タンパク質のアイデンティティをモチーフにより決める試み
Trial to Determine Identity of Proteins by Motifs.

C1

瀬戸保彦，池内義典，川北栄継(1)，西川建(1)，金久實(2)
（財）蛋白質研究奨励会 (1)蛋白工学研 (2)京大化研
Seto,Y., Ikeuchi,Y., Kawakita,S.(1), Nishikawa,K.(1), and Kanehisa M.(2)
Protein Res.Found., (1)Protein Eng.Res.Inst., and (2)Kyoto Univ.

We propose a semi-automatic method to classify protein seqeunces into superfamilies. The method is a combination of the global sequence homology search and the local sequence identifier. The former is performed by a pairwise homology search method developed by Nishikawa et al(2), which jointed five different measures to improve sensitivity for detecting seqeunce similarity. The latter is done by applying the fragment peptide library(FRAP) compiled by Seto et al(3). We evaluate our method by applying it to 89 superfamilies in the PIR database.
Out of 89 superfamilies, 65 superfamilies are classified by the combination method. We find that 20 superfamilies are multi-superfamilies:Different superfamilies belong to a same group by our criteria. We find difficulty to classify 4 superfamilies.

1.序

我々はタンパク質の特定のグループに特徴的な断片配列、モチーフの収集について第2回ゲノム情報ワークショップで報告した。蓄積した断片配列は現在12198件に及ぶ。自然界におけるタンパク質の種類は数千種と見積もる事もできる。つまり殆どのタンパク質に固有のモチーフがそろそろ揃ってきたと考えられる。また長年の間に収集してきたタンパク質の配列は13000件に及ぶ。これらの配列の分類を始める時期が来たと考え、数年前よりその方法論の検討を始めた。

分類する事は同一性の規範、アイデンティティを決める事と言える。ふたつのタンパク質が同族か否かは物理的、化学的、生物学的な性質等から総合的に判断される。ゲノムプロジェクトの進行に伴い生産される膨大な情報はまず配列であり、タンパク質としてのその他の性質は不明である。

タンパク質の分類はDayhoff,Barkerらにより進化的視点から系統的になされてきた(1)。しかし、分類の難しさから、配列データの増加に比較して分類作業は順調には進んでいない。また長い配列や多機能性タンパク質の配列が増加してきたため、分類も複雑になっている。本報告において、我々はできる限り配列の情報だけから自動的に分類する事を試みた。その方法は全体的な類似性と局所の同一性を組み合わせるものである。分類分けのされているPIRの配列データベースを利用して、多くの生物種から得られたタンパク質を含むスーパーファミリーをモデルとしてその方法論を評価した。

2.方法

分類は2段階で行う。第1段階で西川らの開発した5パラメータ法(2)により、配列全体で比較して、タンパク質を類似した配列を持つグループに分ける。そのグループで弱い類似性しか示さず同族判定のむずかしい場合、第2段階として瀬戸-金久らの断片ペプチドライブラリー(FLAP)(3)のモチーフを共通に持つタンパク質を同族とする。

5パラメータ法は5個のスコアを利用してタンパク質の類似性を評価する。まず比較するタンパク質の配列をGoad-Kanehisa法により、整列表記する。進化的に変わり易いアミノ酸であるかどうかを考慮し、つまりDayhoff 等のPAM行列数値を使い比較してスコア1を得る。窪田らの方法により、アミノ酸の物理化学的な性質を考慮してスコア3を得る。ここに得たスコア1,3について配列の長さを考慮したスコア2,4をそれぞれ得る。その整列配列で、アミノ酸残基を直接比較して一致度の程度によりスコア5を得る。

西川はこれら5個のスコアについて同族のタンパク質になるための域値を立体構造データバンクPDBのデータを利用して決めた。ふたつのタンパク質を比較して域値を満たすスコアがN個のものをランクNの類似度とする。ランク5の類似度のタンパク質は同族として大体間違いない。ランクを4にする事により、殆どの同族タンパク質を検索できる。しかし同族でないタンパク質(ノイズ)が類似配列として検索される。従って、検索で拾われたこれらのノイズが実際に同族でない事を別の手法で判断する必要がある。

