A study of carboxylic ester hydrolases: Structural classification, properties, and database

by

Yingfei Chen

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Chemical Engineering

Program of Study Committee:
Peter J. Reilly, Major Professor
Mark S. Gordon
Richard B. Honzatko
Monica H. Lamm
Ian C. Schneider

Iowa State University

Ames, Iowa

2015

TABLE OF CONTENTS

ACKNOWLEDGMENT

First of all, I want to thank Dr. Reilly, my major professor, for his guidance, support, and patience over the years. He always gave me insightful suggestions on the research, and also provided warm support in life.

I'd like to thank group members David Cantu and Luis Peterson, who helped me on my research projects, and all the undergraduates who worked with me. Thanks to my committee members, Dr. Gordon, Dr. Honzatko, Dr. Lamm, and Dr. Schneider, for their helpful discussions and suggestions on the research and the dissertation.

I also want to thank my friends in Ames, Xiaofei, Le, Fuyuan, Xunpei, Feng, Yanjie, Yanxiang, Tao, and Yingxi, for their companionship.

Last but not least, I'm so grateful to my family. To my husband Yunjie, as without him, I would not have been able to achieve what I did. He was always there when I needed him, and he helped me to be better and stronger. And to my parents, who gave me all they could, to help me and to encourage me to pursue my dream.

CHAPTER 1. INTRODUCTION

Protein structures are the main topic for all my research projects. Structures are important for proteins because they are closely related to their functions. Protein structures provide us clues about how the proteins work with their substrates, and where the reaction happens.

Initially, I started with protein structure projects related to fatty acid and polyketide synthesis. Three enzyme groups, acyl-CoA carboxylases, ketoacyl synthases, and thioesterases, were classified into families and clans according to similarities in amino acid sequences (primary structures) and three-dimensional structures (tertiary structures). Active sites and mechanisms of enzymes were also compared. Members of each family have very similar primary and tertiary structures and the same reaction mechanisms and active sites, suggesting that they have common protein ancestors. Clan members share similar tertiary structures, although they have different primary structures, which indicates that they are from distant common ancestors. Phylogenetic analysis was performed on these enzymes as well, to reveal the subtle diversity of sequences and producing organisms within families.

Besides the structural classification, I also participated in constructing the Thioester-active enzYme (ThYme) database. These three enzyme groups, together with other five enzyme groups and one molecule in the fatty acid/polyketide synthetic system, can be found in the ThYme database.

The classification of carbohydrate binding modules (CBMs) is a recent research project. Using similar methods, CBMs are grouped into tribes according to the secondary and tertiary structures. All the work above is summarized in Chapter 2.

Chapter 3 introduces the existing ways to classify carboxyl ester hydrolases (CEHs). EC numbers, substrate specificities, and primary and tertiary structural classification of CEHs are summarized in this chapter. Several databases, especially CAZY, ESTHER, LED, and MELDB, cover various parts of CEH classification.

The structural classification of CEHs is covered in Chapter 4 in more detail. This project is related to my previous study of amino acid sequences and crystal structure similarities. As CEHs constitute a widely used enzyme group in industry, their systematic classification in this project will help further understanding of their enzyme mechanisms, active sites, and other valuable properties.

After doing computational work on protein structures, I became interested in the experimental side of protein tertiary structures. It is well known that experimentally obtaining protein tertiary structures can be a challenging process. Thus, it occurred to me that I could use computational methods and available protein structures to observe the trends and infer some conclusions by mining structural data, in order to bridge the computational and experimental data in various ways. Aspects like helping wet laboratory researchers to eliminate any redundant combination of crystallization conditions, or improving computational simulation results of protein structures, was to be my goal. This work, using crystallization condition data from the Protein Data Bank, appears in the Appendix of this dissertation. However, after conducting extensive work in this area, it appeared unlikely to lead to significant results.

Some of the work dealing with fatty acid and polyketide synthases and the ThYme database was conducted with my fellow graduate student David Cantu, the computer specialist Matthew Lemons, and the undergraduates Erin Kelly, Ryan Masluk, Christopher Nelson, and Armando Elizondo-Noriega. The work on CBMs was in collaboration with visiting scholar Caio Carvalho from Brazil, along with the undergraduate student Ngoc Phan. All the projects in this dissertation were advised and mentored by Dr. Reilly, who gave me helpful suggestions, discussion, and valuable advice over the years.

