รายการอ้างอิง

[1] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2006). GenBank. Nucleic Acids Res 34: D16-20.

[2] Barker, W.C., Garavelli, J.S., McGarvey, P.B., Marzec, C.R., Orcutt, B.C., Srinivasarao, G.Y., et al. (1999). The PIR-International Protein Sequence Database. Nucleic Acids Res 27: 39-43.

[3] George, D.G., Barker, W.C., Mewes, H.W., Pfeiffer, F., and Tsugita, A. (1994). The PIR-International Protein Sequence Database. Nucleic acids research 22: 3569-3573.

[4] Barker, W.C., Pfeiffer, F., and George, D.G. (1996). Superfamily classification in PIR-International Protein Sequence Database. Methods in enzymology 266: 59-71.

[5] Wu, C.H., Zhao, S., and Chen, H.L. (1996). A protein class database organized with ProSite protein groups and PIR superfamilies. J Comput Biol 3: 547-561.

[6] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the Universal Protein knowledgebase. Nucleic acids research 32: D115-119.

[7] Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., et al. (2005). The Universal Protein Resource (UniProt). Nucleic acids research 33: D154-159.

[8] Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., et al. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic acids research 34: D187-191.

[9] Andreeva, A., Prlic, A., Hubbard, T.J., and Murzin, A.G. (2007). SISYPHUS--structural alignments for proteins with non-trivial relationships. Nucleic acids research 35: D253-259.

[10] Stebbings, L.A., and Mizuguchi, K. (2004). HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. Nucleic acids research 32: D203-207.

[11] Woodcock, S., Mornon, J.P., and Henrissat, B. (1992). Detection of secondary structure elements in proteins by hydrophobic cluster analysis. Protein engineering 5: 629-635.

[12] Mechin, M.C., Bertin, Y., and Girardeau, J.P. (1995). Hydrophobic cluster analysis and secondary structure predictions revealed that major and minor structural subunits of K88-related adhesins of Escherichia coli share a common overall fold and differ structurally from other fimbrial subunits. FEBS Lett 364: 319-324.

[13] Gaboriaud, C., Bissery, V., Benchetrit, T., and Mornon, J.-P. (1987). Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett 224: 149-155.

[14] Krishna, S.S., Sadreyev, R.I., and Grishin, N.V. (2006). A tale of two ferredoxins: sequence similarity and structural differences. BMC Struct Biol 6: 8.

[15] Raimbaud, E., Buleon, A., Perez, S., and Henrissat, B. (1989). Hydrophobic cluster analysis of the primary sequences of alpha-amylases. International journal of biological macromolecules 11: 217-225.

[16] Callebaut, I., Dulin, F., Bertrand, O., Ripoche, P., Mouro, I., Colin, Y., et al. (2006). Hydrophobic cluster analysis and modeling of the human Rh protein three-dimensional structures. Transfus Clin Biol 13: 70-84.

[17] Whitford, D. (2005). Proteins: Structure and Function. England: John Wiley & Sons Ltd. 542.

[18] Foobar. (2006). Schematic representation of the different categories of polytopic membrane proteins [Online]. Available from: http://en.wikipedia.org/wiki/Transmembrane_protein [31 August 2008]

[19] Foobar. (2008). Schematic representation of the different types of interaction between monotopic membrane proteins and the cell membrane [Online]. Available from: http://en.wikipedia.org/wiki/Peripheral_membrane_protein [31 August 2008]

[20] Rulez, E. (2005). Fancy cartoon model of the collagen triple helix [Online]. Available from: [31 August 2008]

[21] Liebecq, C. (1992). Biochemical Nomenclature and Related Documents. Portland Press. 347.

[22] Heinau, V., and Kirste, B. (1994). Amino Acids [Online]. Available from: http://www.chemie.fu-berlin.de/chemistry/bio/amino-acids_en.html [30 August 2008]

[23] Lesser, G.J., Lee, R.H., Zehfus, M.H., and Rose, G.D. (1987). Hydrophobic Interactions in Proteins: in Protein Engineering. New York: Alan R. Liss.

[24] Ruddle, F.H. (1998). Mapping and sequencing of the human genome. Jpn J Cancer Res 89: inside front cover.

[25] Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443-453.

[26] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol 215: 403-410.

[27] Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics (Oxford, England) 21: 951-960.

[28] Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M., Lautrup, B., Norskov, L., et al. (1988). Protein secondary structure and homology by neural networks. The alpha-helices in rhodopsin. FEBS Lett 241: 223-228.

[29] Karplus, K., Karchin, R., Barrett, C., Tu, S., Cline, M., Diekhans, M., et al. (2001). What is the value added by human intervention in protein structure prediction? Proteins Suppl 5: 86-91.

[30] Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics (Oxford, England) 14: 755-763.

[31] Henrissat, B., Popineau, Y., and Kader, J.-C. (1988). Hydrophobic-cluster analysis of plant protein sequences. A domain homology between storage and lipid-transfer proteins. Biochem J 255: 901-905.

[32] Lemesle-Varloot, L., Gaboriaud, C., Morgat, A., Pantel, G., Mornon, J.-P., Lavaitte, S., et al. (1993). MANSEK and SUNHCA. Two interactive programs for the

hydrophobic cluster analysis of protein sequences. Comput Appl Biosci 9: 37-44.

[33] Laget, M.P., Callebaut, I., de Launoit, Y., Stehelin, D., and Mornon, J.-P. (1993). Predicted common structural features of DNA-binding domains from Ets, Myb and HMG transcription factors. Nucleic Acids Res 21: 5987-5996.

[34] Callebaut, I., Prat, K., Meurice, E., Mornon, J.-P., and Tomavo, S. (2005). Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. BMC Genomics 6: 100.

[35] Girardeau, J.P., Bertin, Y., and Callebaut, I. (2000). Conserved structural features in class I major fimbrial subunits (Pilin) in gram-negative bacteria. Molecular basis of classification in seven subfamilies and identification of intrasubfamily sequence signature motifs which might be implicated in quaternary structure. J Mol Evol 50: 424-442.

[36] Breton, C., Oriol, R., and Imberty, A. (1998). Conserved structural features in eukaryotic and prokaryotic fucosyltransferases. Glycobiology 8: 87-94.

[37] Callebaut, I., Courvalin, J.C., Worman, H.J., and Mornon, J.-P. (1997). Hydrophobic cluster analysis reveals a third chromodomain in the Tetrahymena Pdd1p protein of the chromo superfamily. Biochem Biophys Res Commun 235: 103-107.

[38] Geremia, R.A., Petroni, E.A., Ielpi, L., and Henrissat, B. (1996). Towards a classification of glycosyltransferases based on amino acid sequence similarities: prokaryotic alpha-mannosyltransferases. Biochem J 318 (Pt 1): 133-138.

[39] Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. Journal of molecular biology 147: 195-197.

[40] Goonesekere, N.C., and Lee, B. (2004). Frequency of gaps observed in a structurally aligned protein pair database suggests a simple gap penalty function. Nucleic acids research 32: 2838-2843.

[41] Lemesle-Varloot, L., Gaboriaud, C., Morgat, A., Pantel, G., Mornon, J.P., Lavaitte, S., et al. (1993). MANSEK and SUNHCA. Two interactive programs for the

hydrophobic cluster analysis of protein sequences. Comput Appl Biosci 9: 37-44.

[42] Heijne, G.v. (1986). The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. Embo J 5: 3021-3027.

[43] Kannasut, P., Pichyangkura, R., and Ratanamahatana, C.A. (2007). Towards 2-D Automatic Hydrophobic Cluster Alignment. National Conference on Biomedical Engineering, pp. 91-94. Bangkok.

[44] Sali, A., and Blundell, T.L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. Journal of molecular biology 212: 403-428.

[45] Zhu, Z.Y., Sali, A., and Blundell, T.L. (1992). A variable gap penalty function and feature weights for protein 3-D structure comparisons. Protein engineering 5: 43-51.

[46] Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S., and Overington, J.P. (1998). JOY: protein sequence-structure representation and analysis. Bioinformatics (Oxford, England) 14: 617-623.

ภาคผนวก

ภาคผนวก ก

ข้อมูลที่ใช้ในการทดลอง

ข้อมูลลำดับกรดอะมิโนที่ใช้ในการทดลองนั้นมาจาก 3 แหล่ง คือ ฐานข้อมูล PIR [2-5] ฐานข้อมูล HOMSTRAD [10] และฐานข้อมูล SISYPHUS [9] นำมาสร้างเป็นชุดข้อมูลทั้งหมดจำนวน 2 ชุดข้อมูล

ก.1 ชุดข้อมูลจาก SISYPHUS

ฐานข้อมูล SISYPHUS เป็นฐานข้อมูลที่มีการทำการจับคู่ลำดับกรดอะมิโนโดยผู้เชี่ยวชาญ เป็นการจับคู่แบบดูโครงสร้างเป็นหลัก ดังนั้นจึงเหมาะสำกับเป็นข้อมูลสำหรับใช้ทดสอบระหว่างการจับคู่กลุ่มโดยผู้เชี่ยวชาญ กับการใช้เครื่องมือจับคู่กลุ่มที่ไม่ชอบน้ำแบบ 2 มิติโดยอัตโนมัติ ซึ่งเป็นการจับคู่โดยใช้หลักการจับคู่โครงสร้างที่ตรงกันเป็นหลักเหมือนกัน โดยได้มีการเปรียบเทียบไว้ 119 กลุ่มโครงสร้าง สามารถแยกเป็นคู่ได้ทั้งหมด 8,474 คู่ รายละเอียดของชุดข้อมูลแสดงได้ดังตารางที่ ก.1

ตารางที่ ก.1 รายละเอียดของชุดข้อมูลจาก SISYPHUS

Accession code	Protein name	Class	Length
AL00069572	Activating enzymes of ubiquitin-like proteins	alpha-beta	212
AL00050692	ADC-like	All Beta	171
AL00053617	Adenosine kinase	alpha-beta	346
AL00051906	Adrenodoxin reductase, C-terminal domain-like	alpha-beta	265
AL10069117	AhpD-like	All Alpha	168
AL00069381	Alpha subunit of glutamate synthase, central and FMN domains	alpha-beta	469
AL00064294	Autoinducer-2 production protein LuxS	alpha-beta	146

ตารางที่ ก.1 รายละเอียดของชุดข้อมูลจาก SISYPHUS (ต่อ)

Accession code	Protein name	Class	Length
AL00050876	Avidin/streptavidin-like	All Beta	125
AL00102031	AXH domain	All Beta	127
AL00063492	BAG domain	All Alpha	104
AL00050685	Barwin-like endoglucanases	All Beta	206
AL10050335	biMOP-like	All Beta	140
AL00052441	Biotin carboxylase (BC), N-terminal domain-like	alpha-beta	143
AL00052113	BRCT domain	alpha-beta	99
AL10050875	Calycins	All Beta	167
AL00053473	Cannonical alpha/beta hydrolyses	All Beta	457
AL00050155	Cannonical PDZ domains	alpha-beta	108
AL00053335	Cannonical S-adenosyl-L-methionine-dependent methyltransferases	alpha-beta	306
AL00055637	Cell cycle regulatory proteins	alpha-beta	79
AL00048599	Chorismate mutase II	All Alpha	261
AL00052317	Class I glutamine amidotransferase (GAT) -like	alpha-beta	387
AL00057468	Classical and nonclassical Kazal-type domains.	alpha plus beta	55
AL00088995	CUE domain	alpha plus beta	41
AL00051324	Cyanovirin-N	All Beta	101
AL00054403	Cystatin/Monellin-like	alpha-beta	108
AL00054407	Cystatins	alpha-beta	112
AL00089800	Deoxycytidylate deaminase-like	alpha-beta	152
AL00069394	Deoxyribose-phosphate aldolase DeoC	alpha-beta	245

ตารางที่ ก.1 รายละเอียดของชุดข้อมูลจาก SISYPHUS (ต่อ)

Accession code	Protein name	Class	Length
AL00052141	DNA glycosylases - a common domain core	alpha-beta	215
AL00050465	EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain	All Beta	117
AL00050083	EPS8 SH3 domain	All Beta	59
AL00054790	Eukaryotic type KH-domain (KH-domain type I)	alpha-beta	89
AL00052150	FabD/lysophospholipase-like	alpha-beta	557
AL00051904	FAD/NAD(P)-binding domain	alpha-beta	372
AL00051913	FAD-linked reductases, N-terminal domain-like	alpha-beta	441
AL00068994	FAT domain of focal adhesion kinase	All Alpha	125
AL00050847	Fatty acid binding proteins (FABPs)	All Beta	142
AL00047241	Ferritin-like	All Alpha	152
AL00101117	Flagellar export chaperone FliS	All Alpha	110
AL00050475	FMN-binding split barrel	All Beta	209
AL00051395	FMN-linked oxidoreductases - like	alpha-beta	500
AL00054253	GABARAP/GATE-16/LC3/ATG12 family	alpha-beta	88
AL00051931	GDI/CHM/REP family, N-terminal domain-like	alpha-beta	479
AL10054593	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	alpha-beta	438
AL00057282	Grasshopper inhibitor family	All Beta	32
AL00055563	Growth factor receptor-bound protein 2 (GRB2)	alpha-beta	95

ตารางที่ ก.1 รายละเอียดของชุดข้อมูลจาก SISYPHUS (ต่อ)

Accession code	Protein name	Class	Length
AL00057413	Hairpin-loop-domain	alpha plus beta	83
AL00054624	Heme-binding protein A (HasA)	alpha-beta	173
AL00055594	HPr-like	alpha-beta	84
AL00051412	Inosine monophosphate dehydrogenase (IMPDH) - like	alpha-beta	488
AL00047305	Interferons/Interleukin-10 (IL-10) family	All Alpha	147
AL00050021	ISP domain	All Beta	190
AL10101215	KaiA/RbsU dimer	All Alpha	180
AL00101215	KaiA/RbsU domain	All Alpha	90
AL00102449	KaiB/SasA	alpha-beta	93
AL00069960	Kazal-type proteinase inhibitor LEKTI, domain 1 and 6	alpha plus beta	54
AL00051306	LexA/Signal peptidase	All Beta	229
AL00050815	Lipocalins	All Beta	171
AL10050815	Lipocalins/Triabin	All Beta	138
AL10063410	LuxS/MPP/ThrRS/AlaRS common domain	alpha-beta	176
AL00063410	LuxS/MPP-like metallohydrolases	alpha-beta	206
AL00046695	MAT alpha2	alpha plus beta	77
AL20089447	MazE/MraZ/AbrB	All Beta	122
AL00056022	Mitotic spindle assembly checkpoint protein mad2	alpha-beta	185
AL00055723	Mog1p/PsbP-like	alpha-beta	201
AL00050331	MOP-like	All Beta	64
AL00063412	MPP-like	alpha-beta	206
AL00051984	MurCD N-terminal domain	alpha-beta	89
AL00051735	NAD(P)-binding Rossmann-fold domains	alpha-beta	324