FLAPにあるモチーフはタンパク質を特徴づける短い配列である。そこで、第2段階とし

て ランク4,5の関係にあるタンパク質のグループの中で、そのグループに特徴的なモチーフを持たないタンパク質は同族でないとした。

配列データベースNBRF/PIRでは進化的に同じ起源から生成したと考えられるタンパク質を同じ族にまとめて、スーパーファミリーと呼んでいる。Rel.25のデータにおいて総エントリー約7000件、スーパーファミリー約2000件である。PIRのスーパーファミリー89個 を選 んで、 この2つの方法の組み合わせによる再分類を試みた。これらのスーパーファミリーは次の15種に分けられた生物の中、どれか3種以上の生物から得たタンパク 質 を 含む。従って配列の違いが比較的大きいタンパク質のグループと言える。

Pr:prokaryota Fu:fungi Li:lichenes Pl:plantae Pz:protozoa
Ne:nematoda Mo:mollusca Ar:arthropoda Ag:agnatha Ch:chondrichthyes
Os:osteichthyes Am:amphibia Re:reptilia Av:aves Ma:mammalia

各スーパーファミリーについて上のいづれかの生物から得たタンパク質を一つづつ撰びそ れ をプローブとしてPIR全データに対して類似配列検索を行った。例えばSUPERFAMILY NUMBER=958のACh receptorはそのスーパーファミリーにENTRY=10件のタンパク質がある。ソースとなった生物ごとのタンパク質のエントリー数はaves=2, chondrichthyes=4, mammalia=4で、このなかからACCHG1(Av),ACRYA1(Ch),ACHUA1(Ma)のコードを持つタンパク 質がAch receptor族のプローブとして検索、分類の出発点となった。個々のスーパーファミリーのプローブの記載は省く。

3.結果と考察

スーパーファミリー89個のタンパク質を対象にして5パラメータ法で検索した結果を 表に示す。PROTEIN欄はスーパーファミリーの名前、ENTRY欄はそのスーパーファミリーに属するタンパク質の総数、MISSED欄はプローブタンパク質によって検索できなかった そのスー パーファミリーのタンパク質数、NOISE欄はプローブで検索されたそのスーパーファミリー以外のタンパク質総数を示す。MISSED欄が0のスーパーファミリーはENTRYの全てのタ ンパク質が検索されている。NOISE欄のmanyはスーパーファミリー以外のタンパク質が20以上数多く検索された事を意味する。MOTIF欄の配列の間の数字はふたつのモチーフ の間にどれだけのアミノ酸残基があるかを特定の配列で計算したもので、同族であってもタンパク質により異なる事が多い。モチーフ間の距離の目安として記載してある。

対象となったスーパーファミリーは4グループに分ける事ができる。

グループI:同族タンパク質を5パラメータ法で全て検索できる。MISSED欄が0でないホルモン類とhistone H1は検索漏れがある。しかしタンパク質として考えた場合、それらが短い部分配列であることから分類の対象からははずす方が適当である。NOISE欄に示すようにスーパーファミリー以外のタンパク質が拾われるが、グループIにおいてはMOTIF欄に示すモチーフ配列がこれらのタンパク質にない事からノイズを落とす事ができる。

グループII:プローブ配列で検索されるノイズは少ないが、検索できない同族タンパク質がある。IIのノイズはモチーフの有無で除く事ができる。しかし、検索漏れの同族タンパク質を拾うためには5パラメータ法でランクを落とし更に弱い類似検索をする必要が ある。検索にもれたものもその多くはモチーフを持っている。しかしこのIIのタンパク質には配列だけから同族とするのは困難なものがPIRにある事を示す。

グループIII:PIRでは別のスーパーファミリーに属しているタンパク質が我々の規範 からは同族となるものを含むグループである。類似配列の検索で整列したときの同一アミノ酸が10から20パーセントという弱い類似度の場合、局所に同一のモチーフを持つか否かで同族にするかどうかを決める。従って、特定のグループのタンパク質に特徴的であるべきモチーフの質が分類の信頼度を決める。

グループIV:このグループの配列の分散度は大きく、従って弱い類似性を持った多く のタンパク質が検索される。同族性の判断が困難なグループである。これらのスーパーファミリーのモチーフがFRAPライブラリーにあるが、いづれもスーパーファミリー内の特定な族のモチーフであると考えられる。モチーフをスーパーファミリー全体に一般化するにはより洗練されたパターンにする事が必要である。

