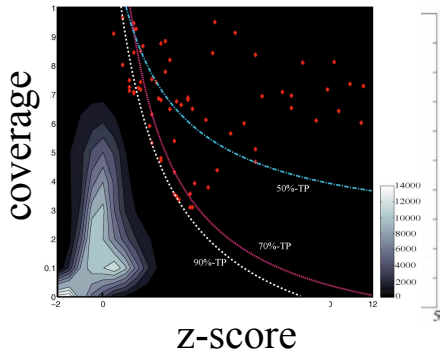
Normalization of similarity score

N hetero compound binding sites appeared in PDB

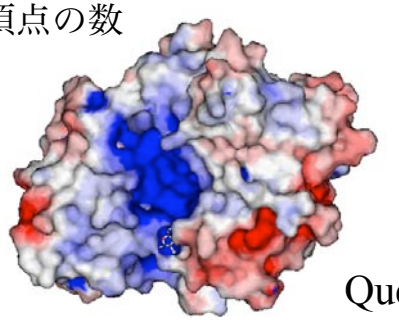
Functional site patches

Normalization from
query protein's view.

Z-score = (score – mean)/std

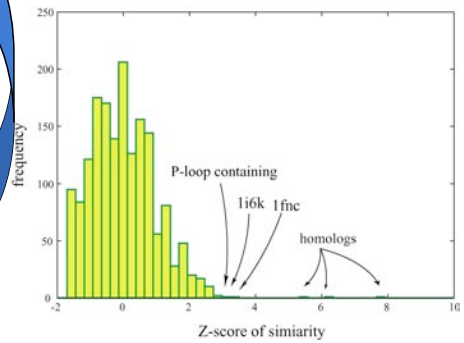


similarity score
= 対応する頂点の数



Query protein

Larger patches would get larger Z-score.



Normalization from
functional sites' view

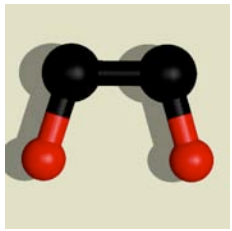
$$\text{coverage} = \frac{\#\ \text{of corresponding vertices}}{\#\ \text{of vertices in each patch}}$$

Results will be shown in **coverage vs. Z-score plot**.

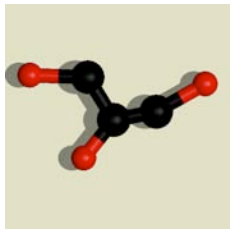
The number of corresponding vertexes is used as similarity score.

Threshold line determination

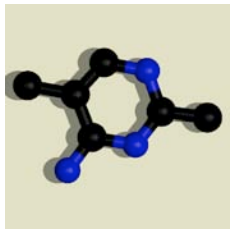
- 10 randomly selected representative with free and complex structures.
  - Homologous proteins with similar ligands are considered to be “correct”.