ตารางที่ ก.1 รายละเอียดของชุดข้อมูลจาก SISYPHUS (ต่อ)

Accession code	Protein name	Class	Length
AL00051971	Nucleotide-binding domain	alpha-beta	333
AL00055464	Origin of replication-binding domain, RBD-like	alpha-beta	205
AL00110848	ParB/Sulfiredoxin	alpha-beta	73
AL00053701	Phosphoglucose isomerase, PGI	alpha-beta	513
AL00052787	Phosphotyrosine protein phosphatases type I -like	alpha-beta	162
AL00052799	Phosphotyrosine protein phosphatases type II	alpha-beta	306
AL00050799	PK beta-barrel domain-like	All Beta	186
AL00100897	Plant proteinase inhibitors - PotII family (known also as Pin2 family)	All Beta	48
AL00089048	Polcalcin	All Alpha	76
AL00054097	Prion-like proteins	All Alpha	100
AL00054814	Prokaryotic type KH domain, KH-domain type II	alpha-beta	97
AL10050155	Protease specific and cannonicalPDZ domains	All Beta	116
AL10074933	Protease specific PDZ-domains	All Beta	113
AL00053354	Protein-L-isoaspartyl O-methyltransferases (PIMT)	alpha-beta	215
AL00100877	Pseudouridine synthases	alpha-beta	277
AL00103575	PSI-domain	alpha-beta	66
AL00109622	PurS-like	alpha-beta	157
AL00050779	Radixin, third FERM domain	All Beta	106

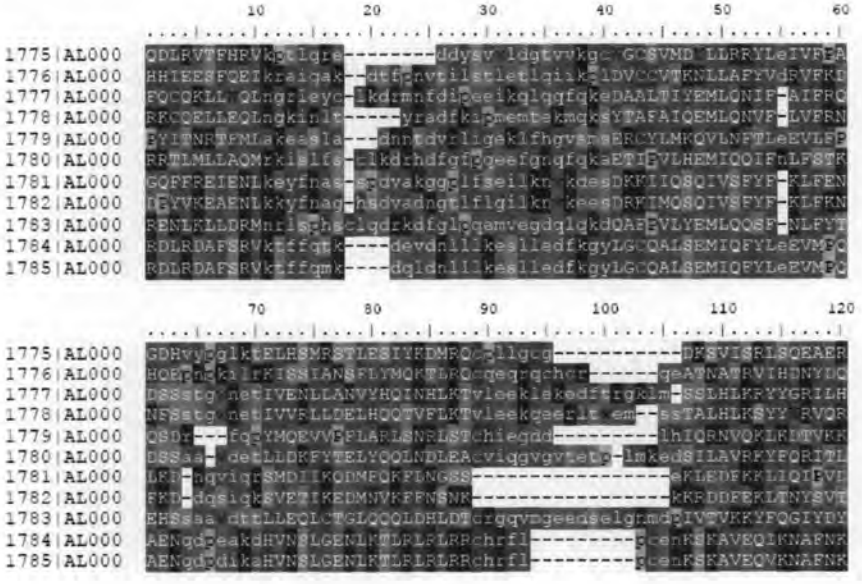
ตารางที่ ก.1 รายละเอียดของชุดข้อมูลจาก SISYPHUS (ต่อ)

Accession code	Protein name	Class	Length
AL00048098	Regulator of G-protein signaling, RGS	All Alpha	122
AL00103120	Relaxase domain	alpha-beta	301
AL00052821	Rhodanese/Cell cycle control phosphatase	alpha-beta	162
AL00053614	Ribokinase-like	alpha-beta	380
AL00089623	Ribose isomerase RpiB	alpha-beta	145
AL00050715	Ribosomal protein L25/GlnRS	All Beta	119
AL00051366	Ribulose-phoshate binding beta/alpha-barrels	alpha-beta	366
AL00063563	RNA polymerase omega subunit	All Alpha	91
AL00055294	RPB6	alpha-beta	73
AL00056276	S-adenosylmethionine decarboxylase (AdoMetDC)	alpha-beta	302
AL10056276	S-adenosylmethionine decarboxylase structural repeats	alpha plus beta	127
AL00063763	SAND domain-like	alpha-beta	75
AL00047861	Saposin-like	All Alpha	78
AL00053697	SIS domain-like	alpha-beta	235
AL00049879	SMAD/FHA domain	All Beta	234
AL00052186	Sporulation response regulator Spo0A (N-Spo0A)	alpha-beta	123
AL00054446	ssDNA-binding transcriptional regulator domain	alpha-beta	161
AL00055398	Subtilisin inhibitors	alpha-beta	107
AL00051934	Succinate dehydrogenase/fumarate reductase flavoprotein N-terminal domain	alpha-beta	356

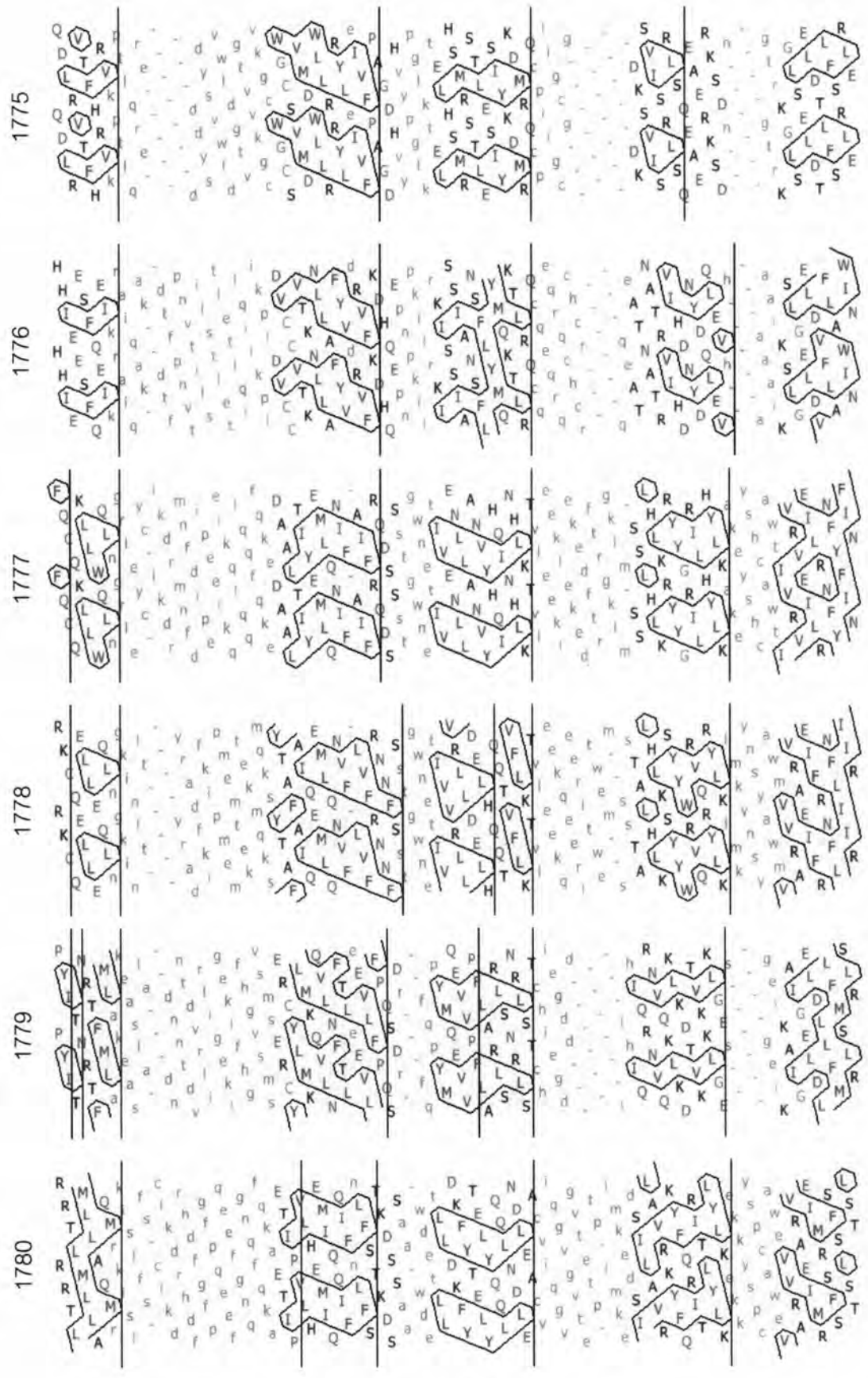
ตารางที่ ก.1 รายละเอียดของชุดข้อมูลจาก SISYPHUS (ต่อ)

Accession code	Protein name	Class	Length
AL00051392	Thiamin phosphate synthase - like	alpha-beta	205
AL00110400	ThiG-like	alpha-beta	242
AL00055186	ThrRS/AlaRS common domain	alpha-beta	162
AL00074653	TolA/TonB C-terminal domain	alpha-beta	69
AL00089155	TorD-like	All Alpha	202
AL00049599	TRAF domain-like	All Beta	153
AL10050464	Translation factors, domains 2 and 3	All Beta	123
AL00057492	Trefoil	alpha plus beta	46
AL00051352	Triosephosphate isomerase (TIM)	alpha-beta	264
AL00103021	tRNA pseudouridine synthase TruD	alpha-beta	409
AL00051713	tRNA-guanine transglycosylase - like	alpha-beta	369

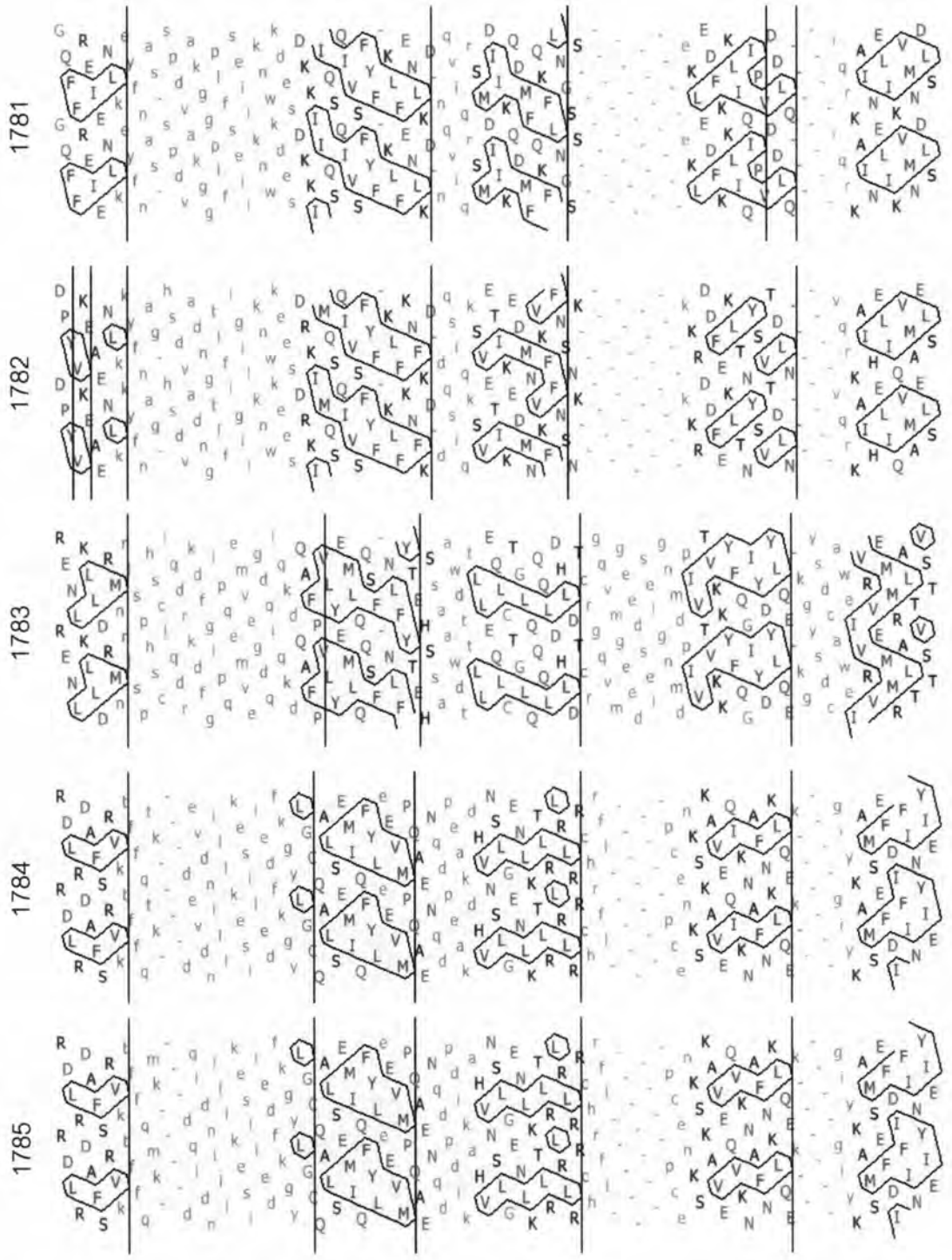
ข้อมูลภายในแต่ละกลุ่มหน้าที่ของโปรตีนได้มีการจับคู่กลุ่มไว้โดยผู้เชี่ยวชาญ ตัวอย่างเช่นในกลุ่มของ AL00047305 Interferons/Interleukin-10 (IL-10) family ผู้เชี่ยวชาญได้จับคู่ลำดับกรดอะมิโนสามารถแสดงเป็นแผนภูมิ 1 มิติดังรูปที่ ก.1 และแผนภูมิ 2 มิติดังรูปที่ ก.2



รูปที่ ก.1 การจับคู่โครงสร้างโดยผู้เชี่ยวชาญใน SISYPHUS แสดงแบบแผนภูมิ 1 มิติ



รูปที่ ก.2 การจับคู่โครงสร้างโดยผู้เชี่ยวชาญใน SISYPHUS แสดงแบบแผนภูมิ 2 มิติ



รูปที่ ก.2 การจับคู่โครงสร้างโดยผู้เชี่ยวชาญใน SISYPHUS แสดงแบบแผนภูมิ 2 มิติ (ต่อ)

ก.2 ชุดข้อมูลจาก HOMSTRAD

ฐานข้อมูล HOMSTRAD เป็นฐานข้อมูลการจับคู่ด้วยโครงสร้างสำหรับโปรตีนที่ทำหน้าที่เหมือนกัน ประกอบไปด้วยกลุ่มโปรตีนจำนวน 130 กลุ่ม (Family) มีโครงสร้างทั้งหมด 530 รูปแบบ โดยสามารถแบ่งเป็นกลุ่มย่อยได้ 1031 กลุ่ม โดยรายละเอียดแสดงไว้ในตารางที่ ก.2 ซึ่งเป็นโปรตีนที่ได้คัดเลือกมาจากฐานข้อมูล PIR โดยในแต่ละโครงสร้างโปรตีนที่เลือกมาจะต้องเป็นโครงสร้างที่ได้มาจากการใช้รังสีเอ็กซ์ในการวิเคราะห์โครงสร้างที่ถูกต้อง ซึ่งการจับคู่ในแต่ละกลุ่มโปรตีนใช้เครื่องมือการจับคู่โดยอาศัยโครงสร้าง COMPARER [44, 45] และใช้เครื่องมือ JOY [46] สำหรับการวิเคราะห์โครงสร้างของลำดับกรดอะมิโนแต่ละตัว โดยตัวอย่างรายละเอียดแสดงไว้ในตารางที่ ก.2