第一段階ではプローブから出発してペアワイズな配列の整列表記で一つ一つのタンパク質 の同族性を判断した。グループIIのcytochrome c,globinのように多くの生物種の配列がある場合、プローブの選び方によって類似配列検索で拾えないタンパク質が多くなる。つ まり、 あ る 族 内の配列の分散が大きいとABC問題が生ずる。A,B,Cがあったとき、A=B,B=C,A=/Cとなったときに変更するのか。A,Cがあり、AはCと同族でないがBが現 れて、 A,CをBがつないでA,B,Cを同族にする事はないのか。現在この問題は人間が判断している。cytochrome c,globinの場合、NOISEが少ない事から分かるように類似配列の検索にミスは少ない。プローブを増す事で検索漏れはなくなる。

SUPERFAMILIES CLASSIFIED BY WEAK HOMOLOGY SEARCH AND MOTIFS FOR EACH SUPERFAMILY

I SUPERFAMILIES EASILY CLASSIFIED fig12a

SUPERFAMILY NUMBER PROTEIN	ENTRY	MISSED	NOISE	MOTIF
958:ACh receptor:	10	0	many	VLYNRADG(20)WLPPAI(11)FPFDQQN(121)PCVLIS(4)LVFYLP (152)WKYVAM
27:alcohol dehydrogenase:	12	0	4	GVCHTD(10)GHEGAG(144)RIIGVD(36)NGGVD
585:antithrombin:	20	0	4	YFKGKW(152)EEGTEA(10)NRPFLF(11)FMGRV
512:argininosuccinate lyase:	3	0	24	LWGGRF(95)TGRSRN(40)MPQKKNLQ
436:ATPase alpha/beta:	16	0	29	AEYFRD(60)EGSIT(37)GIRPAV
441:ATPase 6:	12	0	many	PAGTPTPL(17)ALGVRL(5)AGHLL
440:ATPase lipid binding:	11	0	8	GAGIG(11)ARNPS(0)AILGFA
476:citrate synthase:	3	0	5	HADHE(31)GPLHGLAN
643:corticoliberin:	6	0	0	PPISIDL
661:corticotropin:	17	(2)	2	EFRWG(7)YGGFM
238:creatine kinase M:	7	0	9	PDLIINNN(164)RPLP(16)WPAGRGI
905:crystallin alpha:	55	0	7	LLRVK(1)FKPEDL(1)VKDLD(13)DDHGYISR(26)LSIDGVLT
906:crystallin beta,gamma:	19	0	0	EKPGRG(9)YEQPR(0)YRGIQY
34:cytochrome b5:	9	0	0	VYDLT(5)HPGG(7)AGIDAT (46)MLLTDRN(18)BLPWF(13)IWEAPA
99:cytochrome c oxidase 2:	12	0	many	QWYWSY(57)VPSLGVK(1)DAIPGRL(14)GQCSEICG
100:cytochrome c oxidase 3:	9	0	25	FHAYPHS(15)WPPVGI(9)PLLNT(66)GLRGLKV(32)YWHFVD
1223:elongation factor:	7	0	9	RGITI(10)HIPGH(5)NMITG(26)QTREH
683:gastrin:	12	(4)	1	GWMDFG
678:gastrin releasing peptide:	4	(2)	0	WAVGHLM
69:Glu dehydrogenase:	5	0	11	GYRVQ(8)KGGLR(66)DVPAP
676:glucagon:	25	(6)	1	GTFTSDYSKY(35)HADGSF/NSDAV(15)KYLES
51:glyceraldehydephosphate dehydrogenase:	13	0	0	GFGRIGR(135)ASCTTN(154)SWYDNE
172:glycogen phosphorylase:	4	0	10	FSSDRTI
645:gonadoliberin:	5	(3)	0	AEHWSYGL
771:histone H1:	19	(1)	12	KERKGSS,KPAAK,GASGSF
772:histone H2A:	25	0	1	AGLQFPVGR,LELAGNAA,VLLPKK
773:histone H2B:	24	0	1	EGTKAV,KVLKQVHPD,MNSFIND
774:histone H3:	19	0	1	
775:histone H4:	10	0	1	GKGGKG(4)GAKRH
776:HMG protein:	5	0	4	RRSARL(11)KPKKA(12-35)AENGDAK
