2015年度「バイオインフォマティクス」講義予定表

- 配列アライメント
- データベースサーチ
- タンパク質機能・立体構造予測
- 配列モチーフ・隠れマルコフモデル
- メタゲノム解析（緒方）
- 分子進化・分子系統解析（緒方）
- 遺伝子予測（緒方）
- 機能アノテーション・比較ゲノム（緒方、五斗）
- システムズバイオロジー演習（五斗）

http://goto.kuicr.kyoto-u.ac.jp/lecture/bioinfo.html

配列モチーフ

- 機能ドメイン（機能部位）
  - ***機能的，構造的に重要な部位*** は進化の過程で保存される傾向がある
  - ***進化的に保存された*** ドメイン
- 配列モチーフ
  - 機能ドメイン中の特徴的な ***保存配列パターン***
  - ***マルチプルアライメント*** から抽出
- 配列モチーフの表現方法
  - パターン
  - プロファイル

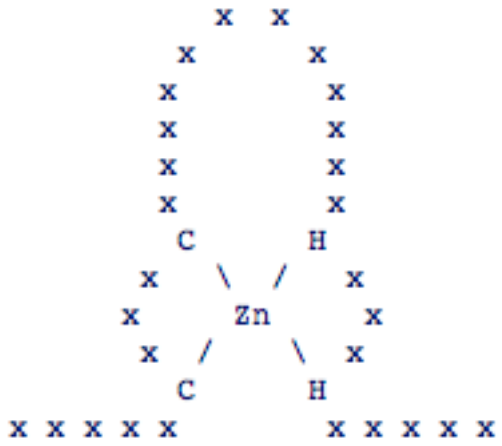
機能ドメインの例

- タンパク質
  - ペプチド鎖切断部位
  - リン酸や糖鎖などの修飾部位
  - シグナル配列
  - DNA結合部位、リガンド結合部位
- RNA
  - スプライス部位：GU/AGルール
  - 翻訳開始点
- DNA
  - 複製開始点
  - プロモーター

ENCODE

機能ドメインの例

- 亜鉛フィンガー DNA 結合部位
  - PROSITE パターンの例
  - C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.



http://www.genome.jp/dbget-bin/www_bget?prosdoc:PDOC00028

ドメインからモチーフの抽出

マルチプルアライメント

Triose phosphate isomerase (TIM) の活性部位

```
E.coli           EEVCARQIDAVLKT----------------QGAAAFEGAVIAYEPVWAIGTGKSATPAQA
H.influenzae     EEVCARQIDAVINA----------------LGVEAFNGAVIAYEPIWAIGTGKSATPAQA
X.fastidiosa     EAILRAQLEPVLSL----------------VGSAGFARAVVAYEPIWAIGTGRTATPDQA
Buchnera         EQVIQRQLNLILKN----------------LGTSAFKNIIIAYEPIWAIGTGYSADPEHV
S.aureus         NDVVGEQVKKAVAG----------------LSEDQLKSVVIAYEPIWAIGTGKSSTSEDA
A.thaliana       MDVVAAQTKAIADR----------------VTN--WSNVVIAYEPVWAIGTGKVASPAQA
H.pylori         FKAVKEFLSEQLEN----------------IDLN-YPNLVVAYEPIWAIGTKKSASLEDI
T.whipplei       LSRFRSVLSHLKAISDKKHSIGYALGSKTHFLDSDQLHMLVAYEPSSAINSGKCANSGDI
                        .                                :*****  **.:  :   .
```

A-Y-E-P-[IVS]-[WS]-A-I-[GN]-[TS]-[GK]
配列パターン

```
A.thaliana       QEVHDELRKWLAKNVSADVAATTRIIYGGSVNGGNCKELGGQADVDGFLVGGASLKP-EF
```

データベースに登録されている配列パターンは

[AVG]-[YLV]-E-P-[LIVMEPKST]-[WYEAS]-[SAL]-[IV]-[GN]-[TEKDVS]-[GKNAD]

http://www.genome.jp/dbget-bin/www_bget?prosite:PS00171

配列パターン

- 保存配列をアミノ酸のパターンとして表現
  - *正規表現*による表現方法
    - 文字列の集合を一つの文字列で表現する方法
  - 例
    - `[AV]-Y-E-P-[LIVM]-W-[SA]-I-G-T-[GK]`
    - `C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H`
  - `x`: 任意のアミノ酸
  - `x(2,4)`: 任意のアミノ酸が２～４個続く
  - `[]`: この中のアミノ酸のどれか
  - `{}`: この中のアミノ酸以外のどれか

- ***見た目に分かりやすいが、アミノ酸の出現頻度情報は失われてしまう***

PROSITE でのパターン表現

```
ID   TIM; PATTERN.
AC   PS00171;
DT   APR-1990 (CREATED); DEC-2004 (DATA UPDATE); JUN-2007 (INFO UPDATE).
DE   Triosephosphate isomerase active site.
PA   [AVG]-[YLV]-E-P-[LIVMEPKST]-[WYEAS]-[SAL]-[IV]-[GN]-[TEKDVS]-[GKNAD].
NR   /RELEASE=53.3,274295;
NR   /TOTAL=236(236); /POSITIVE=236(236); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR   /FALSE_NEG=9; /PARTIAL=4;
CC   /TAXO-RANGE=A?EP?; /MAX-REPEAT=1;
CC   /SITE=3,active_site;
CC   /VERSION=1;
DR   P36204, PGKT_THEMA , T; P36186, TPI1_GIALA , T; P36187, TPI2_GIALA , T;
DR   Q9SKP6, TPIC_ARATH , T; Q9M4S8, TPIC_FRAAN , T; P46225, TPIC_SECCE , T;
DR   P48496, TPIC_SPIOL , T; Q57D01, TPIS1_BRUAB, T; Q8YHF5, TPIS1_BRUME, T;
DR   Q8G0F7, TPIS1_BRUSU, T; Q928I1, TPIS1_LISIN, T; Q71WW9, TPIS1_LISMF, T;
DR   Q8Y4I3, TPIS1_LISMO, T; Q2K869, TPIS1_RHIEC, T; Q98ME7, TPIS1_RHILO, T;
DR   Q92QA1, TPIS1_RHIME, T; Q2YIQ6, TPIS2_BRUA2, T; P0C119, TPIS2_BRUAB, T;
DR   Q8YCV3, TPIS2_BRUME, T; Q8FVH2, TPIS2_BRUSU, T; Q92EU4, TPIS2_LISIN, T;
DR   Q723V9, TPIS2_LISMF, T; Q8YA20, TPIS2_LISMO, T; Q2JZQ2, TPIS2_RHIEC, T;
DR   Q986N6, TPIS2_RHILO, T; Q92NH8, TPIS2_RHIME, T; P96985, TPIS3_RHIEC, T;
DR   P92119, TPIS_AEDTO , T; Q9YBR1, TPIS_AERPE , T; Q8UEY3, TPIS_AGRT5 , T;
DR   Q8YP17, TPIS_ANASP , T; P91895, TPIS_ANOME , T; O66686, TPIS_AQUAE , T;
DR   P48491, TPIS_ARATH , T; O28965, TPIS_ARCFU , T; Q750Y8, TPIS_ASHGO , T;
```