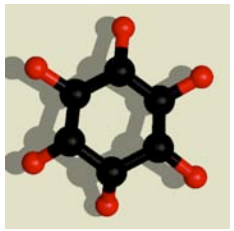
Ethylene glycol



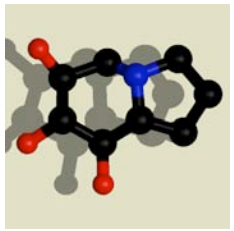
Glycerol



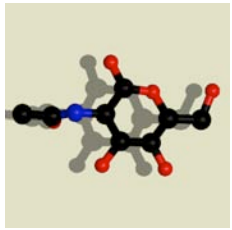
2,5-dimethyl-pyrimidin-4-ylamine



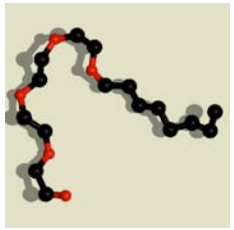
Myo-inositol



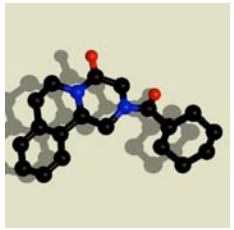
Castanospermine



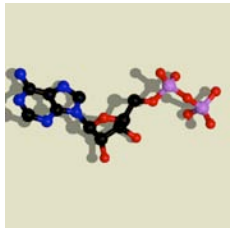
N-acetyl-d-galactosamine



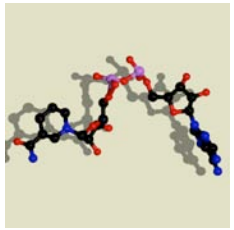
(Hydroxy ethyloxy)tri (ethyloxy)octane



Praziquantel

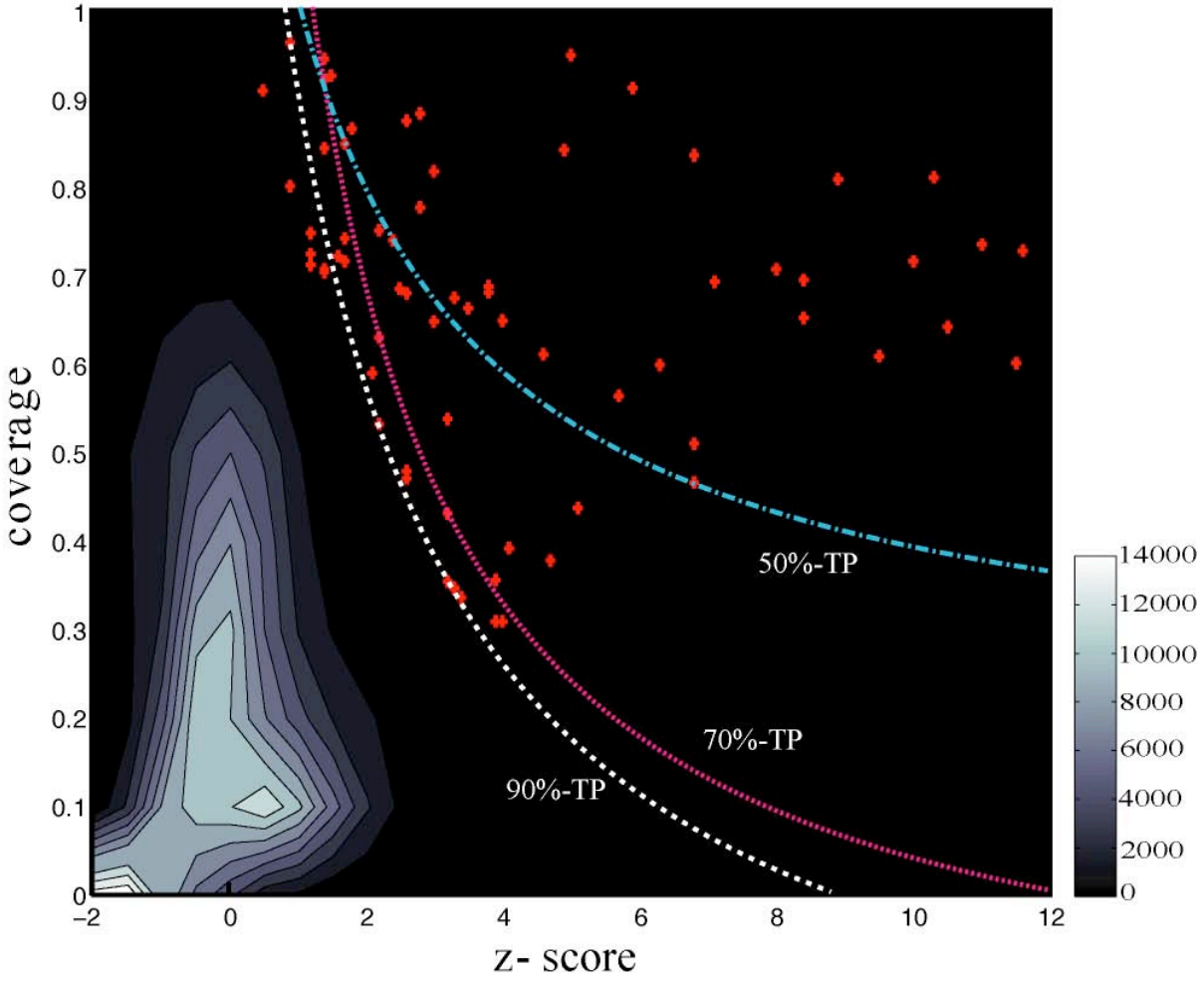


ADP



NAD

Threshold line determination



Maximize *CC*

with a threshold line

$$cov. = \frac{a}{z+b} + c$$

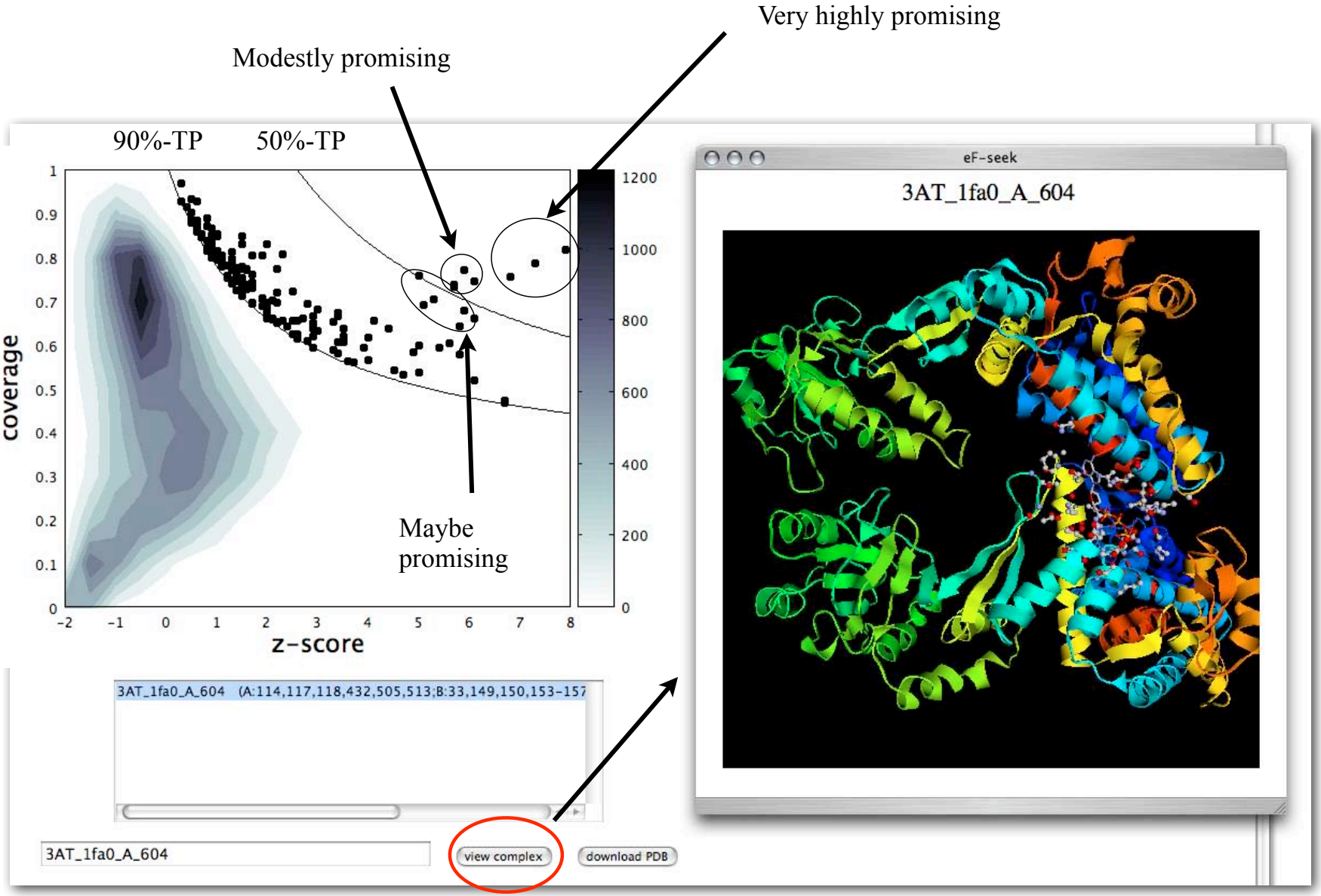
under a constraint that

fraction of TP exceed 90%, 70% or 50%

90%-TP line will be used hereafter.

$$CC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

結果の見方



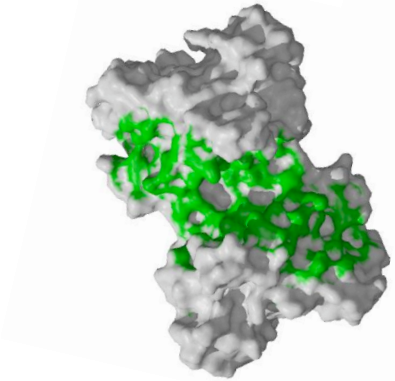
良さそうなのを選んでView complexを押す

関連サーバ

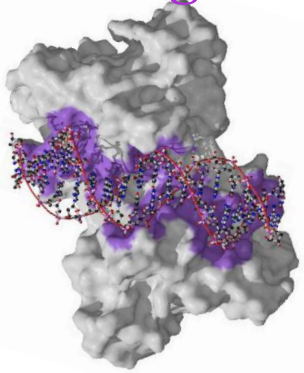
DNA結合部位の予測

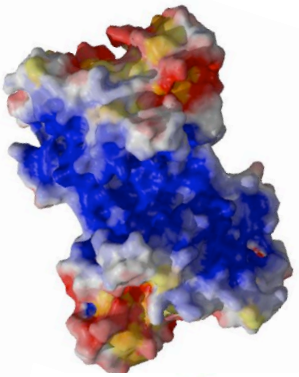
Methionine Repressor
(1MJQ)