ตารางที่ ก.2 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD

PDB Structure	Protein Name	Class	Length
1bgxt,1xo1a	5'-3' exonuclease	multi domain	399
1bif,1k6ma	6-phosphofructo-2-kinase	alpha beta	101
2pgd,1pgjb	6-phosphogluconate dehydrogenases	multi domain	195
1c8ca,1azpa	7kD DNA-binding domain	alpha plus beta	6
1ayoa,1edya	Alpha-2-macroglobulin family A	all beta	187
1qqfa,1c3d	Alpha-2-macroglobulin family B	all alpha	219
1g41a,1e94e,1e32a,1d2na	ATPase family associated with various cellular activities (AAA)	alpha beta	961
1hsla,1lst,1ggga	bacterial extracellular solute-binding proteins, family 3	alpha beta	475
1c1da,1leha,1gtma,1bvua,1euza,1b26a,1hwxa,1bgva	amino acid dehydrogenase	alpha beta	1008
1b7ba,1e19a	Amino acid kinase family	alpha beta	366

ตารางที่ ก.2 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

PDB Structure	Protein Name	Class	Length
1ajsa,2csta,1yaaa,1ama,1ay8a,1ars,3tata,1b8ga,1bjwa,1bw0a	aminotransferases class-I	alpha beta	795
1b0ua,1f3oa,1g291,1g6ha,1f2ua,1e69a	ABC transporter	alpha beta	1355
1auoa,1fj2a	Phospholipase/Carboxylesterase	alpha beta	439
2abd,1hbka	Acyl CoA binding protein	all alpha	195
1maha,2ace,1clea,1trh,1thg,2bce	alpha beta-hydrolase	alpha beta	712
1bo4a,1cm0a,1ygha,1qsta,1b87a,1cjwa	Acetyltransferase (GNAT) family	alpha plus beta	211
1e2ta,1gx3a	N-acetyltransferase	alpha plus beta	154
2hpaa,1rpt,1ihp	Histidine acid phosphatase	alpha beta	373
1c96a,1l5ja	Aconitase family (aconitate hydratase)	alpha beta	385
1qr0a2,1f7la,1qr0a1	4'-phosphopantetheinyl transferase superfamily	alpha plus beta	404
1phza,1psda	ACT domain	alpha plus beta	390
1atna,3hsc,1hjoa,1dkgd	actin/heat-shock cognate	alpha beta	301
1es7b,1btea	Activin types I and II receptor domain	small	48

ตารางที่ ก.2 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

PDB Structure	Protein Name	Class	Length
1aps,2acy	acylphosphatase	alpha plus beta	301
1knb,1qhva,1noba	Adenovirus fiber protein head domain (knob domain)	all beta	97
1dj3a,1dj2a,1qf5a	Adenylosuccinate synthetase	alpha beta	306
1cof,1cnua,1ak6	actin-depolymerizing proteins	alpha plus beta	372
3huda,1cdoa,1teha,2ohxa,1d1ta	alcohol dehydrogenase	multi domain	191
1aky,1akea,1zin,1zaka,2ak3a,3adk,1uky,1uke,1gky	nucleotide kinase	alpha beta	981
1gc5a,1l2la	ADP-specific Phosphofructokinase/Glucokinase conserved region	alpha beta	45
1b25a,1aora	Aldehyde ferredoxin oxidoreductase	multi domain	184
1jlxa1,1jlxa2	Agglutinin	all beta	2
1qq2a,1qmva,1prxa	AhpC/TSA family	alpha beta	429
1b4ka,1aw5	Delta-aminolevulinic acid dehydratase	alpha beta barrel	413
1f8ga,1pjca	Alanine dehydrogenase/pyridine nucleotide transhydrogenase	alpha beta	426

ตารางที่ ก.2 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

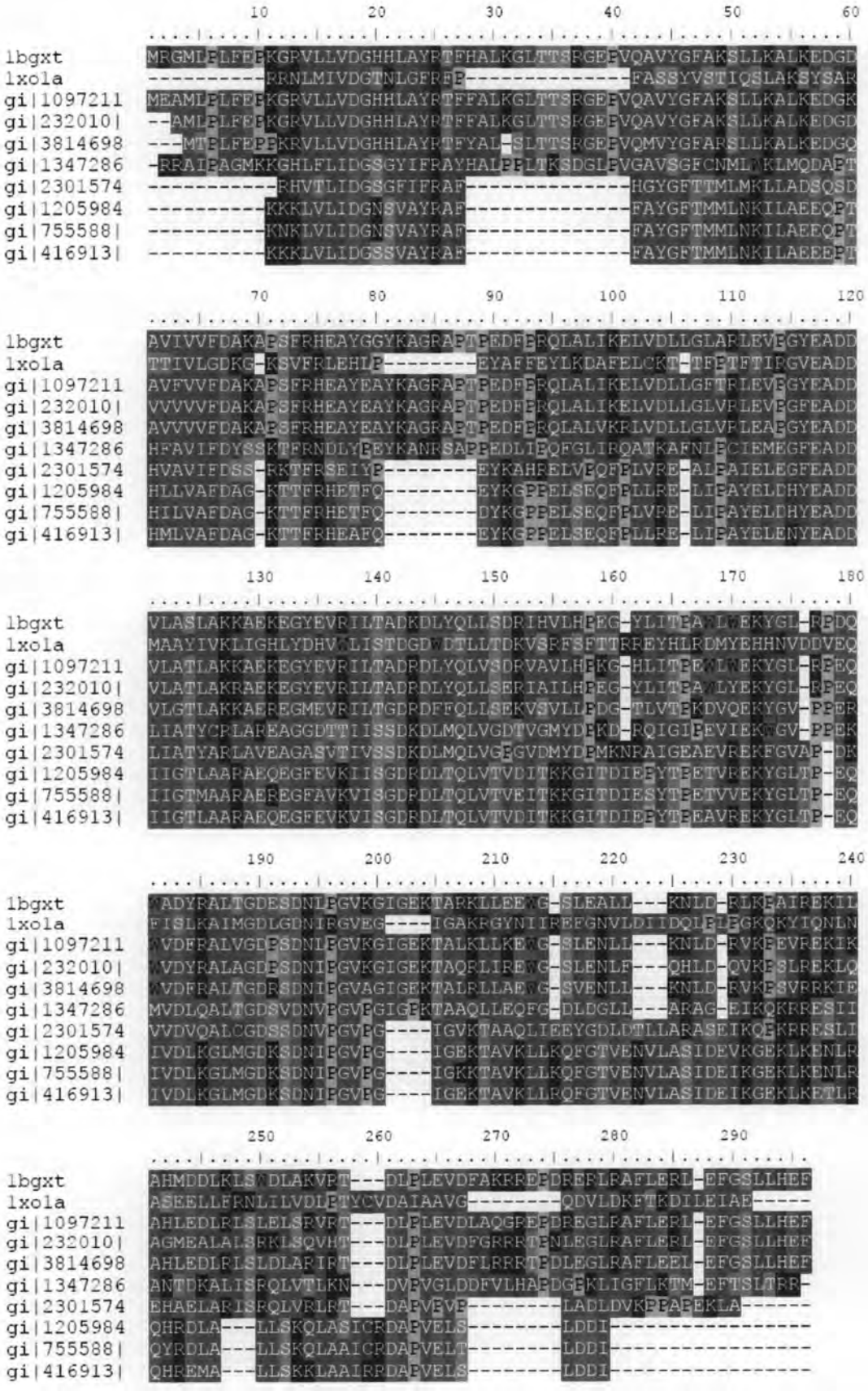
PDB Structure	Protein Name	Class	Length
1f8ga,1pjca	Alanine dehydrogenase/pyridine nucleotide transhydrogenase domain 1	alpha beta	537
1f8ga,1pjca	Alanine dehydrogenase/pyridine nucleotide transhydrogenase domain 2	alpha beta	392
1bd0a,1ct5a	Alanine racemase, N-terminal domain	alpha beta barrel	438
1bj51,1bj52,1bj53	albumin	all alpha	395
1a4sa,1bxsa,1ag8a,1ad3a	Aldehyde dehydrogenase	alpha beta	436
1fua,1jdia	Class II Aldolase	alpha beta	369
1ald,1fbaa,1a5ca	fructose-1,6-bisphosphate aldolase	alpha beta barrel	229
1ah4,1ads,1frb,2alr,1afsa,1a80,1qrqa	aldo/keto reductase	alpha beta barrel	652
1hlra,1dgja,1jroa,1fo4a,1n62a,1ffva	Aldehyde oxidase and xanthine dehydrogenase, domains 1-2	multi domain	737
1dgja,1hlra,1fo4a,1jrob,1ffvb,1n62b	Aldehyde oxidase and xanthine dehydrogenase, domains 3-4	multi domain	1127
1ed9a,1ew2a	Alkaline phosphatase	alpha beta	317

ตารางที่ ก.2 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

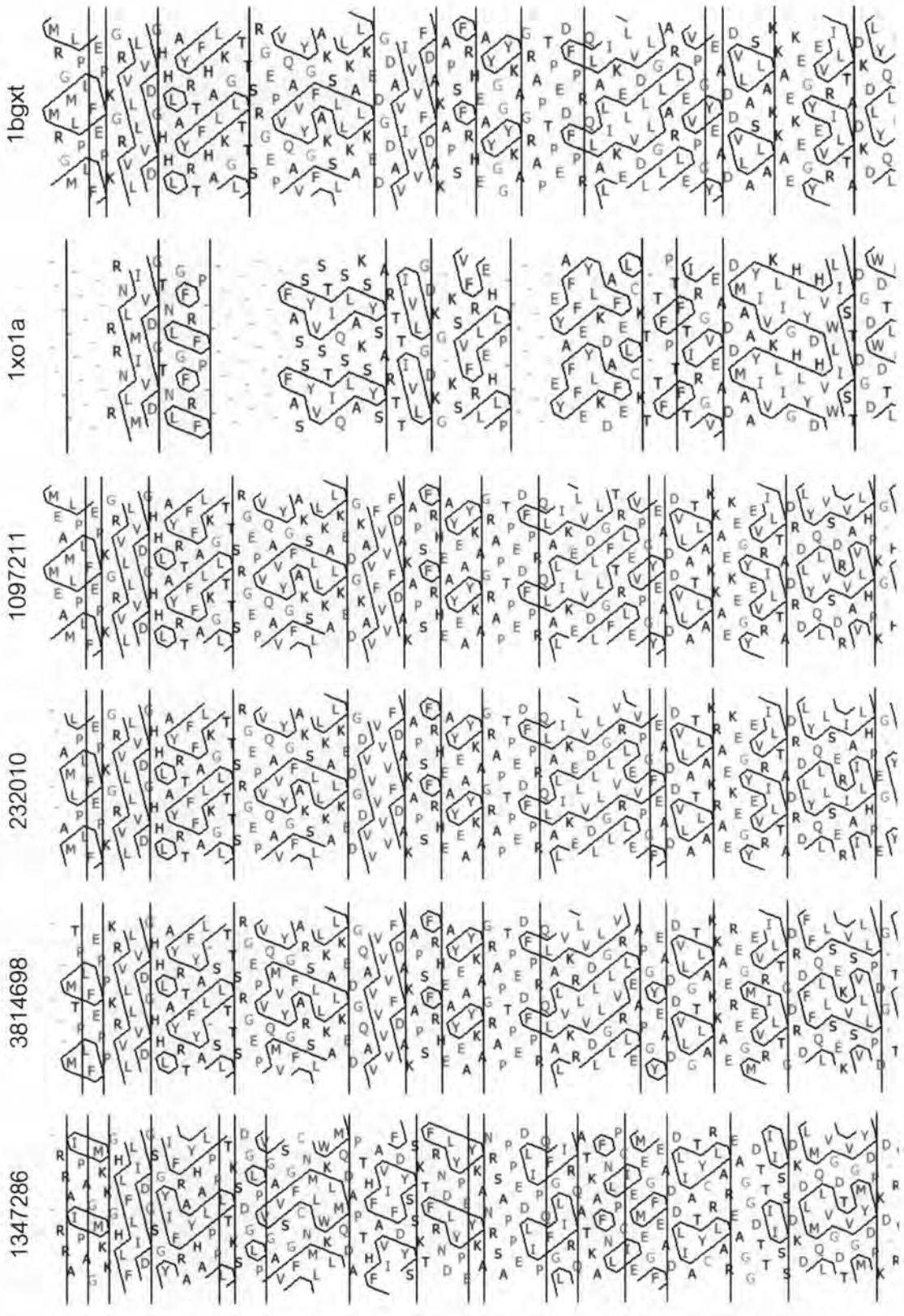
PDB Structure	Protein Name	Class	Length
1hx0a,1smd,1jae,1g94a,1bag,1smaa,1bvza,1uok,2aaa,7taa,1cgt,1d3ca,1ciu,1cyg,1qhpa,1hvxa,1vjs,1gcya,1avaa,1ehaa,1bf2,1g5aa,1gjwa	Alpha amylase, catalytic domain	alpha beta barrel	3264
1hvxa,1vjs,2aaa,7taa,1gjwa,1smaa,1uok,1g5aa,1bvza,1ciu,1qhpa,1pama,1d3ca,1cgt,1cyg,1bag,1smd,1hx0a,1jae,1g94a,1avaa	Alpha amylase, C-terminal domain	all beta	471
1bvza,1smaa	Alpha amylase, N-terminal ig-like domain	all beta	81
1bf2,1ehaa,1hx0a,1smd,1jae,1g94a,1bag,1smaa,1bvza,1uok,2aaa,7taa,1cgt,1d3ca,1ciu,1cyg,1qhpa,1hvxa,1vjs,1gcya,1avaa,1g5aa,1gjwa	Alpha amylase, catalytic and C-terminal domains	multi domain	3272
1qtsa,1e42a	Alpha adaptin AP2, C-terminal region (consists of 2 domains)	multi domain	44
1qtsa,1e42a	Alpha adaptin AP2, N-terminal domain of C-terminal region	all beta	34