- 機能部位の説明
- パターン
- パターンを持つ配列のID
- 機能部位を持つ配列のID
- 短いパターンは偽陽性(False positive) の問題がある

POSITIVE: 機能部位に当該パターンを持つ配列の数
FALSE_POS: 機能部位を持たないが、パターンを持つ配列の数
FALSE_NEG: 機能部位を持つが、パターンを持たない配列の数

PROSITE でのパターン表現

- 疑陽性が多いパターンは、SKIP-FLAG で区別できるようにしている

Database: PROSITE
Entry: PS00001
LinkDB: PS00001
Original site: PS00001

```
ID   ASN_GLYCOSYLATION; PATTERN.
AC   PS00001;
DT   APR-1990 (CREATED); APR-1990 (DATA UPDATE); APR-1990 (INFO UPDATE).
DE   N-glycosylation site.
PA   N-{P}-[ST]-{P}.
CC   /TAXO-RANGE=??E?V;
CC   /SITE=1,carbohydrate;
CC   /SKIP-FLAG=TRUE;
CC   /VERSION=1;
PR   PRU00498;
DO   PDOC00001;
//
```

頻度情報の導入

Triose phosphate isomerase (TIM) の活性部位

```
E.coli          EEVCARQIDAVLKT----------------QGAAAFEGAVIAYEPVWAIGTGKSATPAQA
H.influenzae    EEVCARQIDAVINA----------------LGVEAFNGAVIAYEPIWAIGTGKSATPAQA
X.fastidiosa    EAILRAQLEPVLSL----------------VGSAGFARAVVAYEPIWAIGTGRTATPDQA
Buchnera        EQVIQRQLNLILKN----------------LGTSAFKNIIIAYEPIWAIGTGVSADPEHV
S.aureus        NDVVGEQVKKAVAG----------------LSEDQLKSVVIAYEPIWAIGTGKSSTSEDA
A.thaliana      MDVVAAQTKAIADR----------------VTN--WSNVVIAYEPVWAIGTGKVASPAQA
H.pylori        FKAVKEFLSEQLEN----------------IDLN-YPNLVVAYEPIWAIGTKKSASLEDI
T.whipplei      LSRFRSVLSHLKAISDKKHSIGYALGSKTHFLDSDQLHMLVAYEPSSAINSGNCANSGDI
                         .                                ::**** **.:  :  .
```

A-Y-E-P-[IVS]-[WS]-A-I-[GN]-[TS]-[GK]

```
X.fastidiosa    QAVHAFIRGEVAKADAR-IADSLPILYGGSVKPDNASELFSQPDVDGGLVGGASLVAEDF
Buchnera        QLIHVFIKNYILKYSSI-NRNDIIIQYGGSINHTNVKKFIEQPDINGLLIGNSSLSAKEF
S.aureus        N   VRQTIADLSSKEVS  IDGALVGGASLK---V
A.thaliana      Q   LRKWLAKNVSADVA  DGFLVGGASLKP-EF
H.pylori        Y   LKQILN--------  DGLLIGSASWELENF
T.whipplei      V   KDIVN-  -----  DGILVGRASLNASDF
                                   :* *:* :*    .
```

V: 2/8
I: 5/8
S: 1/8
+ pseudocount

配列プロファイル

- マルチプルアライメントの各残基位置のアミノ酸出現頻度をカウント
- Pseudocount などを導入して正規化
- 表現方法
  - 位置特異的スコアマトリックス（PSSM: Position Specific Score Matrix または SSSM: Site Specific Score Matrix）
  - 隠れマルコフモデル
  - ブロック