CHAPTER 2. EARLIER RESEARCH

Five projects will be summarized in this chapter. They include four projects conducted during my work toward my M.S. degree and one recent project about carbohydrate binding modules (CBMs). The four previous projects related to the fatty acid synthesis system include the construction of the ThYme database and structural classification of thioesterases (TEs), ketoacyl synthases (KSs), and acyl-CoA carboxylases (ACCs), according to their primary and tertiary structure similarity. The recent project classified CBMs into tribes by their available three-dimensional structures. I was second author of papers on the TEs (Cantu et al., 2010) and the ThYme database (Cantu et al., 2011), first author on the KS (Chen et al., 2011) and ACC (Chen et al., 2012) papers, and third author of the CBM paper (Carvalho et al., 2015).

ThYme database

The ThYme database was created to obtain insights into the fatty acid synthesis-related enzymes (Cantu et al., 2011). Each enzyme group, including acetyl-CoA synthases (ACSs), ACCs, acetyl transferases (ATs), KSs, ketoacyl reductases (KRs), enoyl reductases (ERs), hydroxyacyl dehydratases (HDs), and TEs, was classified into families. Primary structures within each family are similar to each other, as are their tertiary structures. The enzyme active sites and catalytic mechanisms are also conserved in each family, indicating these enzymes come from the same ancestor. Different families can be grouped into single clans, where their primary structures are not related, but their tertiary structures and active sites remain the same. Families in the same clan come from a more distant ancestor.

Protocols and automation scripts to identify families and clans were developed. We gathered query sequences for each enzyme group that are labeled as “evidence at protein level” from the UniProt database (The UniProt Consortium, 2009). Then we used our in-house scripts to run the BLAST program (Altschul et al., 1997) continuously using an E-value

threshold of 0.001, to classify query sequences into potential families (Cantarel et al., 2009). The lower the E-value in BLAST, the more similar sequences are to each other. The family classification was verified by multiple sequence alignment in MUSCLE (Edgar, 2004) and tertiary structure superposition in MultiProt (Shatsky et al., 2004) and PyMOL (DeLano, 2002).

Thioesterases

According to the protocols that we developed, TEs fall into 25 families, with 12 of them being found in four clans (Cantu et al., 2010). Families TE1 to TE13 consist of acyl-CoA hydrolases. TE14 to TE19 members are acyl-ACP hydrolases. TE20 enzymes are protein-palmitoyl hydrolases, TE21 members are protein-acyl hydrolases, and TE22 and TE23 enzymes are glutathione hydrolases. TE24 and TE25 members are nearly all uncharacterized.

HotDog and α/β-hydrolase folds are two most common folds among TE family structures. TE4 to TE15 members, as well as TE24 and TE25 enzymes, have HotDog folds, whereas TE2 and TE16 to TE22 enzyme structures have α/β-hydrolase folds. Other protein folds also exist, including NagB folds for TE1 proteins, flavodoxin-like folds for TE3 structures, and lactamase folds for TE23 members. Most TE families have two or more PDB structures, except for families TE5, TE7, TE12, TE15, and TE19, where their root mean square deviations (RMSDs) of the distances between α-carbon atoms of different tertiary structures, and P_{avg} values, indicating the average percentages of α-carbon atoms that could be compared (Cantu et al., 2010), were calculated. All these families have average RMSD values lower than 1.8 Å and P_{avg} values greater than 75%, with two exceptions.

Clans group two or more families together when they share similar tertiary structures, active sites, and reaction mechanisms. Tertiary structures were superimposed, and RMSDs and P_{avg} values were recorded.

Clan TE-A includes families TE5, TE9, TE10, and TE12, where they all share a similar HotDog fold. Clan TE-B consists of families TE8, TE11, and TE13, with HotDog folds as

well. However, TE-A and TE-B share only limited sequence similarities and limited secondary structure elements, which make them two separate clans. α/β-Hydrolase clans TE-C and TE-D consist of families TE16 to TE18 and families TE20 and TE21, respectively, according to secondary and tertiary structure analysis.