ตารางที่ ก.2 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

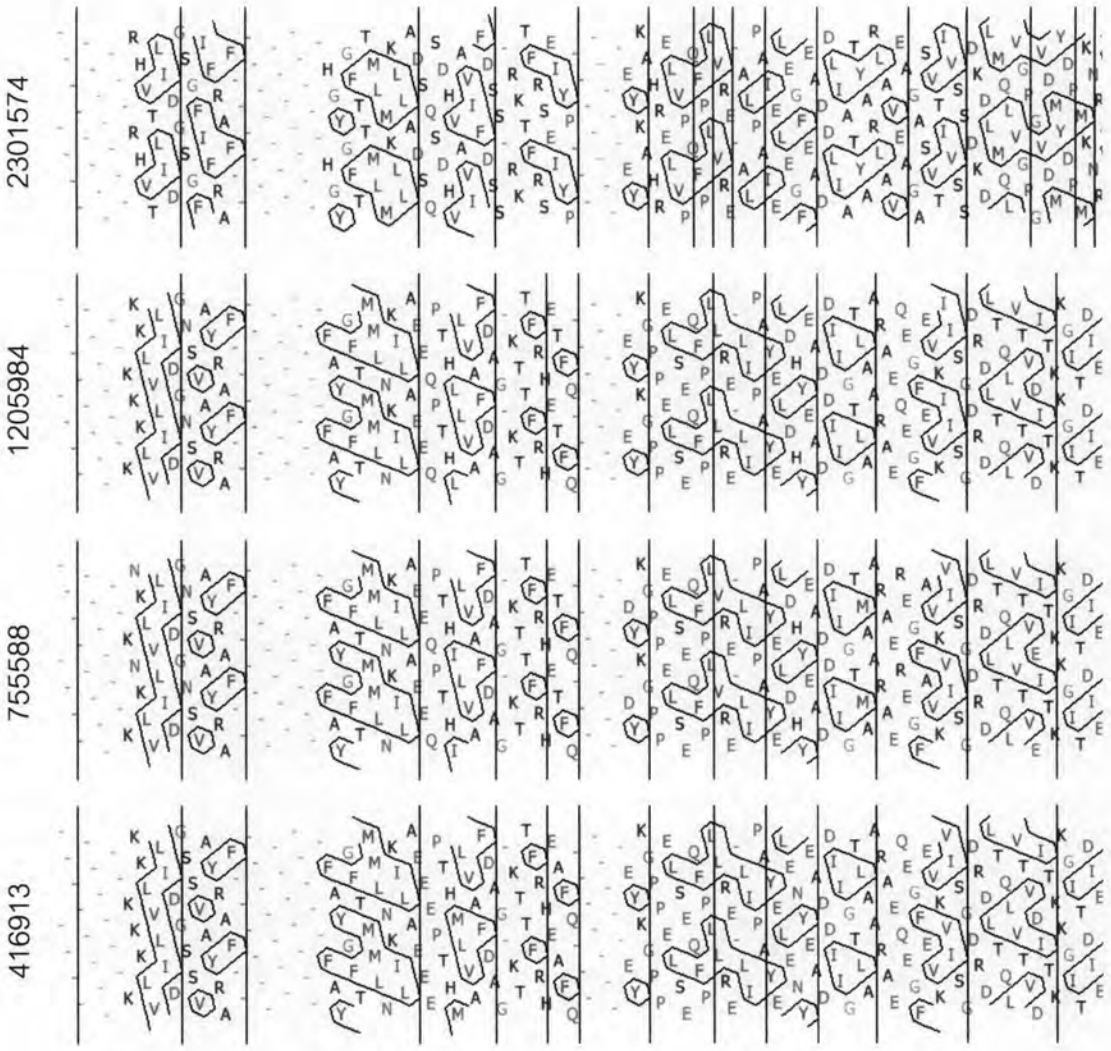
PDB Structure	Protein Name	Class	Length
1qtsa,1e42a	Alpha adaptin AP2, C-terminal domain of C-terminal region	alpha plus beta	35
1ocka,1m22a	Amidase	alpha beta	438
1lba,1j3ga	N-acetylmuramoyl-L-alanine amidase	alpha plus beta	438
1jdw,1bwda	Amidinotransferase	alpha plus beta	276
2lbp,2liv	periplasmic binding protein - branched-chain amino acid	alpha beta	320
5daaa,1a3ga	Aminotransferase class IV	multi domain	430
1bjoa,1bt4a	Aminotransferases class-V	alpha plus beta	277
1h83a,1f8sa	Flavin containing amine oxidase	multi domain	385
1h83a,1f8sa	Flavin containing amine oxidase, domain 1	alpha beta	412



รูปที่ ก.3 การจับคู่โครงสร้างใน HOMSTRAD แสดงแบบแผนภูมิ 1 มิติ



รูปที่ ก.4 การจับคู่โครงสร้างใน HOMSTRAD แสดงแบบแผนภูมิ 2 มิติ



รูปที่ ก.4 การจับคู่โครงสร้างใน HOMSTRAD แสดงแบบแผนภูมิ 2 มิติ (ต่อ)

ภาคผนวก ข

ผลงานตีพิมพ์

งานประชุมวิชาการ "ประชุมวิชาการวิศวกรรมชีวการแพทย์แห่งชาติครั้งที่ 5" ซึ่งจัดขึ้น ณ โรงแรมทวินทาวเวอร์ กรุงเทพมหานคร ประเทศไทย ในวันที่ 8 กรกฎาคม 2550 ในหัวข้อเรื่อง "Towards 2-D Automatic Hydrophobic Cluster Alignment" โดย ภิสิทธิ์ กรรณสูต รัฐพิชญางกูร และ โชติรัตน์ รัตนามหัทธนะ

90

CO-3

วันที่ 8 กรกฎาคม 2550 เวลา11:30-11:50 น
ห้องจรัสเมือง 2 ชั้น 2 โรงแรมทวินทาวเวอร์

Towards 2-D Automatic Hydrophobic Cluster Alignment

Phisit Kannasut[1*], Rath Pichyangkura[2], Chotirat Ann Ratanamahatana[3]
[1] Department of Computer engineering, Faculty of Engineering, Bangkok, Thailand
g49pks@cp.eng.chula.ac.th

Current techniques in protein homology testing involve a 1-dimensional alignment of Nucleotide or Amino acid sequencing. Due to its various constraints and low sequence identity values, Hydrophobic Cluster Alignment has been used to predict the structure and functionality of protein. However, this method still needs to be done manually and solely depends on experience and expertise of a researcher. In this work, we implement a 2-D visualization tool for amino acid and propose a new protein representation that could be used effectively in Hydrophobic Cluster Alignment software.

Keywords: automatic, 2-D alignment, hydrophobic cluster analysis, representation, bioinformatics

Towards 2-D Automatic Hydrophobic Cluster Alignment

Phisit Kannasut[1*], Rath Pichyangkura[2], Chotirat Ann Ratanamahatana[1]

[1] Department of Computer engineering, Faculty of Engineering, Bangkok, Thailand
[2] Department of Biochemistry, Faculty of Science, Bangkok, Thailand

* *g49pks@cp.eng.chula.ac.th*

Abstract: *Current techniques in protein homology testing involve a 1-dimensional alignment of nucleotide or amino acid sequencing. Due to its various constraints and low sequence identity values, Hydrophobic Cluster Alignment has been used to predict the structure and functionality of protein. However, this method still needs to be done manually and solely depends on experience and expertise of a researcher. In this work, we implement a 2-D visualization tool for amino acid and propose a new protein representation that could be used effectively in Hydrophobic Cluster Alignment software.*

Introduction

Life science and biological data have grown very rapidly in the past few years. This includes nucleotide/protein sequences available on GenBank and PIR genetic database, the crucial data for protein analyses, e.g., protein homology, structure (Primary, Secondary, Tertiary, or Quaternary), or even functionalities of proteins.

Any form of protein could consist of as many as 20 different amino acids (encoded in a DNA) that are connected with Peptide bonds. Protein must be first extracted from the cells or organisms before we can study its structure, functionality, or identify its amino acid sequence. A renowned project such as the Human genome project has led to an emergence of protein and amino acid sequence database.

The fact that each protein consists of up to 20 different amino acids in random order makes an analysis of protein homology particularly difficult. The current methods are mostly based on Maximum Matching [1], Hidden Markov Model [2], Basic Alignment Search Tool (BLAST) [3] or Hydrophobic Cluster Analysis (HCA) [4], all of which still have major limitations and drawbacks; they could not provide a proof of sequence homology if their sequence identity appears to be too low, which usually occur in proteins from different species (but the proteins themselves have the same functionality) [5].

In this work, we are proposing a tool to help identify amino acid sequences from different organisms that has similar functionality. Our proposed method is based on the idea of Hydrophobic Cluster Analysis, using 2-dimensional diagram that highlights the hydrophobic elements. We then propose a new protein representation that preserves the crucial features that are essential for an effective and meaningful alignment.

Proposed Method

We have adopted HCA approach and illustrated amino acid sequences into 2-D helical pattern by twisting the protein into a smoothed helix, where each twist will contain 3.6 amino acids [4]. Then, this cylinder will be cut lengthwise and spread into 2-dimensional plane, and the hydrophobic amino acid will be highlighted and grouped together, as shown in Fig. 1.

An amino acid sequence can be arranged into 2-dimensional chart according to the following formula

$$ID[i] = (i \times NT) \bmod NR$$

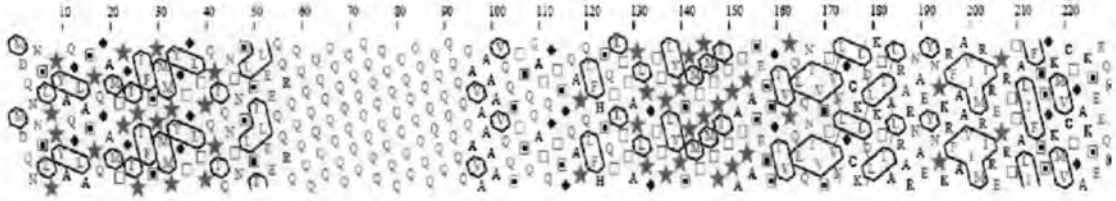


Fig. 1. 2-D representation of the amino acid sequence.

where *NT* is the number of rounds (skewed vertical lines) within one cycle, *NR* is the number of amino acid within one cycle, *i* is the amino acid's ordering in the amino acid sequence.

Visualization Tool

In the process to achieve automatic alignment, we have first implemented a visualization tool on Microsoft Windows environment, using C# .net programming language. Its main functionality is to draw HCA, group hydrophobic cluster, and draw lines to indicate the boundary of the clusters, which will be essential for the automatic alignment algorithm. We also apply the standard color codes to our drawing, while the specialized characters (P,G,T,S) are preserved and not transformed to other symbols as in Lemesle-Varloot L's style [6]. An output sample from our visualization tool is illustrated in Fig. 2.

In the past, after any 2-D HCA visualization is obtained, the rest of the work is given to the human expert to do the hydrophobic cluster alignment. In this work, we attempt to come up with a way to perform an alignment automatically without human intervention. There are several ways that an alignment could be done, such as image recognition, a pattern recognition, string matching, etc. For faster and more accurate alignment, we will avoid an image processing approach. So, we first start off with its representation that can be later used for effective and efficient alignment.

Representation

To measure similarities between two amino acid sequences, we need to somehow group the hydrophobic amino acid into clusters. This can be done by first convert an amino acid string sequence into binary number. All hydrophobic amino acids (7 types: Valine (V), Isoleucine (I), Leucine (L), phenylalanine (F),

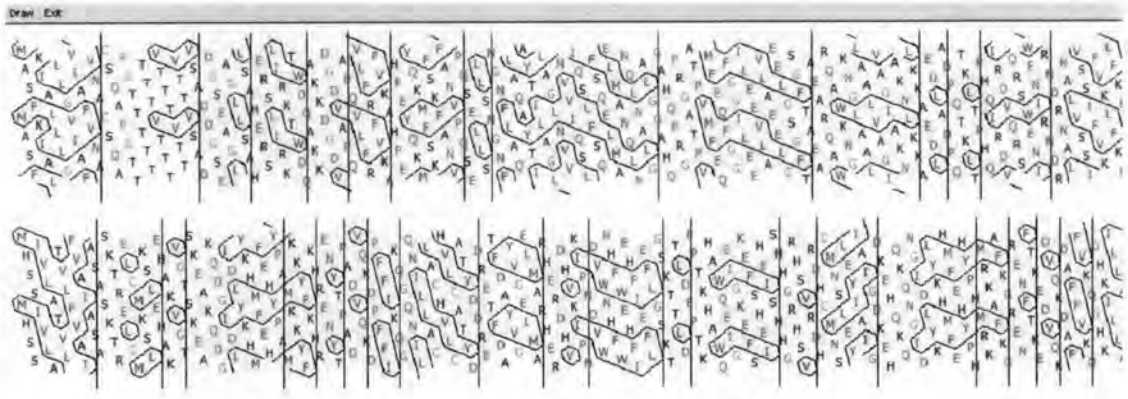


Fig. 2. HCA output samples of the amino acid sequence from our visualization tool. (Top) Hemoglobin amino sequence of *Daphnia magna* (Arthropod, insect); (Bottom) Hemoglobin amino acid sequence of *Pseudoterranova decipiens* (Nematod, worm).

tryptophan (W), methionine (M) and tyrosine (Y)) will be replaced by 1, and all others will be replaced by 0, as illustrated in Fig.3 (b) To identify and split the hydrophobic into individual clusters, we will locate the position where there are 5 or more consecutive 0's. If the number of consecutive 0's is fewer than 5, the 1's on both left and right of these 0's will be combined into one same cluster; a sequence of 1's without any 0's in between is definitely belong to the same cluster . In addition, special treatment needs to be done when a symbol "P" or "Proline" is encountered, by immediately completing its current hydrophobic cluster (if any) and starting a new one. Fig. 3(c) shows the cluster boundaries; consecutive 1's indicate the membership of the corresponding hydrophobic amino acid that belongs in the same cluster.

Once the amino acids have been transformed into binary symbols and clustered into individual clusters, we have to make sure that these data are in appropriate form that can be effectively used for an alignment. Note that all the research up to this point only knows how to cluster HCA; no further research has been done to utilize this information for an automatic 2-D alignment. In fact, the latest work on exploiting HCA for protein homology was dated back in 1990 [6], but perhaps the computer technologies back then were not powerful enough to feasibly solve the protein homology problem. So, this problem has been put aside and seemed to be overlooked by today's researchers, which is our main motivation for this work in an attempt to make 2-D automatic alignment possible.