配列プロファイル：位置特異的スコアマトリックス

- 位置 i におけるアミノ酸 j の出現頻度

$$Frq(i,j) = n(i,j)/N$$

  - $n(i,j)$：位置 i においてアミノ酸 j が出現した個数
  - N：アライメントに含まれる配列の本数

- 位置 i におけるアミノ酸 j のスコア

$$pssm(i,j) = \ln \frac{Frq(i,j)}{P(j)}$$

  - $P(j)$：アライメントを構築している配列全体または データベース全体から得られるアミノ酸組成

配列プロファイル：位置特異的スコアマトリックス

PROSITE Profile

```
ID   ACP_DOMAIN; MATRIX.
AC   PS50075;
DT   NOV-1997 (CREATED); NOV-1997 (DATA UPDATE); JUN-2007 (INFO UPDATE).
DE   Acyl carrier protein phosphopantetheine domain profile.
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=71;
MA   /DISJOINT: DEFINITION=PROTECT; N1=6; N2=66;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=2.3; R2=.02281121; TEXT='-LogE';
MA   /CUT_OFF: LEVEL=0; SCORE=271; N_SCORE=8.5; MODE=1; TEXT='!';
MA   /CUT_OFF: LEVEL=-1; SCORE=184; N_SCORE=6.5; MODE=1; TEXT='?';
MA   /DEFAULT: D=-20; I=-20; B1=-80; E1=-80; MI=-105; MD=-105; IM=-105; DM=-105; MM=1; M0=-1;
MA   /I: B1=0; BI=-105; BD=-105;
MA   /M: SY='T'; M=-5,-15,-20,-17,-12,-10,-22,-18,2,-13,-1,0,-13,-6,-10,-13,-5,4,1,-23,-9,-12;
MA   /M: SY='E'; M=-6,-6,-22,-6,9,-13,-21,-9,-11,0,-8,-7,-7,-13,1,1,-4,-3,-8,-24,-10,4;
MA   /M: SY='E'; M=-5,9,-24,11,15,-24,-12,-3,-23,3,-20,-15,6,-9,5,1,4,-2,-19,-29,-16,9;
MA   /M: SY='E'; M=-5,2,-26,4,8,-22,-13,-7,-21,7,-17,-12,0,-13,3,7,-2,-6,-16,-22,-12,5;
MA   /M: SY='L'; M=-6,-27,-19,-30,-23,4,-30,-23,26,-25,28,17,-25,-27,-21,-20,-19,-5,23,-23,-3,-23;
MA   /M: SY='R'; M=-3,-10,-10,-11,2,-16,-19,-11,-13,-1,-8,-7,-8,-17,-1,3,-5,-6,-9,-26,-13,-1;
MA   /M: SY='E'; M=-1,3,-23,4,9,-24,-11,-7,-22,8,-19,-13,2,-11,5,6,2,-2,-17,-26,-15,7;
MA   /M: SY='I'; M=-5,-22,-20,-27,-19,-4,-29,-20,20,-19,13,10,-18,-21,-14,-17,-15,-6,14,-20,-4,-18;
MA   /M: SY='I'; M=-8,-30,-24,-33,-27,8,-29,-26,19,-24,15,9,-28,-27,-23,-21,-22,-10,17,9,4,-25;
MA   /M: SY='A'; M=11,-8,-8,-12,-5,-19,-11,-14,-14,-1,-14,-9,-6,-15,-4,-4,2,-2,-6,-25,-15,-5;
MA   /M: SY='E'; M=-5,10,-26,15,22,-28,-12,-2,-26,6,-21,-16,4,-8,10,0,2,-6,-23,-28,-16,16;
MA   /M: SY='V'; M=-5,-14,-15,-16,-6,-11,-23,-14,4,-11,0,1,-13,-19,-5,-12,-7,-5,6,-24,-8,-6;
MA   /M: SY='L'; M=-2,-24,-21,-26,-19,5,-24,-20,10,-23,22,7,-23,-24,-18,-19,-18,-7,6,-7,0,-18;
MA   /M: SY='G'; M=3,-4,-25,-5,-4,-27,24,-12,-29,-6,-25,-16,1,-12,-4,-8,5,-9,-23,-24,-22,-4;
MA   /M: SY='V'; M=-1,-12,-19,-14,-8,-11,-20,-14,4,-12,0,1,-11,-18,-10,-13,-5,-2,7,-25,-9,-10;
MA   /I: I=-4; MI=0; MD=-15; IM=0;
MA   /M: M=-2,-6,-13,-6,-5,-9,-11,-10,-2,-8,0,-2,-7,-7,-7,-8,-5,-3,-1,-18,-8,-7; D=-3;
```

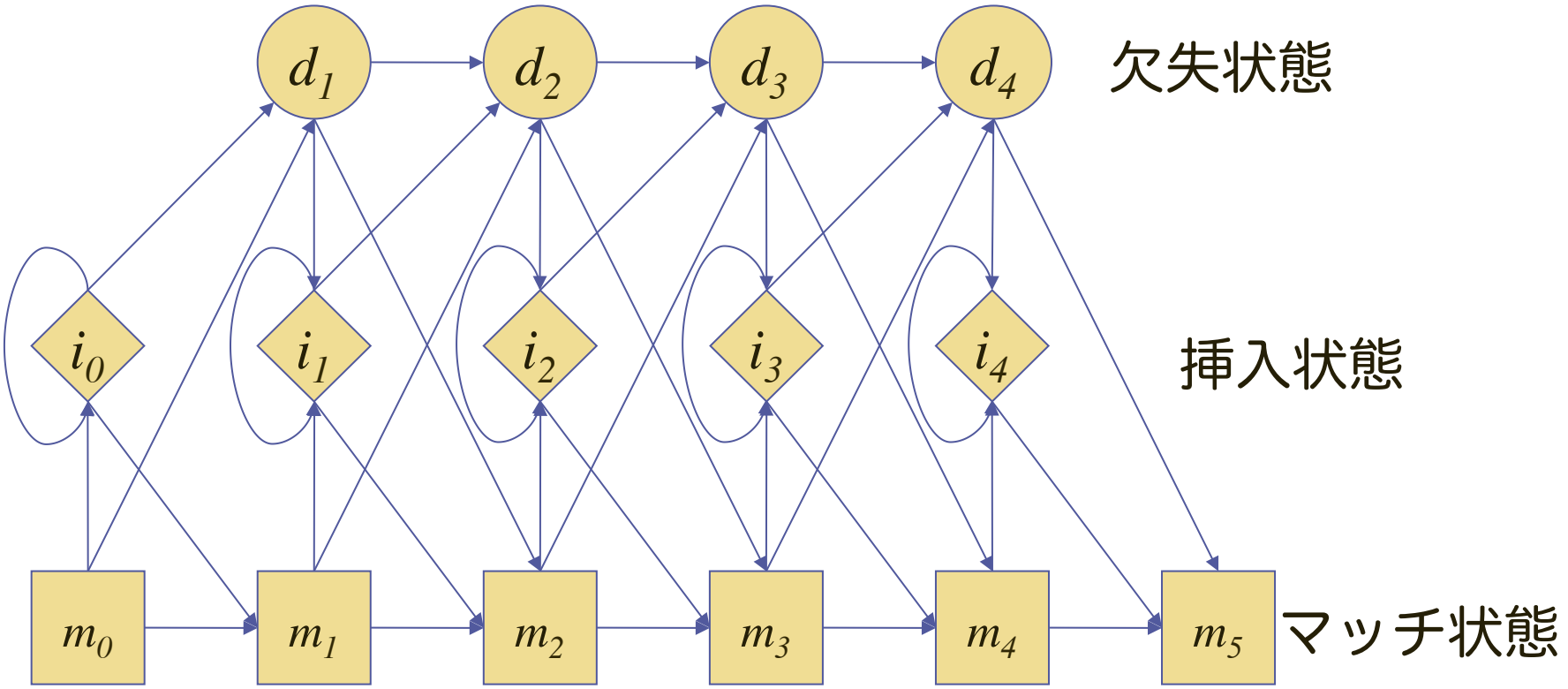
位置 1 2 3 4 :