The catalytic residues and mechanisms for TE families are summarized as well, if available. Various active sites and mechanisms exist in HotDog fold structures. In clan TE-A, Tyr7, Asp11, and His18 were proposed as the catalytic residues in PDB structure 2PZH of TE9. In clan TE-B, TE8 member 3F5O has Asn50, Asp65, Ser83, and Gly57 important for catalysis. TE11 residues Gly65 and Glu73 in 1Q4S were suggested as the catalytic residues. The catalytic residues in TE13 member 1WLU appear to be Gly40 and Asp48. Unlike the HotDog fold enzymes, which have various catalytic residues and mechanisms, α/β-hydrolase fold enzymes have conserved Ser-His-Asp catalytic triads.

Ketoacyl synthases

Ketoacyl synthases were classified into five families, KS1 to KS5 (Chen et al., 2011). KS1 members are mainly 3-ketoacyl-ACP synthase III (KAS III) enzymes, which condense malonyl-ACP with acyl-CoA to produce acetoacyl-ACP. KS2 members are 3-ketoacyl-CoA synthases, fatty acid elongases, and very long-chain fatty acid (VLCFA) condensing enzymes, which come from eukaryota, especially plants. KS3 enzymes are generally 3-ketoacyl-ACP synthases I and II (KAS I and KAS II), and KS domains of large multifunctional type I fatty acid synthases (FASs). KS3 is the largest KS family, containing over 13,000 sequences at the time of publication. KS4 members are mainly chalcone synthases, narigenin-chalcone synthases, stilbene synthases, and polyketide synthases. Most KS4 sequences come from eukaryota, and the rest come from bacteria. KS5 comprises elongation of VLCFA proteins and fatty acid elongases. They all come from eukaryota and most of them are from animals. KS2 and KS5 enzymes are transmembrane proteins, producing VLCFAs.

KS1, KS3, and KS4 families are part of the same clan. They share the same five-layer α-β-α-β-α protein structures. Furthermore, catalytic residues from these three families are found in the same positions. KS1 and KS4 sequences have the same catalytic triad, Cys-His-Asn, where KS3 enzymes have Cys-His-His as the catalytic residues. These three families share the same ping-pong kinetic mechanism (Plowman et al., 1972), using Cys-His-Asn/His residues. KS2 and KS5 members have no available crystal structures. Since they are transmembrane proteins, it is hard to obtain their crystal structures experimentally. However, computational simulation studies have been done on KS2 (Joubès et al., 2008) and KS5 enzymes (Chumningan et al., 2010), respectively, where homology modeling and *ab initio* methods were applied to predict their tertiary structures.

Phylogenetic analysis has been conducted for all five KS families to see the sequence diversity within each family (Chen et al., 2011). Phylogenetic trees were produced by MEGA (Tamura et al., 2007, 2011), with the minimum evolution method (Nei and Kumar, 2000), 250 to 500 bootstrap iterations, and Jones-Taylor-Thronton (JTT) model values (Jones et al., 1992). Potential subfamilies were selected based on the visual divergence, and they were verified by the statistical Z-value test (Mertz et al., 2005). KS1 to KS5 have 12, 10, 14, 10, and 11 subfamilies, respectively. KS1 phylogenetic analysis enabled an experimentalist to rationally select 30 representative genes of interest to characterize (S. Garg, personal communication, 2013).

Acyl-CoA carboxylases

Acyl-CoA carboxylases (ACCs) consist of three functional domains: biotin carboxylases (BCs), biotin carboxyl carrier proteins (BCCPs), and carboxyl transferases (CTs). BCCP is a structural domain that swings between BC and CT domains. The BC domain adds bicarbonate to the BCCP biotin moiety. Then BCCP swings to the CT domain, leaving its carboxyl group to acetyl-CoA, in order to form malonyl-CoA (Knowles, 1989).

One gene or several individual genes may encode BC, BCCP, and CT domains, which leads them to be in the same protein chain or in several different chains. Each domain was extracted from the whole sequences and classified into families according to their primary and tertiary structure similarities (Chen et al., 2012). There is one BCCP family (BCCP1), one BC family (BC1), and two CT families (CT1 and CT2). The CT1 family corresponds to CT_β sequences and the CT2 family contains mainly CT_α sequences.