In this work, we propose an HCA representation that will become a building block in the Hydrophobic Cluster alignment algorithm. An ideal representation should have the following properties: (i) ability to identify and display hydrophobic cluster groupings; (ii) ability to remove unnecessary/redundant information (e.g., the hydrophilic amino acid that links

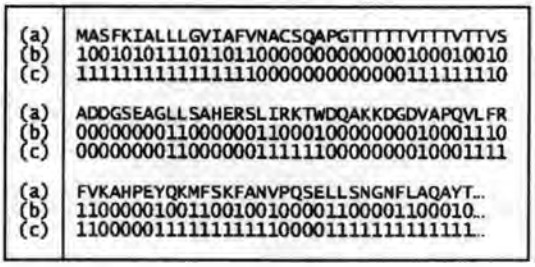


Fig. 3: Hydrophobic cluster representation that has been segmented

the hydrophobic clusters together) from the data; and (iii) preservation of hydrophobic cluster sequence must always hold. Consequently, our proposed representation will take only the identified hydrophobic clusters while preserving all the original cluster ordering; each cluster is separated by a comma ",". Fig. 4 (a) shows the haemoglobin protein representation of *Daphnia magna*, and Fig. 4 (b) shows the haemoglobin protein representation of *Pseudoterranova decipiens* using our proposed method.

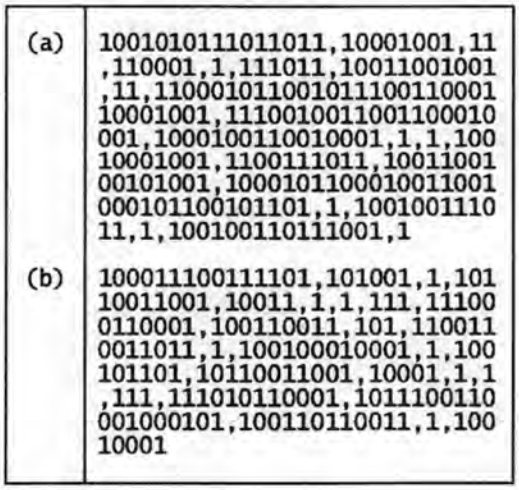


Fig. 4. Hydrophobic cluster representation that has been segmented.

The feature or shape of each segmented cluster representation can be easily recognized and identified, as shown in Fig. 4, using our visualization tool, which can easily transform back an forth between the representation and the hydrophobic cluster visualization. From Fig. 5, we can see that the representations of

examples 2 and 3 are 110001 and 11, respectively, but their hydrophobic cluster visualizations reveal that these two clusters differ by only 1 position. Our representation also supports the homology measurement, by looking at the distance between "1"; if the positions of "1" in the two amino acid sequence clusters are the same, those clusters will have high similarity. Similarly, the representations of examples 1 and 5 from Fig. 5 can be used to find the 2-D alignment and homology within the clusters; we can see that the cluster in example 1 is a subset of the cluster in example 5, i.e. example 5 contains 2 of the example 1 pattern, and for each match, there are 3 similar positions and 2 dissimilar positions, as illustrated in Fig. 6.

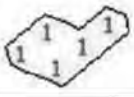
Example	Representation	Hydrophobic Cluster
1	10001001	
2	110001	
3	11	
4	111011	
5	10011001001	

Fig. 5: Hydrophobic Cluster visualization that corresponds to the proposed representation.

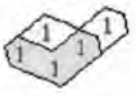
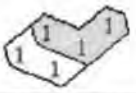
Example	Representation	Hydrophobic Cluster
5	10011001001	
1	10001001	
5	10011001001	
1	10001001	

Fig. 6: Alignment and similarity measurement between two hydrophobic clusters.

Conclusion and Future work

In this work, we have proposed a new hydrophobic cluster representation that satisfies all the requirements and can then be used in the 2-D automatic hydrophobic cluster alignment tool. We strongly believe that this 2-D alignment approach will be superior to the existing 1-D alignment in terms of similarity accuracy among homologous proteins. Our future work will include a development and implementation of the an easy-to-use automatic hydrophobic cluster alignment software that could facilitate and unveil the new avenue of research in protein's homology and functionality.

References

[1] Needelmann, S.B. and Wunsch, C.D. (1969), *J. Mol. Biol.*, **48**, p 443-453

[2] Park, J., Karplus, et al. (1998), *J. Mol. Biol.*, **284**, p 1201-1210.

[3] Altschul, S.F., Gish, W., et al. (1990), *J. Mol. Biol.*, **215**, p 403-410

[4] Gaboriaud, C., Bissery, V., et al. (1987), *FEBS Lett.*, **224**, p 149-155

[5] Ruth T.S., Rath P., et al. (1999), *Virology*, **259**, p 20-33

[6] Lemesle-Varloot L., Gaboriaud, C., et al. (1993), *Comput Appl Biosci.*, **9**, p 37-44

วารสารทางวิชาการในระดับนานาชาติ "International Journal of Biomedical Engineering and Technology (IJBET)" ในหัวข้อเรื่อง "Automatic 2-D Hydrophobic Cluster Alignment" สำนักพิมพ์ Inderscience โดย ภิสิทธิ์ กรรณสูต รัฐ พิชญางกูร และ โชติรัตน์ รัตนามหัทธนะ ได้รับการตอบรับการตีพิมพ์ในวารสารเมื่อวันที่ 19 กันยายน 2550 ตีพิมพ์ในปี 2009

จดหมายตอบรับการตีพิมพ์เมื่อวันที่ 19 กันยายน 2550

From: Jérôme Darmont <jerome.darmont@univ-lyon2.fr>

Date: Nov 19, 2007 11:10 PM

Subject: IJBET special issue notification

Dear Ann,

We are pleased to inform you that your paper "DM03 - 2-D automatic hydrophobic cluster alignment" has been accepted for publication in the upcoming special issue "Warehousing and Mining Complex Data: Applications to Biology, Medicine, Behavior, Health and Environment" of the International Journal of Biomedical Engineering and Technology.

Please find attached to this e-mail your paper reviews. Study the comments carefully and incorporate the changes into your final version. Please follow strictly the guidelines from the attached CfP (page 2) for preparing your final paper, and e-mail it to bdd@eric.univ-lyon2.fr before December 15, 2007 (please indicate your paper ID in subject) as well as a separate file with your response to the reviewers' comments. Should you have any questions regarding your final paper, please do not hesitate to contact us. We look forward to receiving your contribution.

Sincerely,

The guest editors

Jérôme Darmont, ERIC, Université Lumière Lyon 2

Bât. L, 5 Avenue Mendès-France, 69676 Bron Cedex, France

Tél. +33 478-774-403 mailto:jerome.darmont@univ-lyon2.fr

http://eric.univ-lyon2.fr/~jdarmont/

Int. J. Biomedical Engineering and Technology, Vol. x, No. x, xxxx 1

Automatic 2-D Hydrophobic Cluster Alignment

Phisit Kannasut

Department of Computer Engineering,
Chulalongkorn University, Bangkok, Thailand
E-mail: g49pks@cp.eng.chula.ac.th

Rath Pichyangkura

Department of Biochemistry,
Chulalongkorn University, Bangkok, Thailand
E-mail: prath@chula.ac.th

Chotirat Ann Ratanamahatana

Department of Computer Engineering,
Chulalongkorn University, Bangkok, Thailand
E-mail: ann@cp.eng.chula.ac.th

Abstract: Biological databases in the past decade have tremendously grown in size. However, effective retrieval of these data is still a great challenge. In particular, we need a high-quality tool to measure similarity among protein sequences within the database. Current techniques in protein homology testing involve a 1-dimensional alignment of Nucleotide or Amino acid sequencing. Due to its various constraints and low sequence identity values, Hydrophobic Cluster Alignment has increasingly been used to predict the structure and functionality of protein. However, this Hydrophobic Cluster Alignment still needs to be done manually and solely depends on experience and expertise of a researcher. In this work, we implement a 2-D visualization tool for amino acid and propose a new protein representation that could be used effectively in our 2-D automatic Hydrophobic Cluster Alignment software. The 2-D alignment was performed on the amino acid sequences from the PIR database. The resulting evolution trees and the alignment results have demonstrated that homologous proteins with same functionalities obtain high sequence identity score, whereas ones with different functionalities obtain low sequence identity score as anticipated.

Keywords: bioinformatics; hydrophobic cluster analysis; protein homology; automatic alignment

Reference to this paper should be made as follows: Kannasut, P., Pichyangkura, R., and Ratanamahatana, C.A. (2007) '2-D Automatic Hydrophobic Cluster Alignment', *Int. J. Biomedical Engineering and Technology*, Vol. x, No. x, pp.xx-xx.

Biographical notes: P. Kannasut is a Master's student at the Department of Computer Engineering, Chulalongkorn University, Bangkok, Thailand. His current research interest includes similarity tools to identify homologous proteins.

R. Pichyangkura received his Ph.D. from Michigan State University. He is currently a lecturer at the Department of Biochemistry, Chulalongkorn University, Bangkok, Thailand. His current research interests include molecular biology, cloning and enzyme engineering.

C.A. Ratanamahatana received her B.S. and Ph.D. in Computer Science from Carnegie Mellon University and University of California, Riverside, respectively. She is currently a lecturer at the Department of Computer Engineering, Chulalongkorn University, Bangkok, Thailand. Her current research interests include Time series Data mining, Machine Learning, and Artificial Intelligence.

1. Introduction

Life science and biological data have grown very rapidly in the past few years. This includes Nucleotide/Amino acid sequences available on GenBank (Benson et al. 2006) and PIR protein database (Barker et al. 1999) , the crucial data for protein analyses, e.g., protein homology, structure (Primary, Secondary, Tertiary, or Quaternary), or even functionalities of proteins. One of the largest sources of data is collected from various genome projects, which include nucleotide sequences and amino acid sequences. These data are crucial in analyses and extractions of knowledge to help us understand rules of nature. However, effective retrieval of these data is still a great challenge. In particular, we need a high-quality tool to discover protein homology within the database via sequence alignment.

Any form of protein could consist of as many as 20 different amino acids (encoded in DNA) connected with peptide bonds. Protein must be first extracted from the cells or organisms before we can study its structure, functionality, or identify its amino acid sequence. A renowned project such as the Human genome project has led to an emergence of protein and amino acid sequence database.

However, the fact that each protein consists of up to 20 different amino acids in random order makes an analysis of protein homology particularly difficult. The current methods are mostly based on Maximum Matching (Needleman and Wunsch 1970), Hydrophobic Cluster Analysis (HCA), or Basic Alignment Search Tool (BLAST) (Altschul et al. 1990), all of which still have major limitations and drawbacks; they could not provide a proof of sequence homology if their sequence identity appears to be too low, which usually occur in proteins from different species (but the proteins themselves have the same functionality).

The similarity can generally be measured from an alignment of either nucleotide sequences or amino acid sequences. However, nucleotide sequence similarity is not suitable for protein function discovery; amino acid sequences are typically exploited instead since they contain much more information, such as hydrophobic and hydrophilic properties, etc. Unfortunately, current 1-dimentional alignment tool does have some limitation that yields poor alignment results.

In this work, we are proposing a tool to help identify amino acid sequence from different organisms that has similar functionality. Our proposed method is based on the idea of Hydrophobic Cluster Analysis (Lemesle-Varloot et al. 1990), using 2-dimensional diagram that highlights the hydrophobic elements. We then propose a new protein representation that preserves the crucial features that are essential for an effective and meaningful alignment.

2. Background

2.1. Protein Homology Testing Tools

Several tools and methods have been proposed to determine the protein homology, such as Dynamic programming methods (Needleman and Wunsch 1970), BLAST (Altschul et al. 1990), Hidden Markov models (Park et al. 1998), etc. Among these, the dynamic programming method has long been known and very popular nowadays even though it still could not achieve high accuracies. The algorithm starts by creating a matrix/array that contains the symbols in each amino acid string. To find the similarity, another amino acid array will be subsequently compared and matched until the best alignment is achieved, as shown in Figure 2.

The figure represents two protein samples, the top sequence is the hemoglobin of *Daphnia magna*, an organism classified as arthropod phylum (insect), and the bottom sequence is the hemoglobin of *Pseudoterranova decipiens*, an organism classified as nematode phylum (worm). These two creatures are both physically and genetically very dissimilar, which can be seen by different amino acid structures, creating large and various gaps occurrences. This method therefore is not an effective nor suitable method in discovering protein homology; even though both are the same hemoglobin protein, their amino acid sequencing similarity obtained from this dynamic programming approach yields only 36% similarity and merely 17% in sequence identity. ClustalW (Thompson et al. 1994) is another program that is used for multiple alignments (1-dimentional). Its result is similar to that of Figure 2, i.e. Dynamic programming method also cannot achieve meaningful alignment based on hydrophobic elements.

Another approach in protein homology is BLAST – Basic Local Alignment Search Tool – a statistical and knowledge based tool, which was developed in 1980 and still is a tool of choice to most researchers nowadays. However, this approach still faces the same problem as the dynamic programming approach that it is unable to ***realistically*** evaluate the similarity between two protein sequences. Researchers are well aware of this shortcoming and are drawn to the new idea of Hydrophobic Cluster Analysis (HCA), an extremely powerful tool to understand protein' stability and its folding by comparing and analyzing amino acid sequences (Henrissat et al 1988). Unfortunately, this method is not very widespread since it still needs the eyes of an expert to evaluate the similarity of the created chart. However, HCA has later been used widely in the purpose of secondary protein structure prediction.

MANSEK and SUNHCA are the software that constructs HCA of protein sequences, working on Vax and Sun platforms, respectively. They have been further developed to allow on-screen manipulation (zoom, translation, etc.), as well as plotting on papers

(Lemesle-Varloot et al. 1993). However, neither of them has a function to support the protein homology matching.

Subsequently, some researches have started to utilize HCA into the study of protein structure. Henrissat plots out the HCA diagram and discover the structure that is a c-terminal amino acid sequence of wheat (*Triticum aestivum*). This research reveals that there exist some typical domain structure where (i) variable and conserved domains are located along the sequence at precise positions (ii) the conserve domains characteristically reflect a common ancestor, and (iii) the unique properties of a given protein are associated with the variable domain.

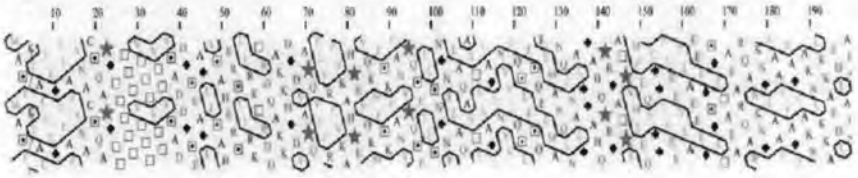
2.2. Hydrophobic Cluster Analysis

In this work, we have adopted an HCA approach and then illustrated amino acid sequences into a 2-D helical pattern by twisting the protein into a smoothed helix, where each twist will contain 3.6 amino acids (Gaboriaud et al. 1987). Then, this cylinder is cut lengthwise and spread into a 2-dimensional plane, and the hydrophobic amino acid will be highlighted and grouped together (Lemesle-Varloot et al. 1990), as shown in Figure 2.

In the past, some researchers have used this method in an application of protein alignment. However, once a 2-dimensional HCA visualization is obtained, the rest of the work was handed over to the human expert to do the actual hydrophobic cluster alignment manually.