SY

SY: コンセンサス　M: マトリックス

配列プロファイル：隠れマルコフモデル

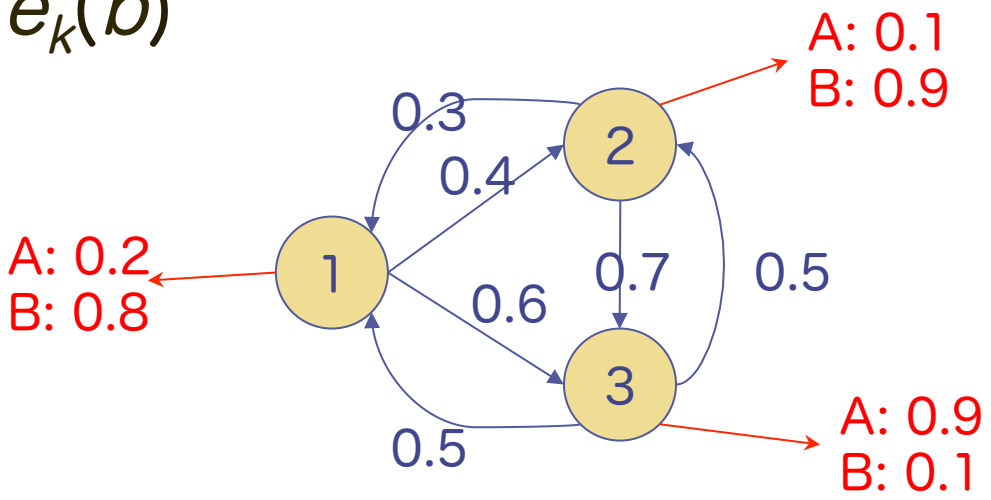
隠れマルコフモデル（HMM: Hidden Markov Model）
　1次のマルコフモデル＋隠れ状態確率
　プロファイルの長さを決めてモデル化