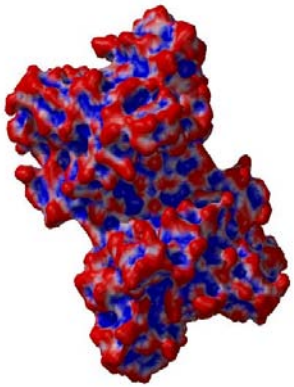
Predicted site

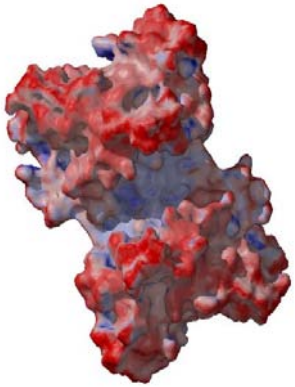


Binding site









Electrostatic potential

Red: negative, blue: positive

Yellow: hydrophobic surface

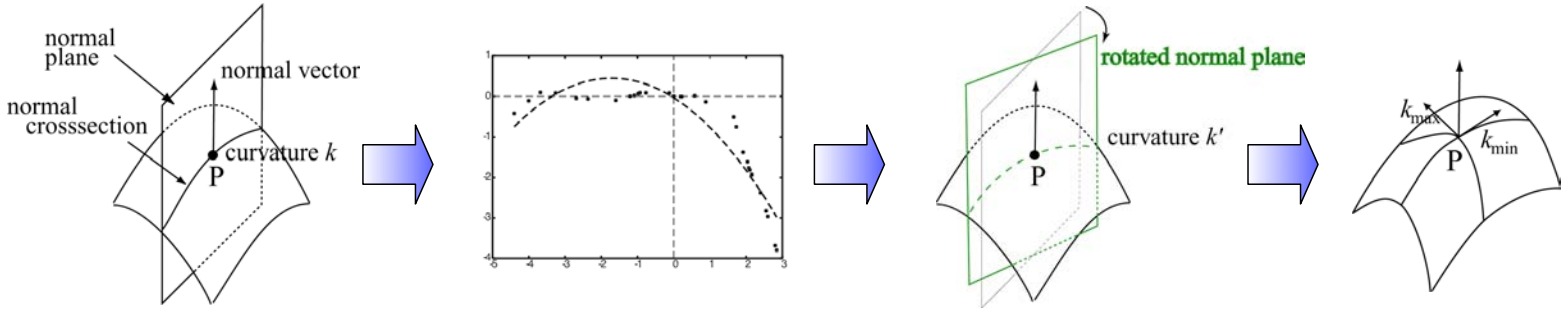
Local curvature relating to DNA-binding directly: slightly protruded

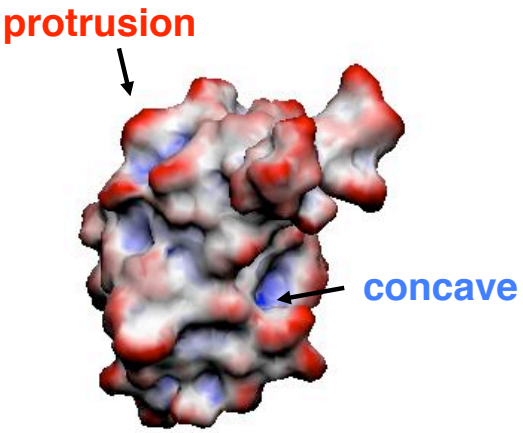
Red: protrusion, Blue: concave

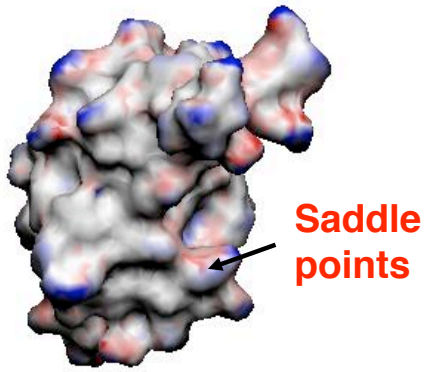
Global curvature relating to entire surface geometry: largely concave

3つの特徴からの予測法

分子表面の曲率計算







Mean Curvature
$= (k_{max}+k_{min})/2$

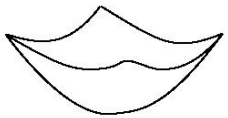
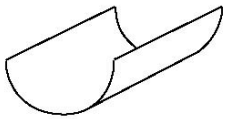
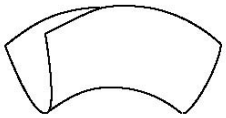
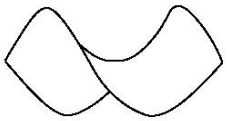
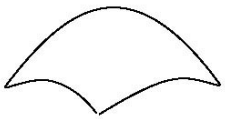
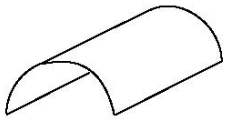
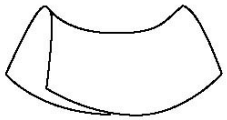
Gaussian Curvature
$= k_{max} \cdot k_{min}$

Kengo Kinoshita
IMS Tokyo University

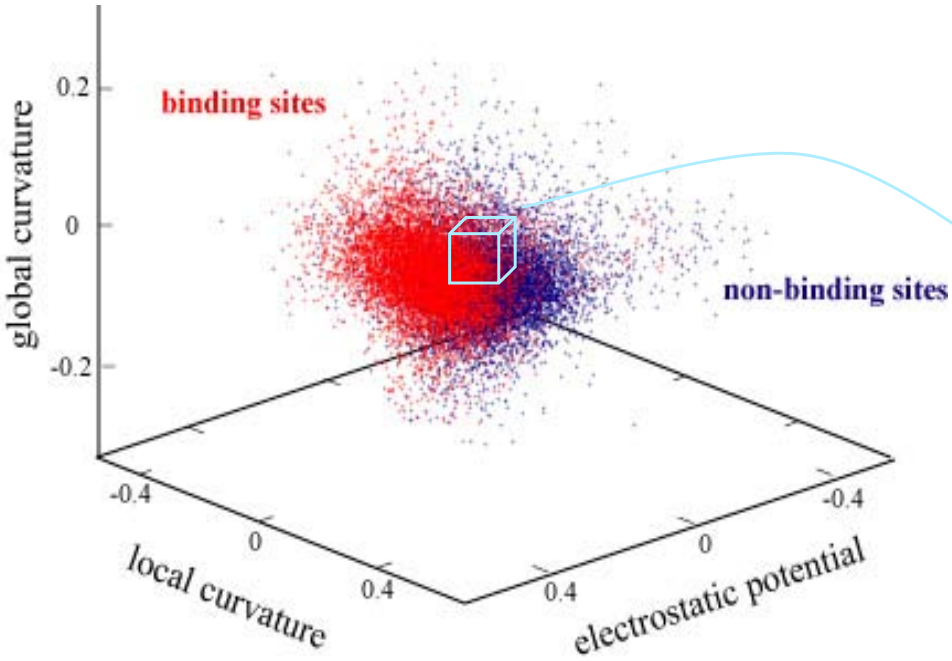
Gauss curvature

Mean Curvature(H)
$= (k_{max}+k_{min})/2$

Gaussian Curvature(K)
$= k_{max} \cdot k_{min}$

	$K > 0$	$K = 0$	$K < 0$
$H > 0$	 concave	 valley	 concave saddle
$H = 0$	*none*	flat	 bending saddle
$H < 0$	 convex	 ridge	 convex saddle

Prediction Scheme-1: Statistical Preference Measure



Distribution of electrostatic potential, local curvature and global curvature for all proteins in dataset-1