1015:hsp70 protein:	5	0	24	GIDLGTT(56)KRLIGR(424)KITITND(35)LESYA(41)AEKEEY(20)GGMPG
1016:hsp90 protein:	8	0	30	EEYGEFYK(47)KLYVRRVF
682:insulin:	43	0	0	HLCG(10)CGDRGFY(38)GIVEQCC,CC(3)C(7)YC
31:lactate dehdyrogenase:	13	0	10	SNPVDIL(21)LHTARF(19)LGEHGD
955:lectin:	20	0	10	CPIKWHGCYIHNIGC(29)WIGLTD(6)WKWVDGT(34)GQWND(6)CYYVC
354:lysozyme:	26	0	0	DGYGGI(3)NWVCL(21)DYGIFQIN(34)CAKKI
1004:metallothionein:	18	0	19	CKCTSCKK(0)SCCSCC(5)CAQGC
83:NADH dehydrogenase 1:	8	0	many	ADGVKL(41)LADL(38)AQTISYE(53)PFDLPE(1)ESELVA(76)RYDQLM
85:NADH dehydrogenase 3:	7	0	many	EKLSKYECGFDP(32)FHA
86:NADH dehydrogenase 4L:	7	0	many	LLCLEG(31)AALEAAIGL
89:NADH dehydrogenase 6:	7	0	many	FLIYLGGM
675:pancreatic hormone:	13	0	0	YPGDDA(19)TRQRY
398:papain:	9	0	1	CGSCW(37)GCQGG(80)YKSGV(28)YWIVKNSWG
401:pepsin:	16	0	5	NPARKY(25)SIGTP(7)FDTGS(141)SKYYTGSL
235:phosphoglycerate kinase:	5	0	4	RVDFNVP(33)SHLGRP(95)DAPKIAKR
215:pyruvate kinase:	4	0	20	NFSHG(27)LDTKGP(340)RGVFP
854:ribosomal protein L14:	14	0	3	ADNSGAR(54)FEDNAAV
804:ribosomal protein S8:	9	0	1	PSSKV(10)EGFIE(36)NPGLRI(20)TSKGLM
810:ribosomal protein S11:	10	0	1	ASPKKT(65)KLPGPG(2)AALRAL(11)DVTPVP
130:superoxide dismutase Cu/Zn:	10	0	0	HGFHVH(30)HVGDLG(38)DDLGKGG(8)TGNAG
131:superoxide dismutase Fe/Mn:	4	0	4	ALEP(9)HHQKHHQ(98)GSGWA(36)WEHAY
209:thymidine kinase:	9	0	1	DRHPIA,DTLPS,GLLGIGNT,EPMIYW,DRHPIA
632:transforming protein myb:	10	0	4	WTREED(29)QCLHRW(9)KKGPWTKEE(31)CRKRWHN(36)LPGRTDN
520:triosephosphate isomerase:	8	0	0	GHSERR(68)NAIGTG
584:trypsin inhibitor(PSTI):	14	0	0	CTKIYDCKIC,YCGTDG(4)NECKLC
915:tubulin:	15	0	3	GGGTGSG(32)VVEPYN(10)DNEALYDIC(193)NNYKGEGM

II SUPERFAMILIES WITH MISSING COMPONENTS

SUPERFAMILY NUMBER PROTEIN	ENTRY	MISSED	NOISE	MOTIF
- 443:ATPase 8:	5	1	0	MPQLN(43)PWKWPWT
- 1:cytochrome c:	129	39	0	CAQCHT(5)GPNLSG(42)YIPGTK(12)DLIAYLK
- 15:cytochrome p450:	22	1	13	GHLLQ(16)YGDVF(15)DTIKEAL(216)DLPGAG(49)PYTDAV (1)HEIQRE(50)PGIFLD(13)FGIGIRVCLG(5)MELFLF
- 765:globin:	429	86	2	LSDLHA(4)VDPVNFK
- 983:lactoglobulin:	20	8	1	ENGEC(20)LDTDYD
-1013:osteocalcin:	8	1	0	REVCELN(2)CDELAD
- 292:phospholipase A2:	44	2	7	YGCYCG(13)CCQIHDC
- 636:transforming protein ras:	34	1	5	GIGLKGKS(42)DTAGQE(52)GNKCDL
- 394:trypsin:	64	4	3	KPCGGSLI(7)VLTAAHC(87)ASGWG(54)GDSGGP(13)GIVSWG