配列プロファイル：隠れマルコフモデル

隠れマルコフモデル ≒ 有限オートマトン＋確率

- 出力記号集合 Σ
- 状態集合 $S=\{1,2,\cdots,n\}$
- 遷移確率（状態 k → 状態 l）a_{kl}
- 出力確率 $e_k(b)$



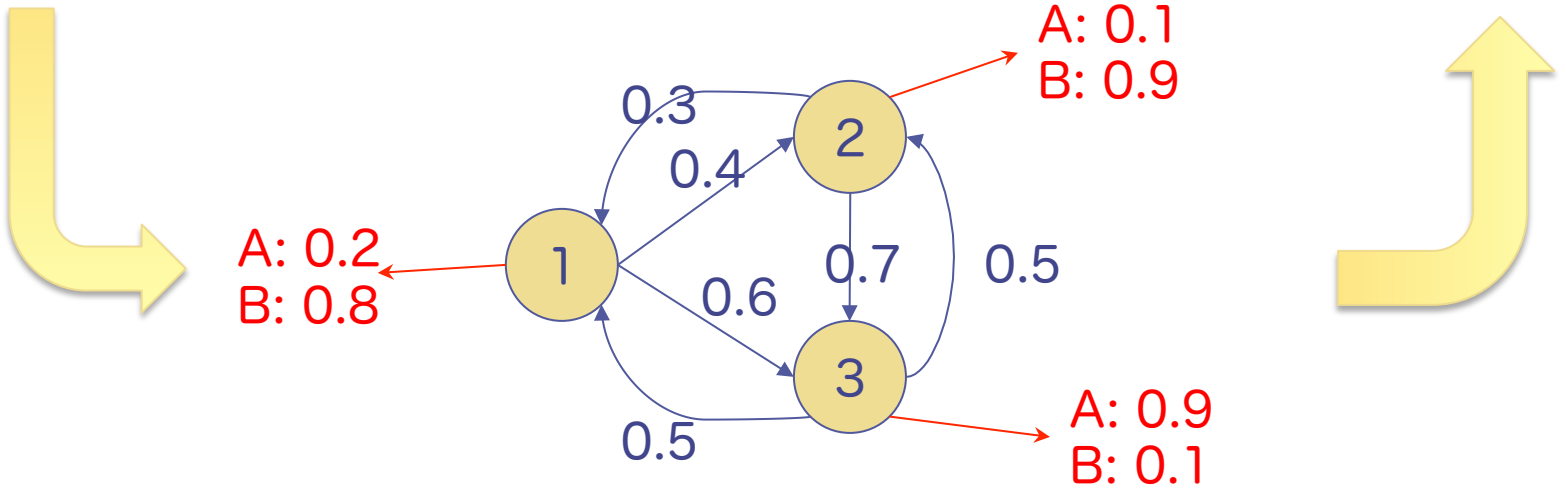
配列プロファイル：隠れマルコフモデル

隠れマルコフモデルのアルゴリズム

- Viterbi アルゴリズム
- 出力記号列から状態列を推定
- 構文解析

BABBABB

2312312

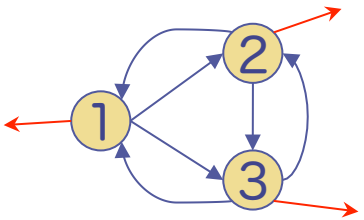


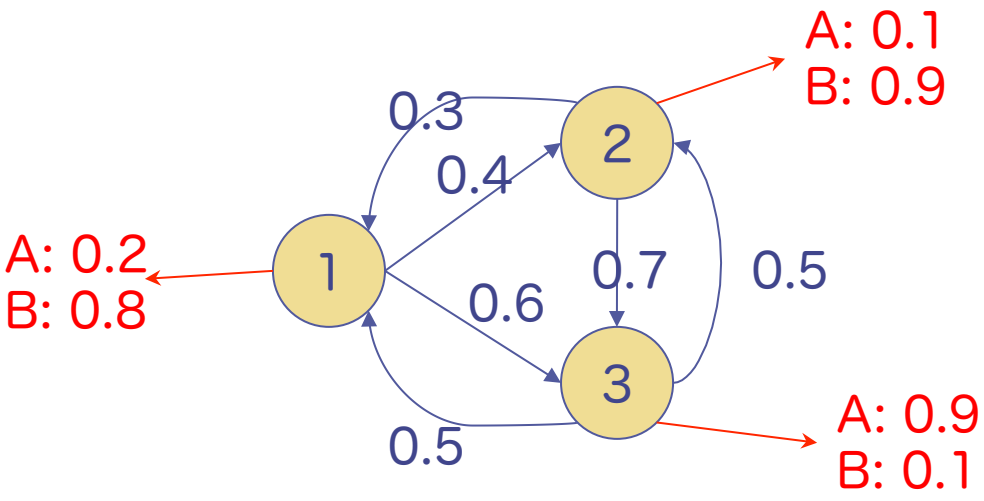
配列プロファイル：隠れマルコフモデル

隠れマルコフモデルのアルゴリズム

- Baum-Welch アルゴリズム
- EM (Expectation-Maximization) アルゴリズム
- 出力記号列からパラメータを推定
- 学習