Relative Frequency

$$F_{bind}(\varphi_e, K_{local}, K_{global}) = N_{bind}(\varphi_e, K_{local}, K_{global}) / N_{bindtotal}$$

$$F_{non\text{-}bind}(\varphi_e, K_{local}, K_{global}) = N_{non\text{-}bind}(\varphi_e, K_{local}, K_{global}) / N_{non\text{-}bindtotal}$$

Statistical Preference Measure

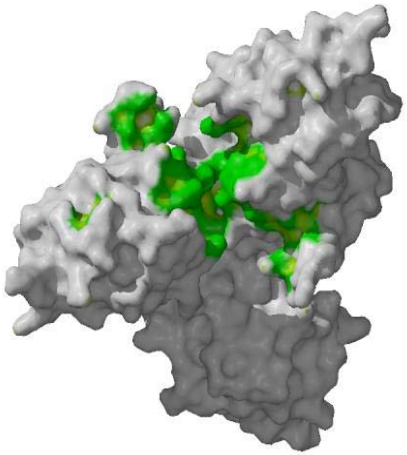
$$P_{bind} / P_{non\text{-}bind}$$

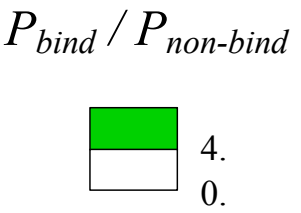
Prediction Scheme-2: Prediction Score (Pscore)

Statistical Preference Measure

$$P_{bind} / P_{non\text{-}bind} > 4.0$$

For each vertex, calculate the measure and colour it when the value exceed 4.0.





Pscore = max (Parea / Whole area)

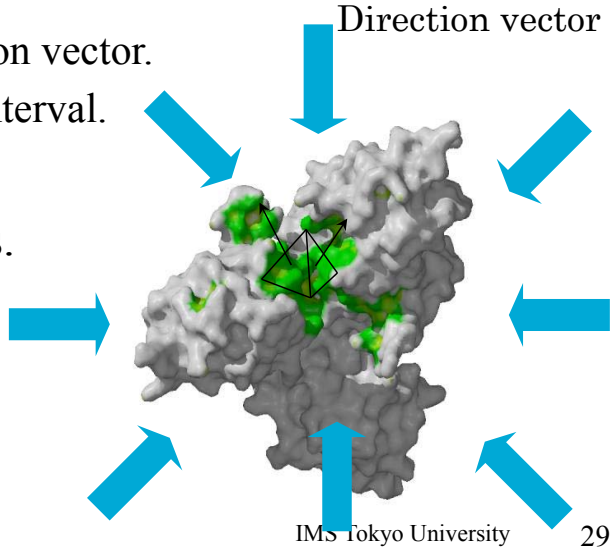
Parea : predicted DNA-binding *weighted* area *for a given direction*

Whole area: whole *weighted* area *for a given direction.*

Weights are calculated as inner product of normal vector and direction vector.

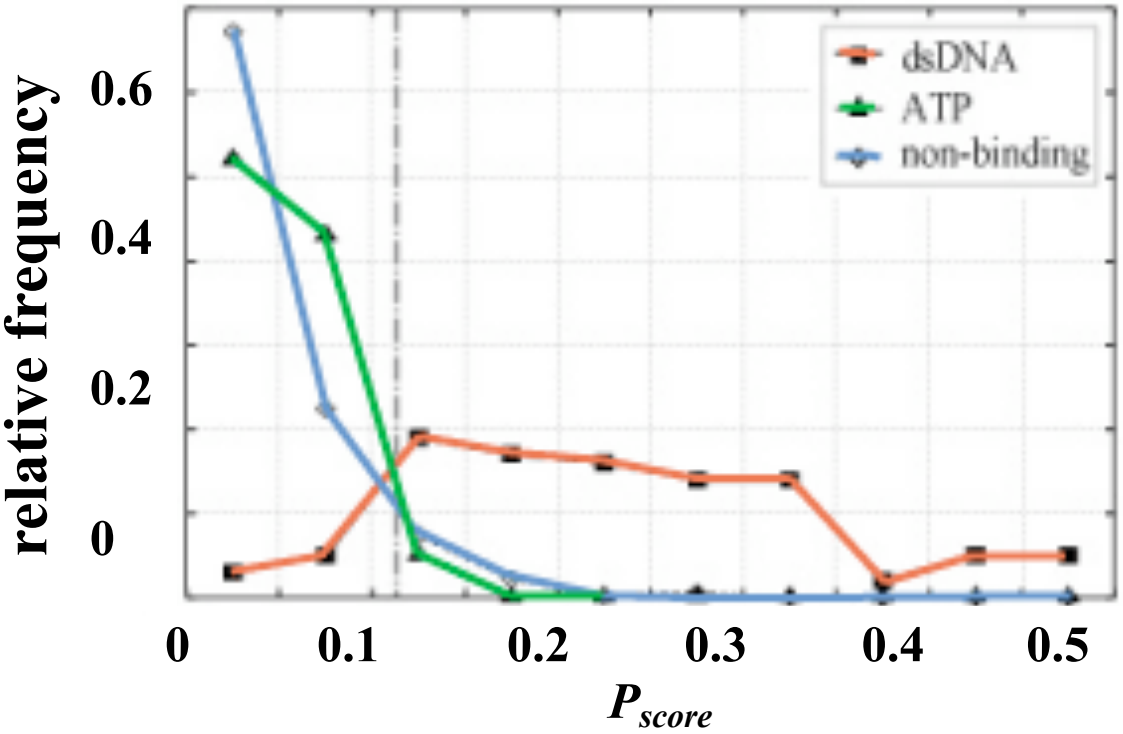
Maximization was done by searching all possible direction by 10° interval.

Pscore will be used as an indicator of the prediction results.



Prediction Results

Tsuchiya et al. (2004) *PROTEINS*, 55, 885-894.



Histogram of Prediction score for dsDNA-binding proteins (63), ATP-binding proteins (21), and non-dsDNA-binding proteins (406)

86% accuracy for predicting dsDNA-binding proteins, and

96% accuracy for predicting non-DNA-binding proteins including ATP-binding proteins.

PreDs: Predition of DNA-binding site

PreDs

Prediction of DNA-bidning site

PreDs is a server making the prediction of dsDNA-binding site on protein surfaces, according to our prediction method developed by focusing our attention on the shape of the molecular surface and the electrostatic potential on the surface.

Please input your E-mail address, your name and job title, and upload a **protein coordinate file with PDB format** of the query structure or **PDB ID**. E-mail address is essential since we will return a mail with a URL at which you will find the results.

HELP

Introduction
Usage
Prediction Method
Reference

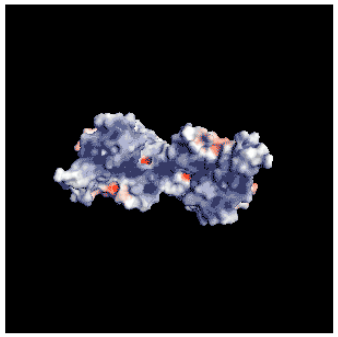
E-mail Adress	
Your Name	
Job Title	

Choose uploading a coordinate file or inputting PDB ID.
If both boxes are filled, the server chooses using the uploaded coordinates.