Our work attempts to come up with a way to perform an alignment automatically without human intervention. There are several ways that an alignment could be done, such as image recognition, pattern recognition, string matching, etc. For faster and more accurate alignment, we prefer a string-matching-based technique and avoid an image processing approach. So, we propose our new representation that can be later used for an effective and efficient 2-D alignment.

Figure 1 Two-dimensioanal representation of the amino acid sequence using HCA method.



2.3. Earlier Representation

To measure similarities between two amino acid sequences, we need to somehow group the hydrophobic amino acid into clusters. This can be done by first convert an amino acid string sequence into binary number. All hydrophobic amino acids (7 types: V, I, L, F, W, M, and Y) will be replaced by 1, and all others will be replaced by 0, as shown in Fig 3 (b). Once the binary representation of the Hydrophobic amino acid is obtained, it is then used in the HCA visualization software (similar to Figure 2). However, this binary representation is not meaningful enough to identify similar protein patterns. In this work, we extend and modify this representation further such that it can be used effectively in the 2-Dimensional hydrophobic cluster alignment.

2.4. Hydrophobic representation

Protein's functionalities mainly depend on its 3-Dimensional structure. Therefore, most proteins would try their best to preserve the structure by revolving in the hydrophobic part facing each other and revolving the hydrophilic part outside for easy access to water.

From previous studies, there are as many as 40 types of Hydrophobicity metrics (Needleman and Wunsch 1970), but we decided to use a more accurate scale proposed by G. Von Heijine (Heijne 1986) shown in Table 1.

According to the Hydrophobicity and other properties, we can classify amino acid into 7 groups, each of which has been assigned its color, as illustrated in Table 2

3. Our Proposed Method

3.1. New representation

We extend the G. Von Heijine metric's (Heijne 1986) representation into hydrophobic cluster representation. To identify and split the hydrophobic into individual clusters, we locate the position where there are 5 or more consecutive 0's. If the number of consecutive 0's is fewer than 5, the 1's on both left and right of these 0's will be combined into one same cluster; a sequence of 1's without any 0's in between is definitely belong to the same cluster. In addition, special treatment needs to be done when a symbol "P" or "Proline" is encountered, by immediately completing its current hydrophybic cluster (if any) and starting a new one. Figure 3 (c) shows the cluster boundaries; consecutive 1's indicate the membership of the corresponding hydrophobic amino acid that belongs in the same cluster.

Once the amino acids have been transformed into binary symbols and clustered into individual clusters, we have to make sure that these data are in an appropriate form that can be effectively used for an alignment. Note that all the research up to this point only knows how to cluster HCA; no further research has been done to utilize this information for an automatic 2-D alignment. In fact, the latest work on exploiting HCA for protein homology was dated back in 1990 (Lemesle-Varloot et al. 1993), but perhaps the computer technologies back then were not powerful enough to feasibly solve the protein homology problem. So, this problem has been put aside and seemed to be overlooked by today's researchers, which is our main motivation for this work in an attempt to make 2-D automatic alignment possible.

Our newly proposed representation will become a building block in the Hydrophobic Cluster alignment algorithm. An ideal representation should have the following properties: (i) ability to identify and display hydrophobic cluster groupings; (ii) ability to remove unnecessary/redundant information (e.g., the hydrophilic amino acid that links the hydrophobic clusters together) from the data; and (iii) preservation of hydrophobic cluster sequence must always hold.

Consequently, our proposed representation will take only the identified hydrophobic clusters while preserving all the original cluster ordering; each cluster is separated by a comma ",". Figure 4 (a) shows the hemoglobin protein representation of *Daphnia magna* (P.Arthropoda), and Figure 4 (b) shows the hemoglobin protein representation of

Pseudoterranova decipiens (P.Nematoda) using our proposed method. Note that both proteins do have the same functionality of oxygen transfer.

The feature or shape of each segmented cluster representation can be easily recognized and identified, as shown in Figure 5, using our visualization tool, which can easily transform back an forth between the representation and the hydrophobic cluster visualization. From Figure 5, we can see that the representations of examples 2 and 3 are 110001 and 11, respectively, but their hydrophobic cluster visualizations reveal that these two clusters differ by only 1 position regardless of the string representations that may somewhat have different lengths. Our representation also supports the homology measurement, by looking at the distance between "1"; if the positions of "1" in the two amino acid sequence clusters are the same, those clusters will have high similarity. For example, the representations of examples 1 and 5 from Figure 5 can be used to find the 2-D alignment and homology within the clusters; we can see that the cluster in example 1 is a subset of the cluster in example 5, i.e., example 5 contains 2 of the example 1 pattern, and for each match, there are 3 similar positions and 2 dissimilar positions, as illustrated in Figure 6. Each of these hydrophobic cluster representations is called a "cluster block."

3.2. Visualization tool

In the process to achieve automatic alignment, we have first implemented a visualization tool on Microsoft Windows environment, using C# .net programming language. Its main functionality is to draw HCA, group Hydrophobic cluster, and draw lines to indicate the boundaries of the clusters, which will be essential for the automatic alignment algorithm. We also apply the standard color codes to our drawing (See Table 2), while the specialized characters (P,G,T,S) are preserved and not transformed to other symbols as in Lemesle-Varloot L's style (Lemesle-Varloot et al. 1993). An output sample from our visualization tool is illustrated in Figure 7

3.3. Cluster alignment program

At this point, we obtain the representation, consisting of cluster block boundaries, from the protein sequences. Then we design and implement a cluster alignment program based on 2-dimensional string matching technique and dynamic programming technique. Cluster blocks from the first sequence are compared with cluster blocks from another sequence in order to search for an optimal alignment. Our algorithm is shown in Figure 8.

From Figure8 *Rep1* and *Rep2* are the whole representations of amino acid sequence 1 and sequence 2, respectively. Every cluster blocks in *Rep1* will be paired with every cluster blocks in *Rep2* in order to search for an optimal alignment, based on a dynamic programming technique.

Our algorithm searches for the highest score between the cluster blocks, *Clus1* and *Clus2*, using HCA_Align function, obtaining three output values, *tmpScore*, *tmpEndResidue*, and *tmpK*, which are the maximum score between *Clus1* and *Clus2*, the ending position of aligned residue in both cluster blocks, and the alignment path corresponding to the achieved maximum score, respectively. This program uses matrix *K* to trace back for the best alignment position.

Figure 9 An example of our proposed cluster alignment algorithm

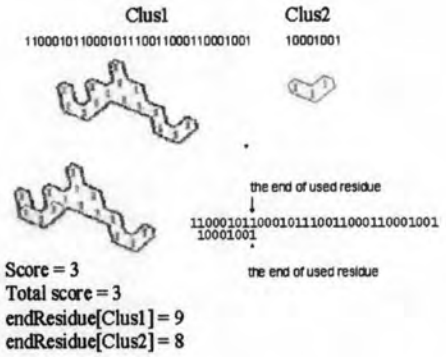
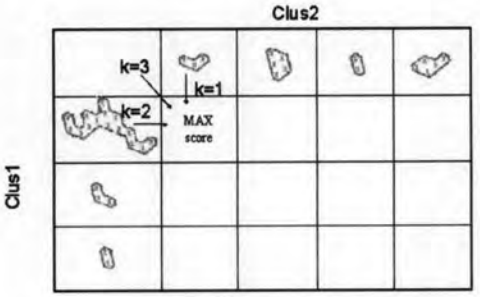


Figure 9 shows an example of two cluster blocks, *Clus1* and *Clus2*, which is used in an alignment. The maximum alignment score is 3, which adds to the initial total score (0) and becomes 3. The ending positions of the aligned residues of each cluster block are also updated; in this example, the ending residue position of *Clus1* is 9 and the ending residue position of *Clus2* is 8.

We use a string matching algorithm based on a dynamic programming approach for similarity matching. To calculate the best score, we first construct a matrix of m by n, where m is the number of cluster blocks in *Rep1* and n is the number of cluster blocks in *Rep2*. Each *Clus1* of *Rep1* will be paired with each *Clus2* of *Rep2*. Each cell in the matrix is a cumulative distance of the maximum score from the three neighboring cells and the score in the current cell (calculate similarly to the example in Figure 9). The three neighboring cells are one from the top (denoted by $k = 1$), one from the left (denoted by $k = 2$), and one (diagonally) from top left (denoted by $k = 3$). An example is shown in Figure 10. We start filling in the matrix from cell (1,1) and finally, the last cell, (m, n) will contain the optimal alignment score.

Figure 10 Our cluster alignment algorithm based on Dynamic Programming approach



Our pseudo codes are provided in Figure 11 and 12. Figure11 computes the maximum alignment score between each pair of cluster blocks based on the cumulative scores from the 3 neighboring directions as mentioned earlier. It also has a sub function called *alignment* that takes care of the alignment at the current cell. If one of the cluster blocks is used up in the previous alignment, the function is terminated and moved on to the next cluster block pair.

Figure 13 A snapshot of the values in the matrices used in the alignment program

endResidue[Clus1]

	1	2	3	4
1	9			
2				
3				

endResidue[Clus2]

	1	2	3	4
1	8			
2				
3				

score	1	2	3	4
1	3			
2				
3				

k	1	2	3	4
1	3			
2				
3				

Figure 13 shows a snapshot of the matrices that store the alignment score, the ending residue positions, and the path's direction (*k*) after the first round of the calculation (see Figure 9).

The alignment algorithm continues with the new cluster block pair, as demonstrated in Figure 14.

Figure 14 The score from 3 directions horizontal (k=1), vertical (k=2), and diagonal direction (k=3)

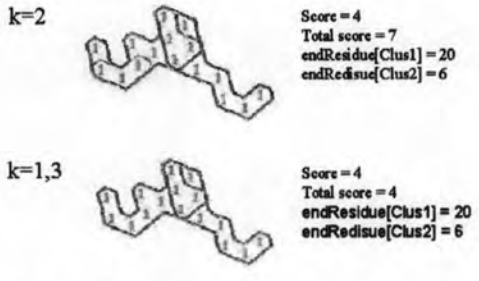


Figure 14 (top) shows the accumulated score from the left direction (k=2). The previous score from the left is 3 (from the score matrix) and the current score is 4, giving the total cumulative score of 7. However, since there is no previous alignments or accumulative scores in the top (k=1) and diagonal (k=3) directions, the total score will just be the alignment at the current cell (Figure 14 (bottom)).

Hence, the maximum score of 7 is selected, as appropriate values of ending residue positions and the direction matrices are updated accordingly, as shown in Figure 15

Figure 15 The updated matrix

endResidue[Clus1]

	1	2	3	4
1	9	20		
2				
3				

endResidue[Clus2]

	1	2	3	4
1	8	6		
2				
3				

score	1	2	3	4
1	3	7		
2				
3				

k	1	2	3	4
1	3	2		
2				
3				

The final score will be obtained when the matrices are filled up; the last cell at position (m, n) of the score matrix will contain the alignment score and the k matrix is used to retrieve all the alignment path, starting from (m, n) and tracing back to position (1, 1).

To evaluate the quality of the sequence alignment in Bioinformatics, some metric must be assessed. In this work, a sequence identity (Lemesle-Varloot et al. 1990) will be calculated to measure how similar the two sequences are. The formula is defined below.

$$Identity = \frac{(2 \times FinalScore)}{(Haa1 + Haa2)} \times 100$$

where $Haa1$ is a number of Hydrophobic amino acid in sequence 1, $Haa2$ is a number of Hydrophobic amino acid in sequence 2, and the *FinalScore* is the alignment score obtained from our program. If the two sequences are very similar, in that every Hydrophobic amino acid of one sequence can be aligned with ones in another sequence, our alignment score will correspond to the number of Hydrophobic amino acid in both sequences, giving the *Identity* score to be 100% similar.

To visualize the result, especially among many types of proteins, we build an evolutionary tree or "Distance Tree," that has been modified a little to suit our distance calculation. The distance measure we use in this work is defined as the number of the Hydrophobic amino acids that are left from the alignment process, shown in the subsequent formula.

$$Dist = (Haa1 + Haa2) - (2 \times FinalScore)$$

If the program can align every single Hydrophobic amino acid between the two sequences, *Dist* value would become zero, which means that these two sequences are extremely similar.

3.4. Validation test

To validate our proposed alignment algorithm, we tested our program on 62 types of protein, classified into more than 2 classes (phosphatase protein, ligase protein, etc.); the data was retrieved from the PIR database (Proclass). The test data was selected in such a way that each protein class would contain various species, such as human, mouse, insect, bacteria, fungi, etc. The lengths of data are between 300 - 330 residues. A full list of protein is shown in Table 3.

We compare our experiment results with ones obtained from the current alignment method, using ClastalW program, which is an alignment program based on string matching and score lookups between residues, using a distance matrix (e.g. Blosom62 matrix).

After the distance matrices from the two methods are obtained, we analyze by looking at their evolutionary tree, using TreeView v.1.6.6.0 program (ref. TreeView), a Neighbor-joining method, to generate a tree.

The result of the current method is shown in Figure 16, and the result of our proposed method is shown in Figure 17, both using TreeView 1.6.6.0 visualization tool.

It can be readily seen in Figure16 that the current ClustalW algorithm is unable to categorize various proteins into correct classes, i.e. various Phosphatase and Ligase proteins are misclassified. Moreover, some classes that are supposed to be very dissimilar are group together into the same node (e.g. P23635 and Q8C5S2). It is apparent that the current method cannot produce meaningful similarity scores when the similar-function proteins are from different organism.

Figure 17 shows the evolutionary tree computed by our proposed method and viewed by TreeView program. It is apparent that the two Phosphatase and Ligase protein classes are classified correctly. More specifically, within the Phosphatase protein class itself, it contains protein groups that are also classified correctly since they have the same functionalities even though they come from different organisms.

4. Conclusion and future work

In this work, we propose an effective and automatic 2-D hydrophobic cluster alignment algorithm, which can generally be applied to numerous Bioinformatics applications, e.g., discovering functionality of proteins, finding protein domains, predicting similar structure, finding protein evolution, etc. We have demonstrated the utility of our proposed algorithm and visualization tools. The experiment results reinforce our claim that the 1-D alignment is inappropriate for the purpose of protein homology discovery while 2-D hydrophobic cluster alignment is much more effective and accurate. Our future work will include a development and implementation of a more user friendly software that could facilitate and unveil the new avenue of research in protein's homology and functionality.

References

Altschul, Stephen F., et al. (1990), 'Basic local alignment search tool', *J Mol Biol*, Vol. 215, No. 3, pp. 403-10.

Barker, W. C., et al. (1999), 'The PIR-International Protein Sequence Database', *Nucleic Acids Res*, Vol. 27, No. 1, pp. 39-43.

Benson, D. A., et al. (2006), 'GenBank', *Nucleic Acids Res*, Vol. 34, pp. D16-20.