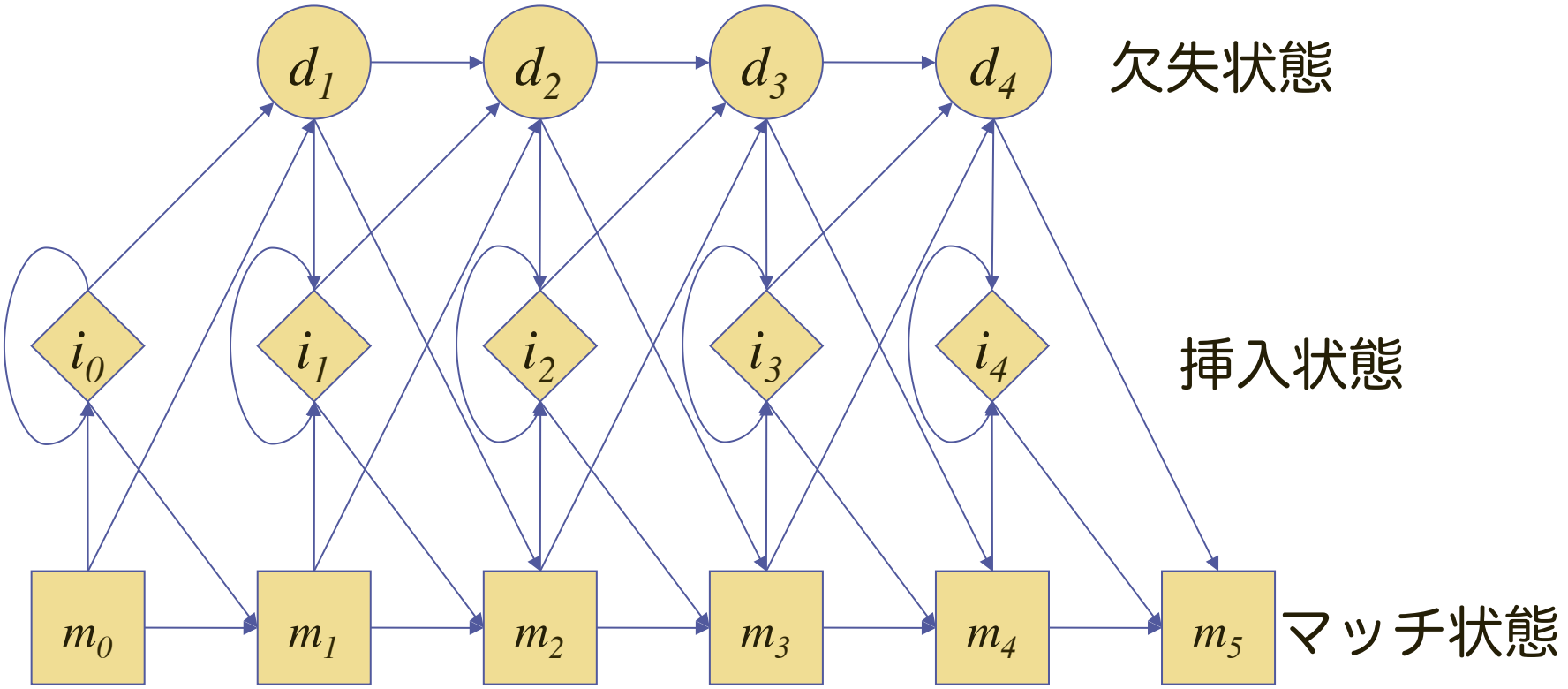
BABBABB
BBAABBABB
ABBABBB
BABAABB





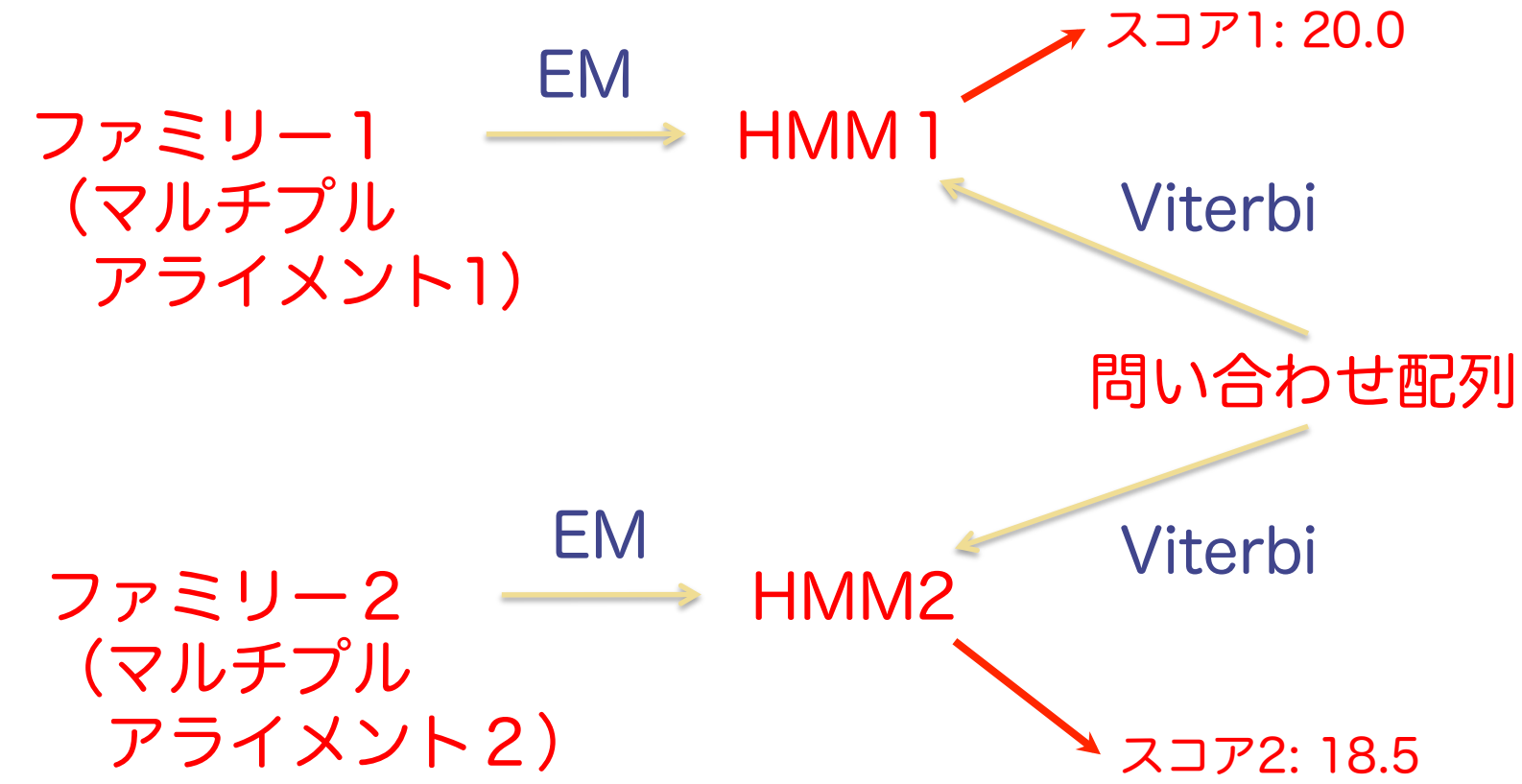
配列プロファイル：隠れマルコフモデル

プロファイル HMM
　ファミリーごとにプロファイルの長さを決めてモデル化



配列プロファイル：隠れマルコフモデル

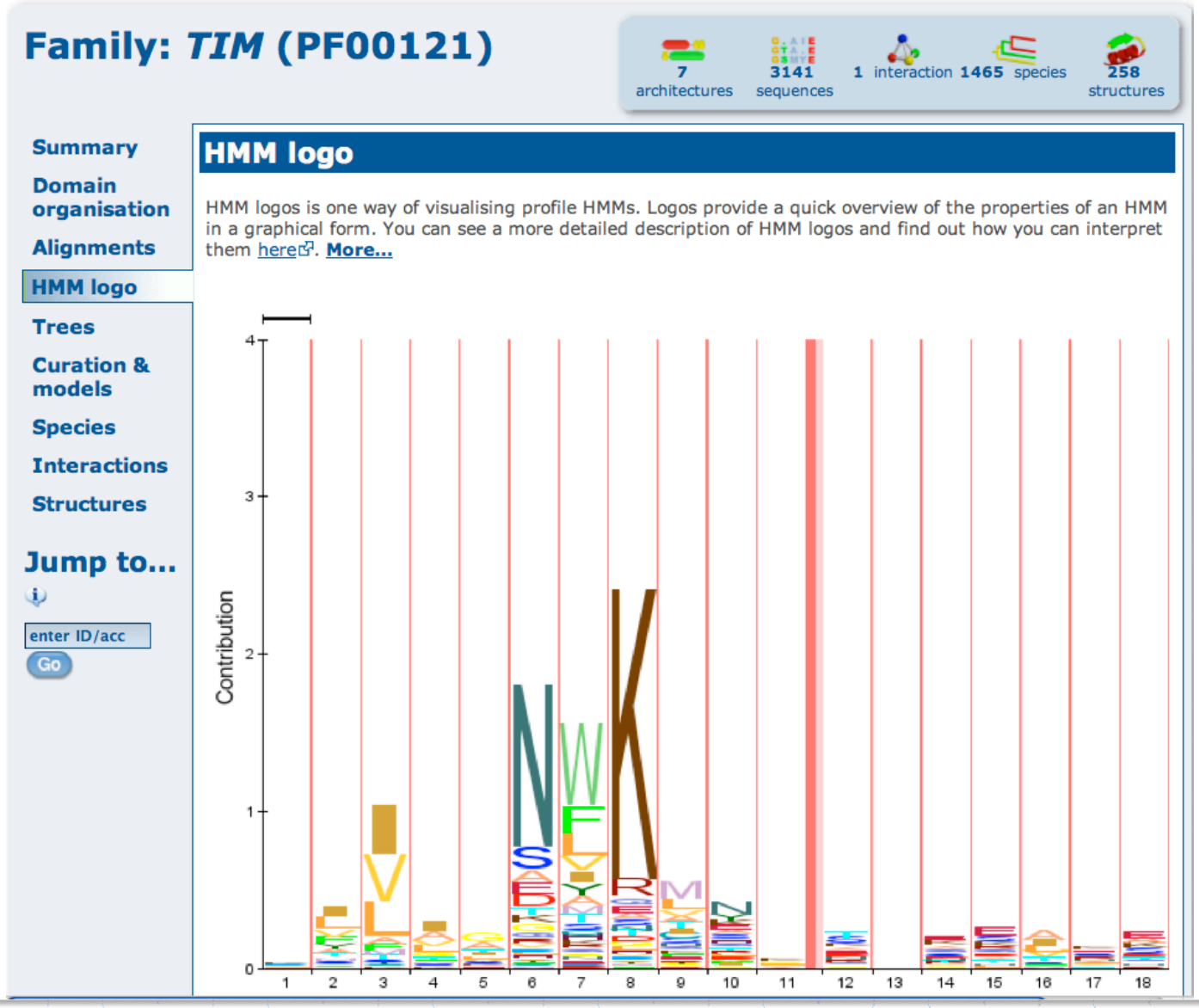
隠れマルコフモデルによるプロファイル作成とモチーフ検索



配列モチーフデータベース

- 配列モチーフの表現方法による分類
  - 配列パターン
    - PROSITE Pattern
  - 配列プロファイル：位置特異的スコアマトリックス
    - PROSITE Profile, NCBI-CDD
  - 配列プロファイル：隠れマルコフモデル
    - PFAM
  - その他
    - BLOCKS, PRODOM
  - 統合
    - InterPro(, NCBI-CDD)