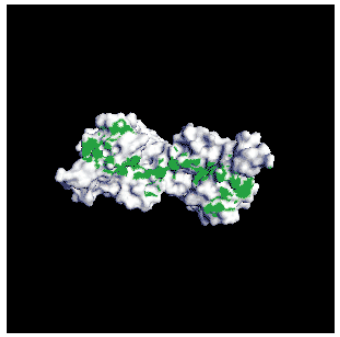
○ Upload Coordinate File	Browse...
○ Input PDB ID	

SUBMIT
RESET

Electrostatic Potential



Predicted DNA-Binding Surface



Predicted DNA-Binding Region (Residue Numbers (Chain ID))

A2(B) A47(A) A48(A) A48(B) A55(A) A55(B)
A96(A) A96(B) R61(A) R61(B) R66(A)
R66(B) R74(A) R74(B) R104(B) R140(A)
R140(B) N57(A) N57(B) N119(A) N119(B)
N123(A) N123(B) D117(A) D117(B) Q63(A)
Q63(B) Q70(B) Q141(A) G33(A) G51(A)
G51(B) G53(A) G53(B) G76(A) G76(B)
H27(B) H73(B) H78(A) H78(B) H98(A) I68(B)
L46(A) L46(B) L116(A) L116(B) K56(A)
K56(B) K65(A) K65(B) K92(A) K92(B)
K120(A) K120(B) F54(A) F54(B) P49(A)
P49(B) P59(A) P59(B) P126(A) P126(B)

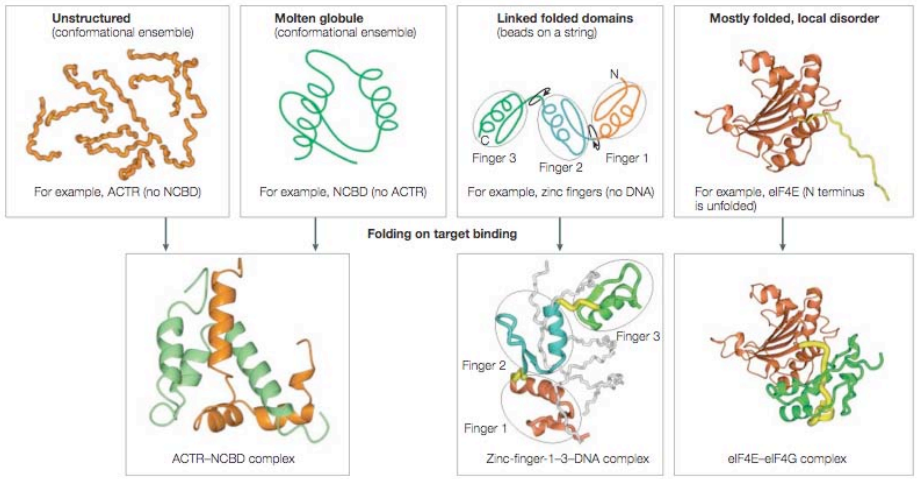
http://pre-s.protein.osaka-u.ac.jp/~preds/

Disorder region とは何か？

- 機能していない状態では一定の構造をとらない領域
  - 定量的な定義はまだ無いが、いくつかの実験手法により決定される
  - ヒトなどの高等な生物に多い
    - 薬剤開発などの際に重要になる
  - 機能を果たす際には決まった構造をとる (Dyson, HJ and Wright, PE, 2005)
    - 様々なタイプが知られている

Disorder regionを予測できれば機能部位が分かる

様々なDisorderのタイプ



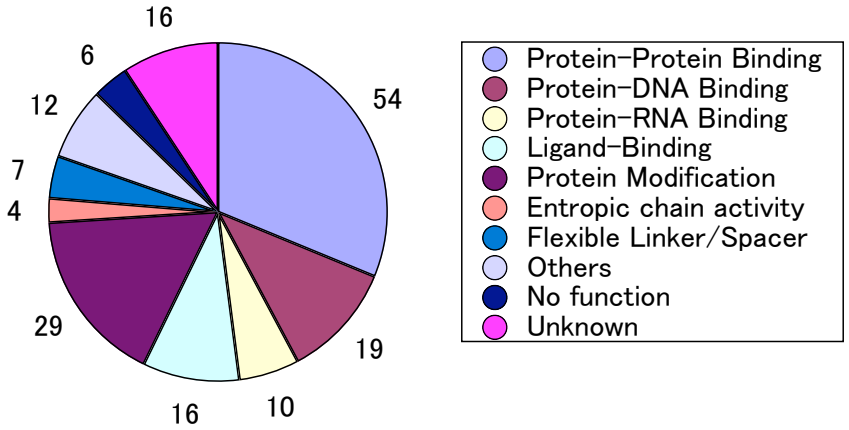
タンパク質disorder領域とその機能

- タンパク質の機能に関わるdisorder領域が存在する
  - Dunker et .al, *Biochemistry*, 2002

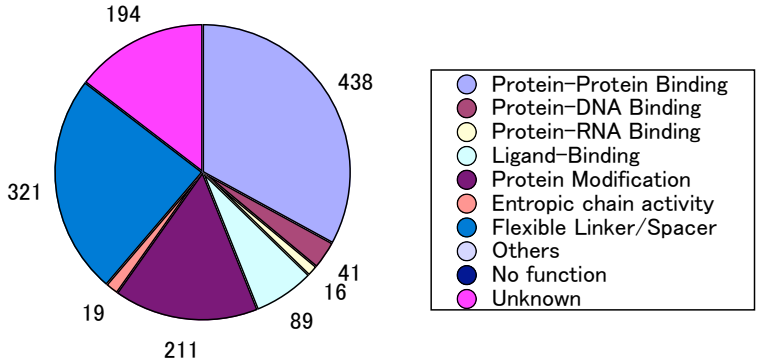
1. 分子認識(Molecular recognition)
   1. タンパク質-タンパク質結合
   2. タンパク質-DNA結合
   3. タンパク質-RNA結合 (t, r, m)
   4. リガンド結合

2. 分子集合体(Molecular assembly)

3. タンパク質修飾(Protein modification)
   - リン酸化
   - アセチル化
   - グリコシル化



90タンパク質での機能の分布 (Dunker et .al, *Biochemistry*, 2002)



DisProt (http://www.disprot.org/), 469 proteins, 1114 regionsでの機能の分布

Statistics of disordered region

高等生物になるほどDisordered 領域が多い

Kingdom	# of proteins	Disorder freq. (% of aa)	Length > 30 (% of chains)	Length > 50 (% of chains)
Archaea	11,742	3.8	2.0	0.7
Bacteria	35,389	5.7	4.2	1.6
Eukaryota	88,531	18.9	33.0	19.6