III MULTIPLE SUPERFAMILIES ARE CLASSIFIED INTO A SAME FROUP

SUPERFAMILY NUMBER PROTEIN	ENTRY	MISSED	NOISE	MOTIF
+ 917:actin:	17	0	5	AGDDLTD,CVVTT,YCILSS
+ 194:Asp aminotransferase:	7	0	2	HNPTG(26)AYQGF(31)ERVGA(25)YSKPP
+ 921:calmodulin:	77	2	8	(D/E)X4GX5(D/E)
+ 578:carbamoylphosphate synthetase:	3	0	3	TAQNEG(231)RPSYVL(65)GVHTGD(49)VIECN(78)STGEV
+ 10:cytochrome b:	18	0	many	LAMHY(23)RANGAS(25)THIGVI(11)GYVLPW(30)TLTRF(91)PEWYF(11)RKLLGVL
+ 98:cytochrome c oxidase 1:	11	0	many	RNKDIG(44)TAHGLIM(25)DMAFPR(214)IAVPTG(59)SSILGA
+ 73:dihydrofolate reductase:	16	0	3	DGHLPW(27)MGRKTW(7)RPLKDRIN(68)EPEGDT
+ 247:DNA dependent RNA polymerase:	6	0	many	ECIGHNG(250)KGKEGR(6)GKRVDF(100)NADFDGD(10)EAQAEAR (330)DTAVKT(3)GYLKRRLV(570)SFEET(20)ENVILG
+ 18:ferredoxin:	78	47	1	CIXCIXCXXXCP(33)SRLXCQI
+1019:homeo protein:	6	0	3	ELEKE(36)KIWFQNRR
+ 84:NADH dehydrogenase 2:	9	8	many	PFHFWLP(1)AMEGP(48)GGLNQT(6)AYSSI(25)LMIANGL
+ 87:NADH dehydrogenase 4:	10	0	many	WLPKAHVE(64)IAYSSIS
+ 88:NADH dehydrogenase 5:	10	0	many	PFHFWLP(67)LAYSSI
+ 313:protein kinase:	91	4	many	VALYDY(77)FLVRES(102)VAVKTLK(90)IHRDL(16)DFGLAR (17)KWTAPE(10)SDVWA
+ 843:ribosomal protein L12:	13	0	2	DMGFGLF
+1020:RNA dependent RNA polymerase:	43	8	33	VLDLKDAF(33)LPQGKKISP(25)YMDDLLI(117)PDFTKKFT GVKPRG(6)MQMDVT,RIVIDYR(61)RMPFGLK
+ 142:thymidylate synthase:	13	3	7	MGKKTW(35)DRTKYGT(31)ELLWFL(15)IWDEWA(94)
+ 614:transforming protein myc:	33	0	10	DPSVVFP(60)EIDVV(132)PKVVIL
+ 582:trypsin inhibitor(BPI):	21	8	1	TGPCKA(16)FXYGGCXGN(11)CXRTCGG
+ 768:ubiquitin:	13	0	0	QQRLIF(1)GKKLSDYKI

IV SUPERFAMILIES DIFFICULT TO BE CLASSIFIED

SUPERFAMILY NUMBER PROTEIN	ENTRY	MISSED	NOISE	MOTIF
# 752:Ig:	267	—	1	
# 757:Ig:	69	—	many	
# 908:keratin:	25	1	many	
# 917:myosin H:	5	8	many	

Goad-Kanehisa法はローカル類似配列、弱い類似配列も強力に検索する。このことは一方で、ローカルの程度を個々に人間が判断する必要があった。またアミノ酸組成や並びに片寄りがある場合や繰り返し配列を持つ場合にはこの類似性のスコアは適当な同族性の判断にはならない。例えばLeuが多い場合それがLeu rich regionにおいてどの様なパターンで並んでいようと類似性は高くなる。GlyGlyProの繰り返しがあると、GlyGlyLysの繰り返し配列と整列させると66パーセントは同じになる。このような場合も自動化を目指す場合には判断の論理をプログラムする必要があり、今回は人間が判断した。