PFAM での配列プロファイル表現



InterPro による配列モチーフデータベースの統合

InterPro: IPR000652 Triosephosphate isomerase

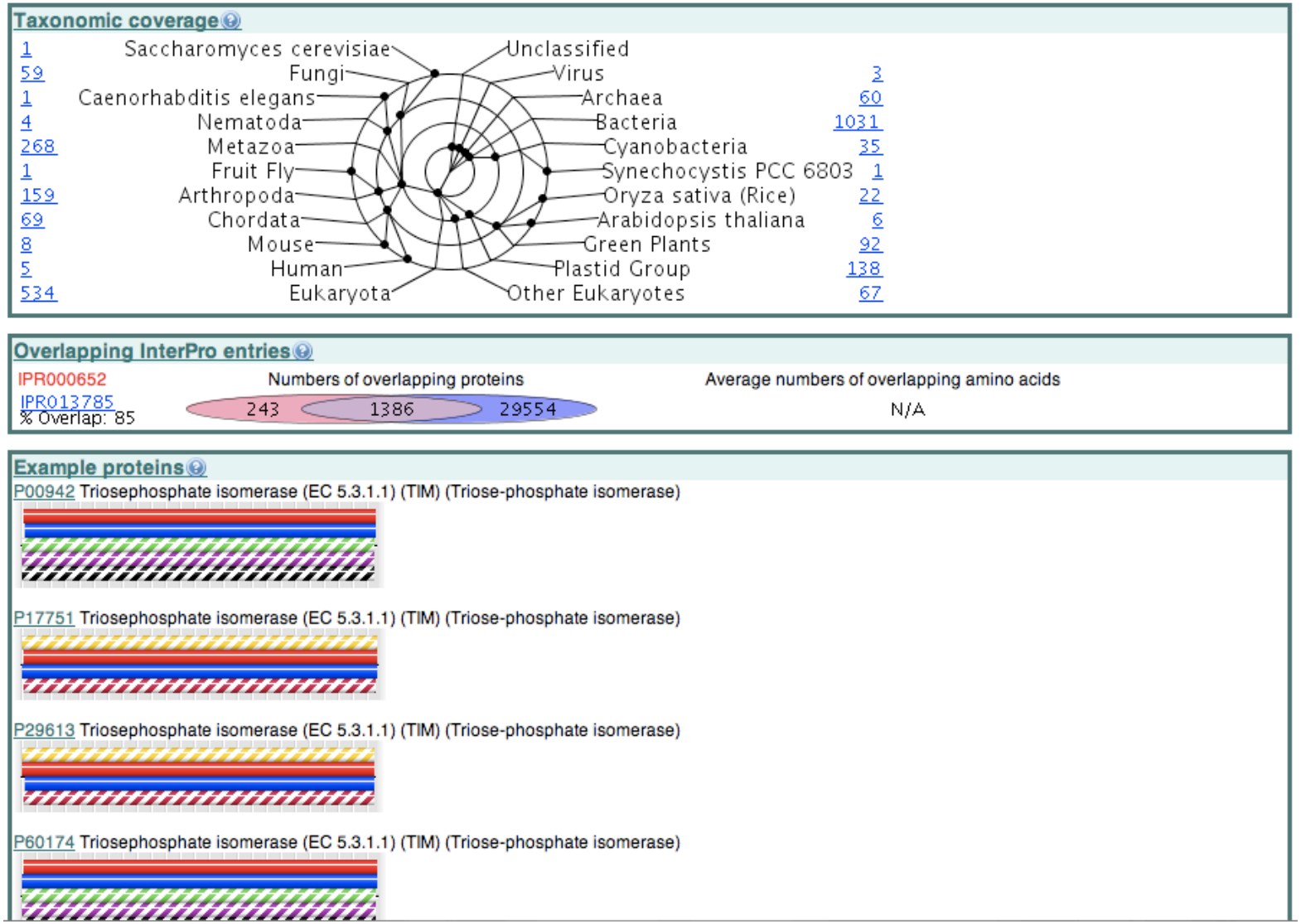
Matches	**Overview:** sorted by AC, sorted by name, of known structure, proteins with splice variants **Detailed:** sorted by AC, sorted by name, of known structure proteins with splice variants **Table:** For all matching proteins, of known structure Architectures
Accession	IPR000652 Triophos_ismrse Matches: 1629 proteins
Type	Family