Breton, C., Oriol, R., Imberty, A. (1997) Conserved structural features in eukaryotic and prokaryotic fucosyltransferases. *Glycobiolog* Vol. 8, pp. 87-94

Callebaut, I., Courvalin, JC., Worman, HJ., Mornon, JP. (1997) Hydrophobic cluster analysis reveals a third chromodomain in the Tetrahymena Pdd1p protein of the chromo superfamily. *Biochem Biophys Res Commun*, Vol. 235, pp. 103-7

Callebaut, I., Prat, K., Meurice, E., Mornon, JP., Tomavo, S. (2005) Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. *BMC Genomics*. Vol. 6, pp. 100

Gaboriaud, Christine, et al. (1987), 'Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences', *FEBS Lett*, Vol. 224, No. 1, pp. 149-55.

Geremia, RA., Petroni, EA., Ielpi, L., Henrissat, B. (1996) Towards a classification of glycosyltransferases based on amino acid sequence similarities: prokaryotic alpha-mannosyltransferases. *Biochem J*. Vol. 318, pp. 133-138

Girardeau, JP., Bertin, Y., Callebaut, I. (2000) Conserved structural features in class I major fimbrial subunits (Pilin) in gram-negative bacteria. Molecular basis of classification in seven subfamilies and identification of intrasubfamily sequence signature motifs which might Be implicated in quaternary structure. *J Mol Evol*. Vol. 50, pp. 424-42

Krishna, SS., Sadreyev, RI., Grishin, NV. (2006) A tale of two ferredoxins: sequence similarity and structural differences. *BMC Struct Biol*. Vol. 6, pp, 8

Heijne, Gunnar von (1986), 'The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology', *Embo J*, Vol. 5, No. 11, pp. 3021-27.

Henrissat, B., Popineau, Y., Kader, JC. (1988) Hydrophobic-cluster analysis of plant protein sequences. A domain homology between storage and lipid-transfer proteins. *Biochem J*. Vol. 255, pp. 901-905

Lemesle-Varloot, L., et al. (1990), 'Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequences', *Biochimie*, Vol. 72, No. 8, pp. 555-74.

Lemesle-Varloot, L., et al. (1993), 'MANSEK and SUNHCA. Two interactive programs for the hydrophobic cluster analysis of protein sequences', *Comput Appl Biosci*, Vol. 9, No. 1, pp. 37-44.

Needleman, Saul B. and Wunsch, Christian D. (1970), 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *J Mol Biol*, Vol. 48, No. 3, pp. 443-53.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol*, Vol. 284, pp. 1201-1210.

Thompson, JD., Higgins, DG., Gibson, TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, Vol. 22, pp. 4673-4680

TreeView V 1.6.6.0 http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Figure 2 A 1-demensional alignment of amino acid sequences using Dynamic programming method

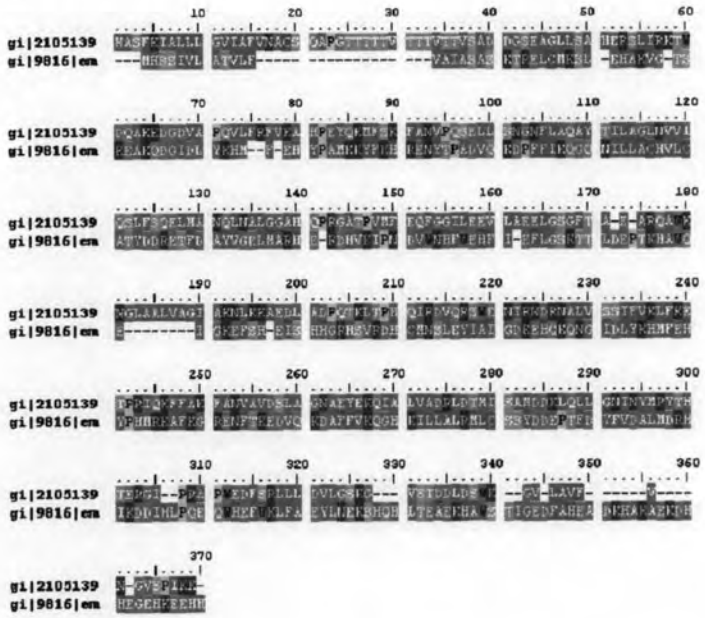


Figure 3 The presence/absence of Hydrophobic amino acid is transformed into binary numbers, represented by 1's and 0's, respectively. (a) amino acid sequence; (b) hydrophobic amino acid clusters–previous representation; (c) our proposed representation that shows cluster boundaries

```
(a)  MASFKIALLLGVIAFVNACSQAPGTTTTTVTTTVTTVSADDGSEAGLLSAHERSLIRKTWD
(b)  10010101110110110000000000000010001001000000000110000001100010
(c)  11111111111111110000000000000011111111000000000110000001111110

(a)  QAKKDGDVAPQVLFRFVKAHPEYQKMFSKFANVPQSELLSNGNFLAQAYTILAGLNVVIQS
(b)  00000001000111011000001001100100100001100001100010110010111100
(c)  00000001000111111000001111111111110000111111111111111111111111

(a)  LFSQELMANQLNALGGAHQPRGATPVMFEQFGGILEEVLAEELGSGFTAEARQAWKNGLAA
(b)  11000110001001000000000011100100110011000100010000001000100
(c)  11111111111111000000000001111111111111111111111100000001111111

(a)  LVAGIAKNLKKAEDLADPQTKLTPHQIRDVQRSWENIRNDRNALVSSIFVKLFKETPR...
(b)  11001000100000100000010000100100010010000001100111011000000...
(c)  11111111100000100000010000111111111110000001111111111100000...
```

Figure 4 Hydrophobic cluster representation that has been segmented

(a)	1001010111011011,10001001,11,110001,1,111011,10011001001,11,110001011001011100110001100011001,1110010011001100010001,1000100110010001,1,1,10010001001,1100111011,1001100100101001,1000101100010011001000101100101101,1,100100111011,1,100100110111001,1
(b)	1000111001111101,101001,1,10110011001,10011,1,1,111,111000110001,100110011,101,1100110011011,1,100100010001,1,100101101,10110011001,10001,1,1,111,111010110001,10111001100010000101,100110110011,1,10010001

Figure 5 Hydrophobic Cluster visualization that corresponds to the proposed representation

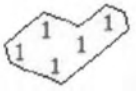
Example	*Representation*	*Hydrophobic Cluster*
1	10001001	
2	110001	
3	11	
4	111011	
5	10011001001	

Figure 6 Alignment and similarity measurement between two hydrophobic clusters

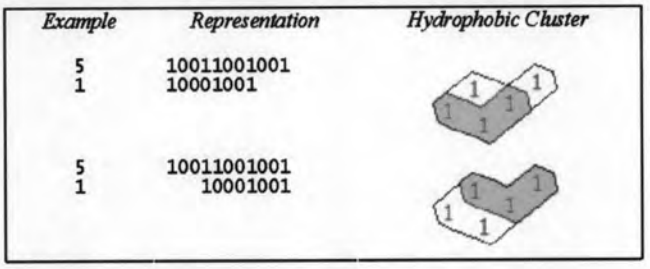


Figure 7 HCA output samples of two amino acid sequences from our visualization tool. (Top) Hemoglobin amino acid sequence of *Daphnia magna* (insect); (Bottom) Hemoglobin amino acid sequence of *Pseudoterranova decipiens* (worm). Hydrophobic cluster blocks and their boundaries are also identified.

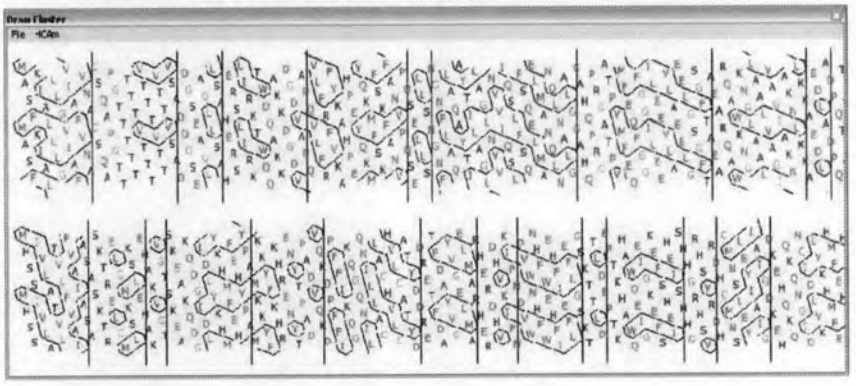


Figure 8 Our cluster alignment algorithm

```
Rep1 = Representation from sequence 1
Rep2 = Representation from sequence 2
for (Clus1 = Cluster block from i to m in Rep1)
  for (Clus2 = Cluster block from j to n in Rep2)
    tmpScore = 0
    (tmpScore, tmpEndResidue. tmpK) =
      HCA_Align(Clus1, Clus2, i, j, endResidue, Score)
    Score[i][j] = tmpScore
    endResidue[i][j] = tmpEndResidue
    K[i][j] = tmpK
  end for
end for

Trace = RetrievePath(K)
FinalScore = score[m][n]
```

Figure 11 HCA_Align algorithm

```
function HCA_Align(Clus1,Clus2,i,j,endResidue,Score)
//Dynamic Programming

//from vertical (top) direction
tmpEndResidue1 = endResidue[i][j-1]
(tmpScore1,tmpEndResidue1) =
  alignment(tmpEndResidue1,Clus1,Clus2)
tmpScore1 += Score[i][j-1]

//from horizontal (left) direction
tmpEndResidue2 = endResidue[i-1][j]
(tmpScore2,tmpEndResidue1) =
  alignment(tmpEndResidue2,Clus2,Clus1)
tmpScore2 += Score[i-1][j]

//from diagonal direction
tmpEndResidue3 = 0
(tmpScore3,tmpEndResidue3) =
  alignment(tmpEndResidue3,Clus1,Clus2)
tmpScore3 += Score[i-1][j-1]

if(tmpScore3 is maximum)
  tmpScore = tmpScore3
  tmpEndResidue = tmpResidue3
  tmpK = 3
elseif(tmpScore2 is maximum)
  tmpScore = tmpScore2
  tmpEndResidue = tmpEndResidue2
  tmpK = 2
else
  tmpScore = tmpScore1
  tmpEndResidue = tmpEndResidue1
  tmpK = 1
endif

return (tmpScore,tmpEndResidue,tmpK)
```

Figure 12 Function alignment algorithm

```
function alignment(tmpEndResidue,ClusA,ClusB)
tmpScore = 0
if(tmpEndResidue < length(ClusA))
  (ClusA,ClusB) = setStartPosition(ClusA,ClusB)
  while(compare all possible pattern)
    newScore = compareScore(Clus1,Clus2);
    newEndResidue =
      findEndResidue(Clus1,Clus2);
    if(newScore>tmpScore)
      tmpScore = newScore
      tmpEndResidue = newEndResidue
    elseif(newScore = tmpScore)
      tmpEndResidue[ClusB] = newEndResidue[ClusB]
    Endif
  endwhile
endif

return (tmpScore,tmpEndResidue)
```

Figure 16 The evolution tree is computed by using ClastalW Method and represent by TreeView 1.6.6.0

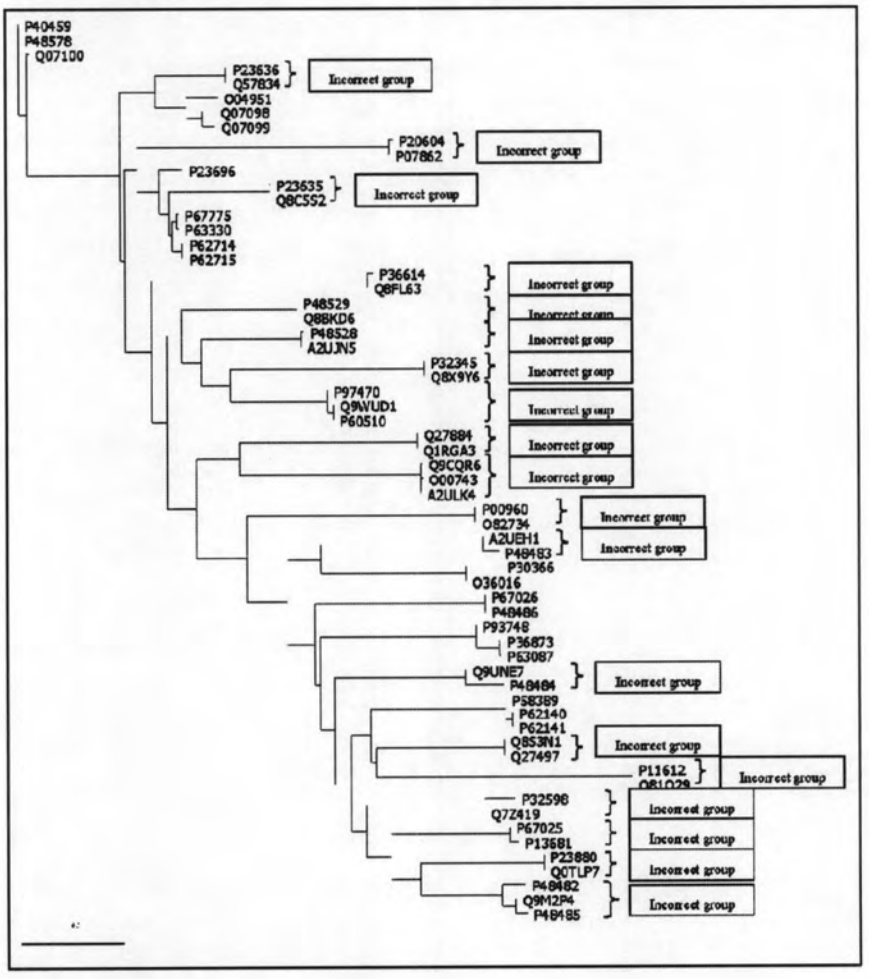


Figure 17 The evolution tree is computed by using our proposed Method and represent by TreeView 1.6.6.0