タンパク質の同一性を決める上でモチーフが鍵になると我々は考えている。あるタンパク質がモチーフを持つかどうかの判断は今回は人間が行った。この判断をコンピュータにさせる事が次の段階であるが、いくつかの解決すべき問題がある。モチーフは通常複数個ある。同族タンパク質がその全てを持つとは限らない。次の図に示すProtein kinaseのスーパーファミリーでは、VALYDY或いはFLVRESのモチーフしかもたない配列が検索され、それらはdomain SH3,domain SH2として知られる。IHRDL,DFGLAR,KWTAPEはkinase domainに特徴的なモチーフであるがこれらがその配列のままいづれのkinaseにもあるわけではない。これらのモチーフの有無の判断は実際には整列表記したときに複数モチーフ全体として10残基程度が一致する場合にモチーフありとした。

自然界ではあるモチーフに起こったミスをほかのモチーフが残基を変化させて補うといったrevertantもある。モチーフ自身の中である残基は実際には他の残基に変わり得るものもある。実験データが数多くたまってもモチーフを確定的な配列に洗練する事は原理的に難しい。

```
                (A)PROTEIN KINASE:  motif
    VALYDY(77)FLVRES(102)VAVKTLK(90)IHRDL(16)DFGLAR(17)KWTAPE(10)SDVWA
     SH3         SH2          IV          VIb        VII       VIII       IX
                (B)PROTEIN KINASE:  domain structure
22---19---15---15---13---11---23---18---16---16---7---16---7---20---7---26---4
I    II   III  IV   ---V---   VIa  VIb  VII VIII  ----IX----     X    ----XI----
                (C)PROTEIN KINASE:  gap position
 10-17/4-17/2-9/ 1-6/  1/ 3-20/ 5-9/7-11/2-19/3-11/ 1/   1/   7/  44/ 1/   1
```

上の図(A)はモチーフがある一つのタンパク質でどれほどのアミノ酸残基数の間隔で並んでいるか、そして(B)のサブドメインのどこに存在しているかを示す。図(B)はHanks,S.K.(4)らが分けているkinase catalytic domainのサブドメイン構造を示し、数字は各サブドメインの長さを残基数で示す。図(C)はサブドメインの間のギャップの長さを示す。整列表記した配列により1から44残基の様々な長さのギャップがある事が分かる。図(A),(B),(C)を比較すると分かるように整列表記におけるギャップの長さに変化があるためモチーフの存在場所の判断が難しい。モチーフの変化と並びをどの様に評価するかが分類の自動化のための課題である。本報告では人間が判断した。

結論

タンパク質の分類を二つの手法の組み合わせで行った。弱い類似性の検索でまず配列をグループ化し、更にFRAPライブラリーのモチーフの有無で分類した。この方法を多くの生物から得られたタンパク質を含むスーパーファミリー89個についてPIRのデータを利用して評価した。この方法で56個は容易に分類できた。9個は更に弱い類似性を検索し、モチーフを利用する事で分類が可能である。20個については、我々の分類規範では複数のスーパーファミリーが同族となった。これはタンパク質がドメイン構造を持っている事とも関係があり、Doolittle(5)が指摘している分類規範の問題である。残り4個は本方法では分類が困難で同一性を決める方法を開発する事が必要であった。

参考文献

1)Dayhoff,M.O., Barker,W.C.: Methods Enzymol., 91,524-545(1983)
2)Nishikawa,K., Nakashima,H., Kanehisa,M., Ooi,T.: Protein Seq Data Anal, 1, 107-116(1987)
3)Seto,Y.,Ikeuchi,Y.,Kanehisa,M.:PROTEINS, 8,341-351(1990)
4)Hanks,S.K.,Quinn,A.M.:Methods Enzymol., 200,38-62(1991)
5)Doolittle,R.F.: Searching through Sequence Databases: Methods Enzymol., 183, 99-110(1990)