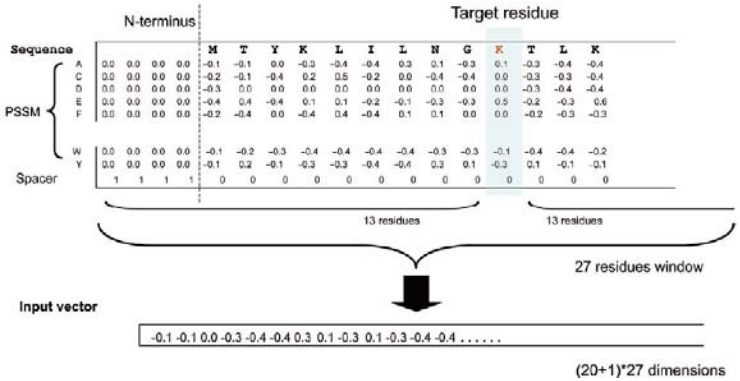
Ward *et al*, JMB, **337**, 635-645, 2004

Estimation by DISOPRED2 (Jones *et al*)

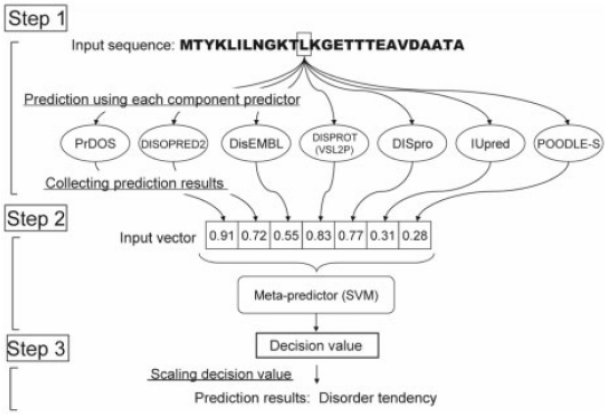
Disorder regionの予測法の開発

Ishida & Kinoshita, 2007, 2008

- Support Vector Machine を利用
- 入力配列→
  - 予測する前後１３残基のアミノ酸配列
  - 近縁蛋白質での置換頻度も考慮 (PSSM)
- その後、複数の方法を組み合わせたMetaアプローチで改良

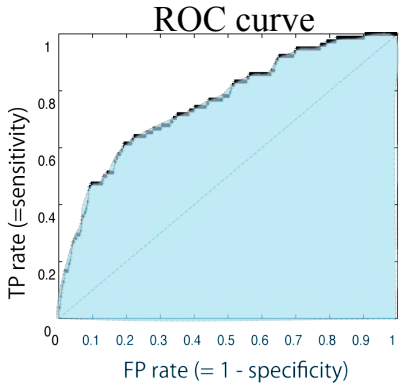


meta-predictorの流れ



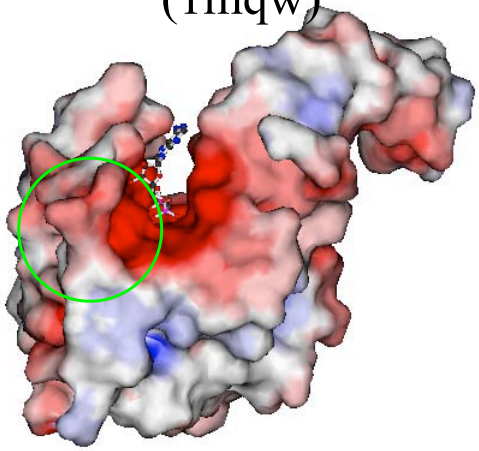
10-fold cross validationによる性能評価

	MCC	ROC
PrDOS	0.502	0.887
DISpro	0.496	0.886
DISPROT(VSL2P)	0.427	0.883
DISOPRED2	0.439	0.864
POODLE-S	0.379	0.850
IUpred	0.406	0.831
DisEMBL	0.397	0.822
meta	0.526 (±0.009)	0.904 (±0.004)

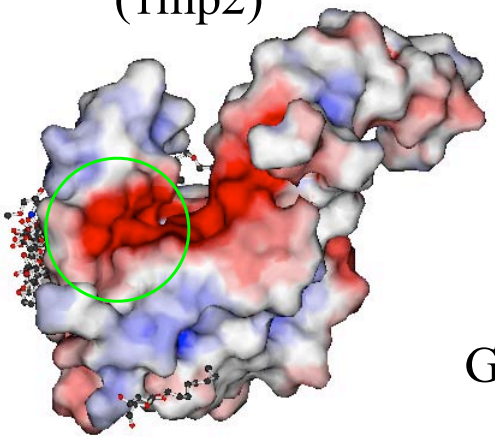


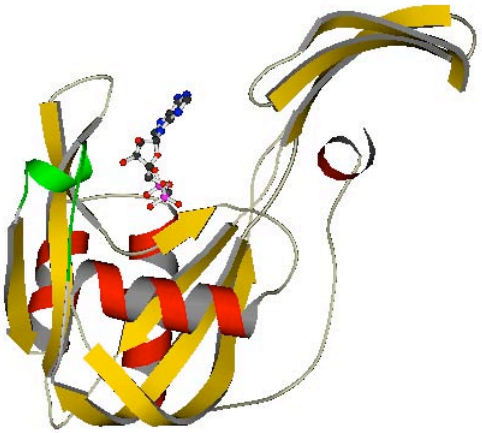
タンパク質リガンドでの例

Complex form
(1mqw)



Free form
(1mp2)



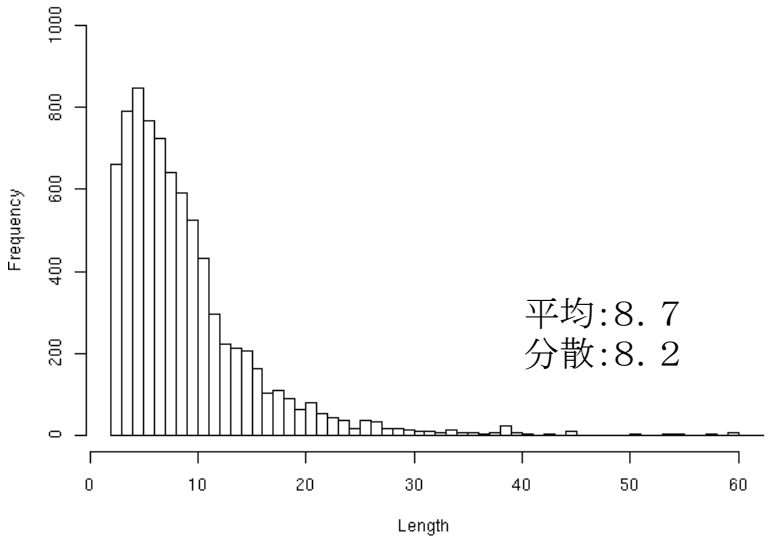


Green parts is flexible and invisible in the free form of the structure, which could prevent our method to predict the binding site.

PDBには見えないループがたくさんある

- Missing loops (or gaps) are identified by comparing SEQRES and ATOM record in each PDB file.
- N-terminal and C-terminal gaps are ignored.
- 7,949 loops are invisible among 41,417 chains in current PDB
  - Apr, 2005
  - Entries with 2.5 Å or better resolution
- About 63% missing loop are 8 or less residue long
  - 8 residues are said to be the threshold value to build the good model.