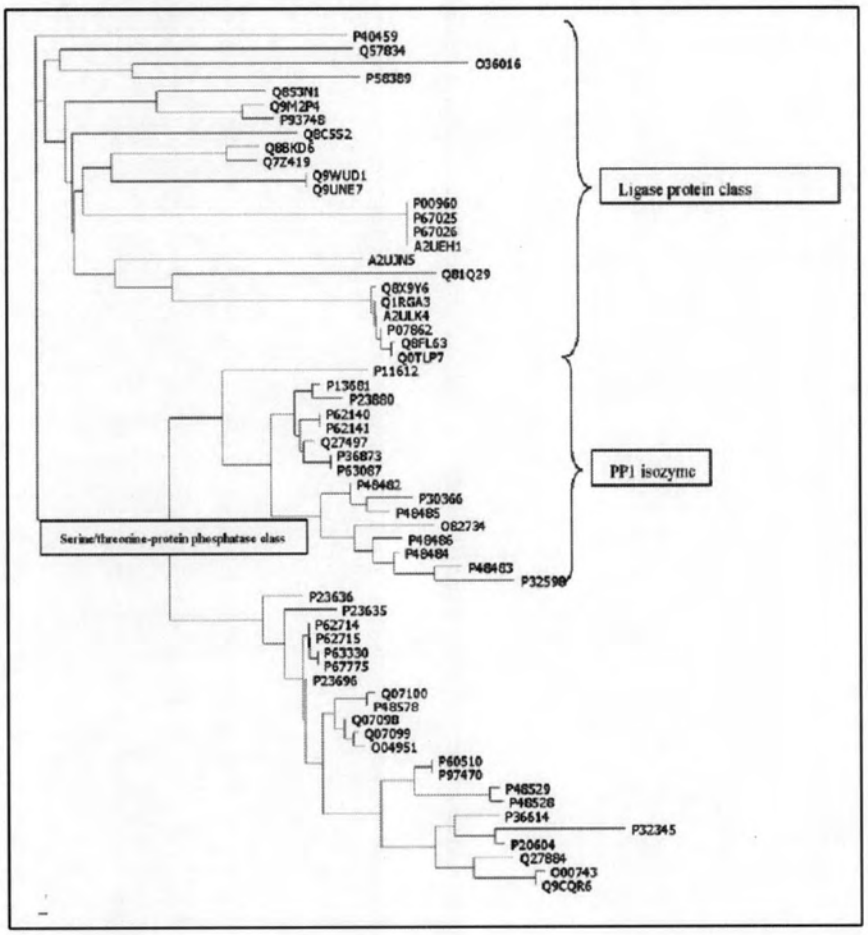


Table 1 Hydrophobicity relatively compared to Glycine

Class	*Hydrophobicity*	*Amino acid*
Hydrophobic	1.5	valine (V)
	2.5	Isoleucine (I)
	1.8	leucine (L)
	2.5	phenylalanine (F)
	3.4	tryptophan (W)
	1.3	methionine (M)
	2.3	tyrosine (Y)
	0.5	alanine (A)
	-2.8	cysteine (C)
Hydrophilic	-7.4	aspartic acid (D)
	-0.3	glutamic acid (E)
	-0.2	asparagine (N)
	-9.9	glutamine (Q)
	-11.2	arginine (R)
	0.5	histidine (H)
	-4.2	lysine (K)
Special	-3.3	proline (P)
	0	glycine (G)
	0.4	threonine (T)
	-0.3	serine (S)

Table 2 Amino acid clusters and their assigned colors

Cluster	*Class*	*Amino Acid*	*Color*
1	hydrophobic	W,I,F,Y,L,V,M	green
2	acidic	D,E,N,Q	red, purple
3	Basic	R,H,K	blue
4		C	yellow
5		S,T	cyan
6		G,P	gray
7		A	dark green

Table 3 List of test data from PIR Database (Barker et al. 1999)

Asscession Number	*Source organism*	*Sequence length*	*Protein name*
Q81Q29	*Bacillus anthracis*	304	D-alanine--D-alanine ligase (EC 6.3.2.4) (D-alanylalanine synthetase) (D-Ala-D-Ala ligase)
Q8X9Y6	*Escherichia coli* O157:H7	306	D-alanine--D-alanine ligase B (EC 6.3.2.4) (D-alanylalanine synthetase B) (D-Ala-D-Ala ligase B)
Q8FL63	*Escherichia coli* O6	306	D-alanine--D-alanine ligase B (EC 6.3.2.4) (D-alanylalanine synthetase B) (D-Ala-D-Ala ligase B)
P07862	*Escherichia coli*	306	D-alanine--D-alanine ligase B (EC 6.3.2.4) (D-alanylalanine synthetase B) (D-Ala-D-Ala ligase B)
Q8BKD6	*Mus musculus* (Mouse)	301	E3 ubiquitin-protein ligase IBRDC2 (EC 6.3.2.-) (IBR domain-containing protein 2)
Q7Z419	*Homo sapiens* (Human)	303	E3 ubiquitin-protein ligase IBRDC2 (EC 6.3.2.-) (IBR domain-containing protein 2) (p53-inducible RING finger protein)
Q9M2P4	*Arabidopsis thaliana* (Mouse-ear cress)	308	E3 ubiquitin-protein ligase SINAT2 (EC 6.3.2.-) (Seven in absentia homolog 2)
Q8S3N1	*Arabidopsis thaliana* (Mouse-ear cress)	309	E3 ubiquitin-protein ligase SINAT5 (EC 6.3.2.-) (Seven in absentia homolog 5)
P67026	*Escherichia coli* O157:H7	303	Glycyl-tRNA synthetase alpha chain (EC 6.1.1.14) (Glycine--tRNA ligase alpha chain) (GlyRS)
P67025	*Escherichia coli* O6	303	Glycyl-tRNA synthetase alpha chain (EC 6.1.1.14) (Glycine--tRNA ligase alpha chain) (GlyRS)
P00960	*Escherichia coli*	303	Glycyl-tRNA synthetase alpha chain (EC 6.1.1.14) (Glycine--tRNA ligase alpha chain) (GlyRS)
P23636	*Schizosaccharomyces pombe* (Fission yeast)	322	Major serine/threonine-protein phosphatase PP2A-2 catalytic subunit (EC 3.1.3.16)
P23635	*Schizosaccharomyces pombe* (Fission yeast)	309	Minor serine/threonine-protein phosphatase PP2A-1 catalytic subunit (EC 3.1.3.16)
P40459	*Saccharomyces cerevisiae* (Baker's yeast)	309	Pantoate--beta-alanine ligase (EC 6.3.2.1) (Pantothenate synthetase) (Pantoate-activating enzyme)
P93748	*Arabidopsis thaliana* (Mouse-ear cress)	305	Putative E3 ubiquitin-protein ligase SINAT1 (EC 6.3.2.-) (Seven in absentia homolog 1)

Asscession Number	*Source organism*	*Sequence length*	*Protein name*
O36016	*Schizosaccharomyces pombe* (Fission yeast)	325	Serine/threonine-protein phosphatase 2A activator 1 (EC 5.2.1.8) (Peptidyl-prolyl cis-trans isomerase PTPA-1) (PPIase PTPA-1) (Rotamase PTPA-1) (Phosphotyrosyl phosphatase activator 1)
P63330	*Mus musculus* (Mouse)	309	Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform (EC 3.1.3.16) (PP2A-alpha)
P67775	*Homo sapiens* (Human)	309	Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform (EC 3.1.3.16) (PP2A-alpha) (Replication protein C) (RP-C)
P62714	*Homo sapiens* (Human)	309	Serine/threonine-protein phosphatase 2A catalytic subunit beta isoform (EC 3.1.3.16) (PP2A-beta)
P62715	*Mus musculus* (Mouse)	309	Serine/threonine-protein phosphatase 2A catalytic subunit beta isoform (EC 3.1.3.16) (PP2A-beta)
P58389	*Mus musculus* (Mouse)	323	Serine/threonine-protein phosphatase 2A regulatory subunit B' (PP2A, subunit B', PR53 isoform) (Phosphotyrosyl phosphatase activator) (PTPA)
P32345	*Saccharomyces cerevisiae* (Baker's yeast)	308	Serine/threonine-protein phosphatase 4 catalytic subunit (EC 3.1.3.16) (PP4C) (Phosphatase PP2A-like catalytic subunit PPH3)
P60510	*Homo sapiens* (Human)	307	Serine/threonine-protein phosphatase 4 catalytic subunit (EC 3.1.3.16) (PP4C) (Pp4) (Protein phosphatase X) (PP-X)
P97470	*Mus musculus* (Mouse)	307	Serine/threonine-protein phosphatase 4 catalytic subunit (EC 3.1.3.16) (PP4C) (Pp4) (Protein phosphatase X) (PP-X)
O00743	*Homo sapiens* (Human)	305	Serine/threonine-protein phosphatase 6 (EC 3.1.3.16) (PP6)
Q9CQR6	*Mus musculus* (Mouse)	305	Serine/threonine-protein phosphatase 6 (EC 3.1.3.16) (PP6)
Q27884	*Drosophila melanogaster* (Fruit fly)	303	Serine/threonine-protein phosphatase PP-V (EC 3.1.3.16)
P48529	*Arabidopsis thaliana* (Mouse-ear cress)	305	Serine/threonine-protein phosphatase PP-X isozyme 1 (EC 3.1.3.16)
P48528	*Arabidopsis thaliana* (Mouse-ear cress)	305	Serine/threonine-protein phosphatase PP-X isozyme 2 (EC 3.1.3.16)
P11612	*Drosophila melanogaster* (Fruit fly)	314	Serine/threonine-protein phosphatase PP-Y (EC 3.1.3.16)

Asscession Number	*Source organism*	*Sequence length*	*Protein name*
P30366	*Arabidopsis thaliana* (Mouse-ear cress)	318	Serine/threonine-protein phosphatase PP1 isozyme 1 (EC 3.1.3.16)
P48482	*Arabidopsis thaliana* (Mouse-ear cress)	312	Serine/threonine-protein phosphatase PP1 isozyme 2 (EC 3.1.3.16)
P48483	*Arabidopsis thaliana* (Mouse-ear cress)	322	Serine/threonine-protein phosphatase PP1 isozyme 3 (EC 3.1.3.16)
P48484	*Arabidopsis thaliana* (Mouse-ear cress)	321	Serine/threonine-protein phosphatase PP1 isozyme 4 (EC 3.1.3.16)
P48485	*Arabidopsis thaliana* (Mouse-ear cress)	312	Serine/threonine-protein phosphatase PP1 isozyme 5 (EC 3.1.3.16)
P48486	*Arabidopsis thaliana* (Mouse-ear cress)	322	Serine/threonine-protein phosphatase PP1 isozyme 6 (EC 3.1.3.16)
O82734	*Arabidopsis thaliana* (Mouse-ear cress)	324	Serine/threonine-protein phosphatase PP1 isozyme 8 (EC 3.1.3.16)
P20604	*Saccharomyces cerevisiae* (Baker's yeast)	311	Serine/threonine-protein phosphatase PP1-1 (EC 3.1.3.16)
P13681	*Schizosaccharomyces pombe* (Fission yeast)	327	Serine/threonine-protein phosphatase PP1-1 (EC 3.1.3.16)
P32598	*Saccharomyces cerevisiae* (Baker's yeast)	312	Serine/threonine-protein phosphatase PP1-2 (EC 3.1.3.16)
P23880	*Schizosaccharomyces pombe* (Fission yeast)	322	Serine/threonine-protein phosphatase PP1-2 (EC 3.1.3.16) (Suppressor protein SDS21)
Q27497	*Caenorhabditis elegans*	329	Serine/threonine-protein phosphatase PP1-alpha (EC 3.1.3.16) (CeGLC-7-alpha) (Glc seven-like phosphatase 1)
P62140	*Homo sapiens* (Human)	327	Serine/threonine-protein phosphatase PP1-beta catalytic subunit (EC 3.1.3.16) (PP-1B)
P62141	*Mus musculus* (Mouse)	327	Serine/threonine-protein phosphatase PP1-beta catalytic subunit (EC 3.1.3.16) (PP-1B)
P36873	*Homo sapiens* (Human)	323	Serine/threonine-protein phosphatase PP1-gamma catalytic subunit (EC 3.1.3.16) (PP-1G) (Protein phosphatase 1C catalytic subunit)
P63087	*Mus musculus* (Mouse)	323	Serine/threonine-protein phosphatase PP1-gamma catalytic subunit (EC 3.1.3.16) (PP-1G) (Protein phosphatase 1C catalytic subunit)
P23696	*Drosophila melanogaster* (Fruit fly)	309	Serine/threonine-protein phosphatase PP2A (EC 3.1.3.16) (Protein microtubule star)
Q07098	*Arabidopsis thaliana* (Mouse-ear cress)	306	Serine/threonine-protein phosphatase PP2A-1 catalytic subunit (EC 3.1.3.16)

Asscession Number	*Source organism*	*Sequence length*	*Protein name*
Q07099	*Arabidopsis thaliana* (Mouse-ear cress)	306	Serine/threonine-protein phosphatase PP2A-2 catalytic subunit (EC 3.1.3.16)
Q07100	*Arabidopsis thaliana* (Mouse-ear cress)	313	Serine/threonine-protein phosphatase PP2A-3 catalytic subunit (EC 3.1.3.16)
P48578	*Arabidopsis thaliana* (Mouse-ear cress)	313	Serine/threonine-protein phosphatase PP2A-4 catalytic subunit (EC 3.1.3.16) (Protein phosphatase 2A isoform 4)
O04951	*Arabidopsis thaliana* (Mouse-ear cress)	307	Serine/threonine-protein phosphatase PP2A-5 catalytic subunit (EC 3.1.3.16)
P36614	*Schizosaccharomyces pombe* (Fission yeast)	305	Serine/threonine-protein phosphatase ppe1 (EC 3.1.3.16) (Phosphatase esp1)
Q9WUD1	*Mus musculus* (Mouse)	304	STIP1 homology and U box-containing protein 1 (EC 6.3.2.-) (STIP1 homology and U-box-containing protein 1) (Carboxy terminus of Hsp70- interacting protein) (E3 ubiquitin-protein ligase CHIP)
Q9UNE7	*Homo sapiens* (Human)	303	STIP1 homology and U box-containing protein 1 (EC 6.3.2.-) (STIP1 homology and U-box-containing protein 1) (Carboxy terminus of Hsp70- interacting protein) (E3 ubiquitin-protein ligase CHIP) (CLL- associated antigen KW-8) (Antigen NY-CO-7)
Q57834	*Methanococcus jannaschii*	306	Tyrosyl-tRNA synthetase (EC 6.1.1.1) (Tyrosine--tRNA ligase) (TyrRS)
Q8C5S2	*Mus musculus* (Mouse)	308	Adult male testis cDNA, RIKEN full-length enriched library, clone: 4933424F19 product: hypothetical HECT domain (Ubiquitin-protein ligase) containing protein, full insert sequence. (Fragment)
A2ULK4	*Escherichia coli* B	306	D-alanine--D-alanine ligase (EC 6.3.2.4)
Q1RGA3	*Escherichia coli* (strain UTI89 / UPEC)	306	D-alanine--D-alanine ligase B (EC 6.3.2.4)
Q0TLP7	*Escherichia coli* O6:K15:H31 (strain 536 / UPEC)	306	D-alanine-D-alanine ligase B (EC 6.3.2.4)
A2UJN5	*Escherichia coli* B	308	Glutamate--tRNA ligase (EC 6.1.1.17)
A2UEH1	*Escherichia coli* B	303	Glycine--tRNA ligase (EC 6.1.1.14)

ประวัติผู้เขียนวิทยานิพนธ์

นายภิสิทธิ์ กรรณสูต เกิดเมื่อวันที่ 22 มิถุนายน พ.ศ. 2527 ที่จังหวัดกรุงเทพมหานครฯ สำเร็จการศึกษาระดับปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) จากคณะวิทยาศาสตร์ สาขาวิชาจุลชีววิทยา จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2548 และได้เข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2549