ACCs include acetyl-CoA carboxylases, propionyl-CoA carboxylases (PCCs), methylcrotonoyl-CoA carboxylases (MCCs), geranoyl-CoA carboxylases (GCCs), acetone carboxylases, and 2-oxoglutarate carboxylases (Chen et al., 2012). The domain arrangements for them were studied. Acetyl-CoA carboxylases from bacteria and plants excluding grasses have individual chains for BC, BCCP, CT_α, and CT_β domains (Chen et al., 2012). Acetyl-CoA carboxylases from eukaryota less plants other than grasses have all four domains in one multi-functional protein chain (Nikolau et al., 2003; Zhang et al., 2003; Sasaki & Nagano, 2004). PCCs, MCC, and GCCs sequences from bacteria and eukaryota have two chains: BC and BCCP domains are in one chain, and the CT_β domain is in another (Toh et al., 1993; Rodríguez & Gramajo, 1999; Nikolau et al., 2003; Lombard & Moreira, 2011). Acetyl-CoA carboxylases and PCCs from archaea have three separate chains, containing BC, BCCP, and CT_β domains, respectively (Chen et al., 2012).

Phylogenetic analysis was performed on each ACC family, including BC, BCCP, CT_α, and CT_β domains (Chen et al., 2012). We used MUSCLE (Edgar, 2004) to conduct multiple sequence alignment (MSA) for representative sequences and MEGA (Kumar et al., 2008) to produce the phylogenetic trees. Potential subfamilies were selected by visual inspection and verified by the statistical Z-value test (Mertz et al., 2005).

The swinging arm mechanism for BCCP to swing between BC and CT domains was first proposed by Waldrop et al. (1994). The facts that crystal structures of BC and CT domains in PCCs and MCCs are separated by 55 Å and 80 Å (Huang et al., 2010, 2011), respectively, supports this swinging arm theory. Bacterial acetyl-CoA carboxylases with CT_α and CT_β

domains on separate chains have their active sites on the interface of the αβ dimers (Bilder et al., 2006). PCCs and MCCs from bacteria and eukaryota have two chains: BC and BCCP are on one chain (α subunit), CT_β is on the other (β subunit) (Chen et al., 2012). They form an $\alpha_6\beta_6$ architecture, where their active sites are located on the dimer interface of two β subunits.

Carbohydrate binding modules

Carbohydrate binding modules (CBMs) are protein domains that associate their corresponding catalytic domains with substrates, in order to increase the catalytic efficiencies of the active sites of the latter. CBMs fall into 67 families according to their amino acid sequence similarity on the Carbohydrate-Active enzyme (CAZy) database (Lombard et al., 2013). Sequences within each family are statistically similar, while sequences from different families are not. Fifty-one of the families have at least one known tertiary structure, which enabled them to be further classified into tribes (Carvalho et al., 2015). Although primary structures from different families are not similar, similar tertiary structures can sometimes be observed, indicating that they are from the same distant common protein ancestor. Since CBMs have no catalytic function, we use the term "tribe", to indicate their tertiary structure similarity and their share of same common ancestors and binding mechanisms, instead of "clan" as previously seen in TEs, KSs, and ACCs.

In general, the same methods were applied to the classification of CBM tribes as the ones used with TEs, KSs, and ACCs. One representative tertiary structure from each CBM family was selected and overlapped with representative structures from other families. RMSDs and P_{avg} values were calculated to show their structural similarity. Secondary structure elements (SSEs) of CBMs were checked as well. Additional criteria for CBM structural classification are the configuration of ligand glycosidic bonds that CBMs bind and the CBM chain length. The criteria for tribe classifications in the order of importance are: SSE order and location, RMSDs and P_{avg} values, binding ligands, chain length, and their producing organisms.