Signatures

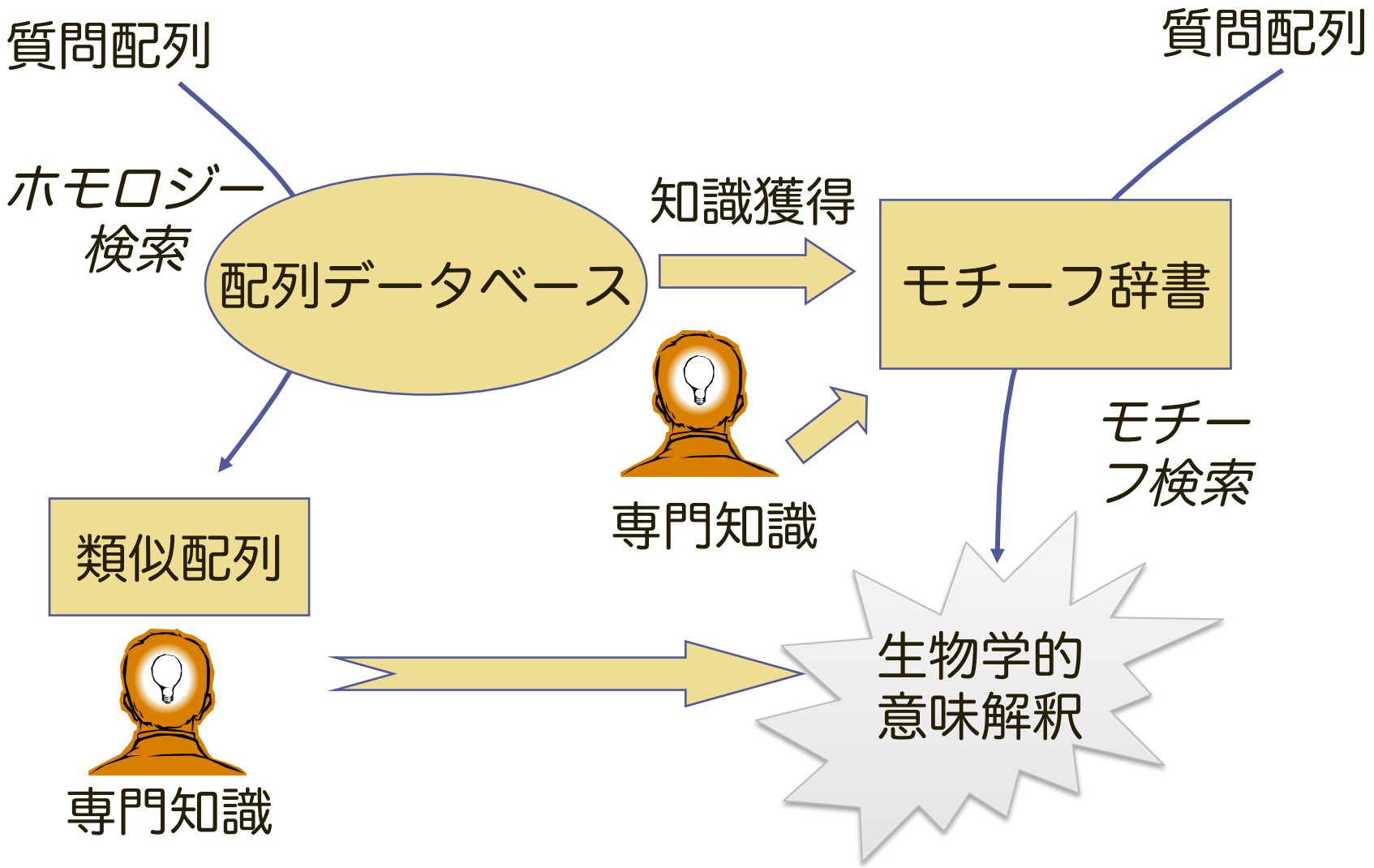
Database	ID	Name	Proteins
ProDom	PD001005	Triophos_ismrse	1442
Pfam	PF00121	TIM	1282
PROSITE pattern	PS00171	TIM	960
PANTHER	PTHR21139	Triophos_ismrse	1241
SuperFamily	SSF51351	Triophos_ismrse	1452
TIGRFAMs	TIGR00419	tim	909

InterPro Relationships	
Contains	IPR013785 Aldolase-type TIM barrel
GO Term annotation	
Process	GO:0008152 metabolic process
Function	GO:0004807 triose-phosphate isomerase activity
InterPro annotation	
Abstract	Triosephosphate isomerase (EC:5.3.1.1) (TIM) [1] is the glycolytic enzyme that catalyzes the reversible interconversion of glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. TIM plays an important role in several metabolic pathways and is essential for efficient energy production. It is a dimer of identical subunits, each of which is made up of about 250 amino-acid residues. A glutamic acid residue is involved in the catalytic mechanism [2]. The sequence around the active site residue is perfectly conserved in all known TIM's. Deficiencies in TIM are associated with haemolytic anaemia coupled with a progressive, severe neurological disorder [3].
Structural links	CATH: 3.20.20.70 SCOP: c.1.1.1 , c.1.31.1 PDB - click here
Database links	PANDIT: PF00121 PROSITE doc: PDOC00155 Blocks: IPB000652 MSDsite: PS00171 Pfam Clan: CL0036.15

InterPro による配列モチーフデータベースの統合



ホモロジー検索とモチーフ検索



関連ウェブサイト

- モチーフデータベース
  - キーワード検索
    - http://www.genome.jp/dbget-bin/www_bfind?motifdic
    - http://www.ebi.ac.uk/interpro/
  - 塩基配列
    - 転写因子結合サイトなど：EPD, TransFac, JASPER
- モチーフ検索
  - http://www.genome.jp/tools/motif/
    - 質問配列が既知のモチーフを持つか
    - 質問モチーフを持つ配列を配列データベースから探す
    - マルチプルアライメントからプロファイルを作る
  - PSI-BLAST: Position Specific Iterated BLAST
    - http://www.ncbi.nlm.nih.gov/BLAST/