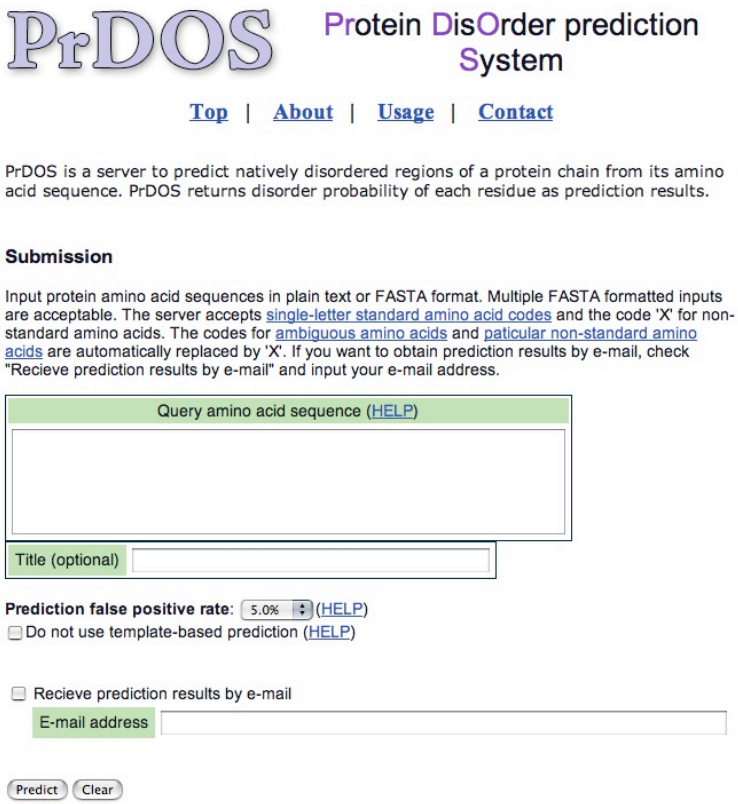
Length distribution of missing loop in PDB



IDPの予測: PrDOS

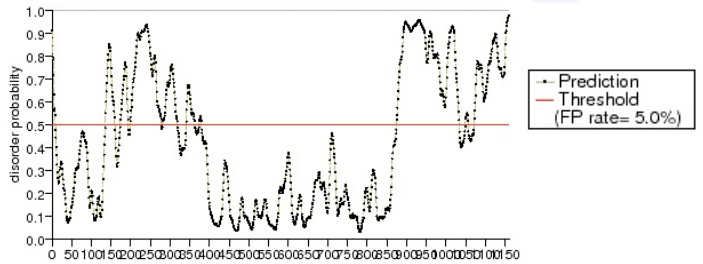
- アミノ酸配で特徴がある
  - 荷電性残基（Gluなど）が多い、配列が単調 ... etc
- 配列から高い精度で予測できる

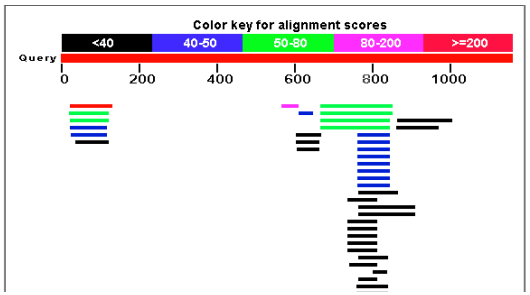
http://prdos.hgc.jp



(例）hERG

1	MPVRRGHVAP	QNTFLDTIIR	KFEGQSRKFI	IANARVENCA	VIYCNDGFCE	50
51	LCGYSRAEVM	QRPCTCDFLH	GPRTQRRAAA	QIAQALLGAE	ERKVEIAFYR	100
101	KDGSCFLCLV	DVVPVKNEDG	AVIMFILNFE	VVMEKDMVGS	PAHDTNHRGP	150
151	PTSWLAPGRA	KTFRLKLPAL	LALTARESSV	RSGGAGGAGA	PGAVVVDVDL	200
201	TPAAPSSESL	ALDEVTAMDN	HVAGLGPAEE	RRALVGPGSP	PRSAPGQLPS	250
251	PRAHSLNPDA	SGSSCSLART	RSRESCASVR	RASSADDIEA	MRAGVLPPPP	300
301	RHASTGAMHP	LRSGLLNSTS	DSDLVRYRTI	SKIPQITLNF	VDLKGDPFLA	350
351	SPTSDREIIA	PKIKERTHNV	TEKVTQVLSL	GADVLPEYKL	QAPRIHRWTI	400
401	LHYSPFKAVW	DWLILLLVIY	TAVFTPYSAA	FLLKETEEGP	PATECGYACQ	450
451	PLAVVDLIVD	IMFIVDILIN	FRTTYVNANE	EVVSHPGRIA	VHYFKGWFLI	500
501	DMVAAIPFDL	LIFGSGSEEL	IGLLKTARLL	RLVRVARKLD	RYSEYGAAVL	550
551	FLLMCTFALI	AHWLACIWYA	IGNMEQPHMD	SRIGWLHNLG	DQIGKPYNSS	600
601	GLGGPSIKDK	YVTALYFTFS	SLTSVGFGNV	SPNTNSEKIF	SICVMLIGSL	650
651	MYASIFGNVS	AIIQRLYSGT	ARYHTQMLRV	REFIRFHQIP	NPLRQRLEEY	700
701	FQHAWSYTNG	IDMNAVLKGF	PECLQADICL	HLNRSLLQHC	KPFRGATKGC	750
751	LRALAMKFKT	THAPPGDTLV	HAGDLLTALY	FISRGSIEIL	RGDVVVAILG	800
801	KNDIFGEPLN	LYARPGKSNG	DVRALTYCDL	HKIHRDDLLE	VLDMYPEFSD	850
851	HFWSSLEITF	NLRDTNMIPG	SPGSTELEGG	FSRQRKRKLS	FRRRTDKDTE	900
901	QPGEVSALGP	GRAGAGPSSR	GRPGGPWGES	PSSGPSSPES	SEDEGPGRSS	950
951	SPLRLVPFSS	PRPPGEPPGG	EPLMEDCEKS	SDTCNPLSGA	FSGVSNIFSF	1000
1001	WGDSRGRQYQ	ELPRCPAPTP	SLLNIPLSSP	GRRPRGDVES	RLDALQRQLN	1050
1051	RLETRLSADM	ATVLQLLQRQ	MTLVPPAYSA	VTTPGPGPTS	TSPLLPVSPL	1100
1101	PTLTLDSLSQ	VSQFMACEEL	PPGAPELPQE	GPTRRLSLPG	QLGALTSQPL	1150
1151	HRHGSDPGS					1200



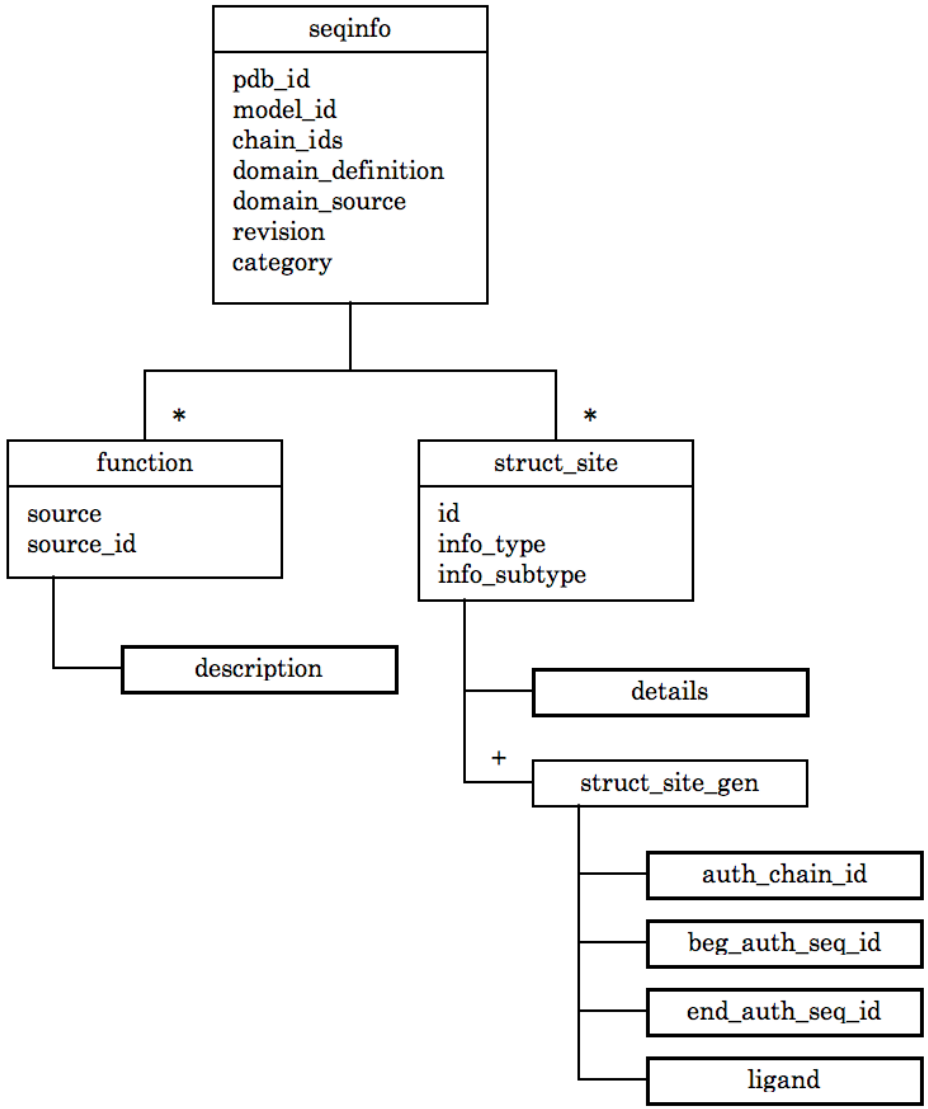
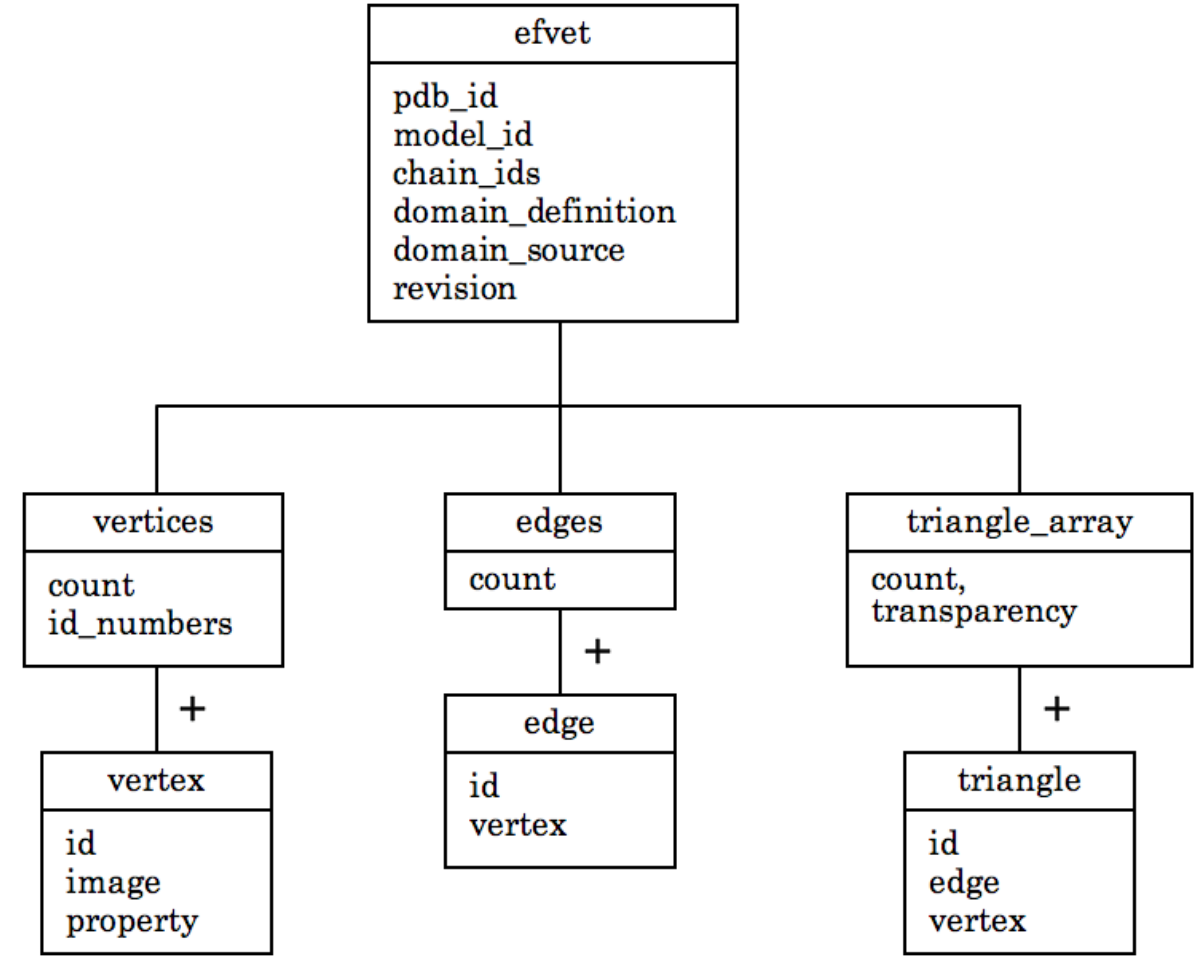


配列を入力してPredictボタンを押すだけ
結果をメールで受け取りたいときのみメールアドレスを入れる

参考資料：eF-site ID

- PID_ModelID-ChainIDが基本
  - 例: 1a1t_3-A
- ModelIDが無い場合は省略する
  - 例: 1tup-C
- 複数のChainを含む場合はアルファベット順で並べる
  - 例: 1tup-ABCEF
- リンクの張り方
  - eF-site IDが分かるとき
    - http://ef-site.hgc.jp/eF-site/servlet/Summary?entry_id=1tup-EF
  - PDB-IDが分かるとき
    - http://ef-site.hgc.jp/eF-site/servlet/Search?pdb=1tup

参考資料：seqinfoファイルの構造



参考資料：efvet-mlの構造