CBM families were classified into nine tribes, CBM-A to CBM-I, containing 27 out of 51 families, when the criteria described above were applied. Each tribe contains two or more families. Members of eight tribes, CBM-A to CBM-H, have β-sandwich protein folds, which are characterized by two antiparallel β-sheets (Sillitoe et al., 2012), although each tribe has its own characteristic β-strand arrangement. Members of the last tribe, CBM-I, have β-trefoil protein folds with threefold axes (Murzin et al., 1992), where β-strands are folded into three similar parts with α-helices at the corners (Carvalho et al., 2015). Members of CBM-A to CBM-C bind α-linked ligands only, such as starch, cyclodextrins, and glycogen, whereas members of CBM-D to CBM-I bind to β-linked polysaccharides, including various β-(1,4)- and β-(1,3)-linked glucans, galactan, mannan, and xylans, and sometimes α-linked ligands as well. Moreover, CBMs in one tribe often associate with more than one glycoside hydrolase (GH) family (Carvalho et al., 2015).

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, *25*(17), 3389–3402.

Bilder, P., Lightle, S., Bainbridge, G., Ohren, J., Finzel, B., Sun, F., et al. (2006). The structure of the carboxyltransferase component of acetyl-CoA carboxylase reveals a zinc-binding motif unique to the bacterial enzyme. *Biochemistry*, *45*(6), 1712–1722.

Cantarel, B., Coutinho, P., & Rancurel, C. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res*, 37(suppl 1), D233–D238.

Cantu, D. C., Chen, Y., & Reilly, P. J. (2010). Thioesterases: A new perspective based on their primary and tertiary structures. *Protein Sci*, *19*(7), 1281–1295.

Cantu, D. C., Chen, Y., Lemons, M. L., & Reilly, P. J. (2011). ThYme: a database for thioester-active enzymes. *Nucleic Acids Res*, *39* (Database issue), D342–346.

Carvalho, C. C., Phan, N. N., Chen, Y., & Reilly, P. J. (2015). Carbohydrate binding module tribes. *Biopolymers*, *103*, 203–214.

Chen, Y., Kelly, E. E., Masluk, R. P., Nelson, C. L., Cantu, D. C., & Reilly, P. J. (2011). Structural classification and properties of ketoacyl synthases. *Protein Sci*, *20*(10), 1659–1667.

Chen, Y., Elizondo-Noriega, A., Cantu, D. C., & Reilly, P. J. (2012). Structural classification of biotin carboxyl carrier proteins. *Biotechnol Lett*, *34*(10), 1869–1875.

Chumningan, S., Pornputtapong, N., Laoteng, K., Cheevadhanarak, S., & Thammarongtham, C. (2010). 3D Structure modeling of a transmembrane protein, fatty acid elongase. In Computational Systems–Biology and Bioinformatics, pp. 36–45, Springer, Berlin and Heidelberg.

DeLano, W. L. (2002). The PyMOL User's Manual. DeLano Scientific, San Carlos, CA, 452.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, *32*(5), 1792–1797.

Huang, C. S., Ge, P., Zhou, Z. H., & Tong, L. (2011). An unanticipated architecture of the 750-kDa α6β6 holoenzyme of 3-methylcrotonyl-CoA carboxylase. *Nature*, *481*(7380), 219–223.

Huang, C. S., Sadre-Bazzaz, K., Shen, Y., Deng, B., Zhou, Z. H., & Tong, L. (2010). Crystal structure of the α6β6 holoenzyme of propionyl-coenzyme A carboxylase. *Nature*, *466*(7309), 1001–1005.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci,* 8(3), 275–282.

Joubès, J., Raffaele, S., Bourdenx, B., Garcia, C., Laroche-Traineau, J., Moreau, P., et al. (2008). The VLCFA elongase gene family in *Arabidopsis thaliana*: phylogenetic analysis, 3D modelling and expression profiling. *Plant Mol Biol*, *67*(5), 547–566.

Knowles, J. R. (1989) The mechanism of biotin-dependent enzymes. *Ann Rev Biochem,* 58(1), 195–221.

Kumar, S., Nei, M., Dudley, J., & Tamura, K. (2008). MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform*, *9*(4), 299–306.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2013). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res*, *42*(D1), D490–D495.

Lombard, J., & Moreira, D. (2011). Early evolution of the biotin-dependent carboxylase family. *BMC Evol Biol*, *11*(1), 232.

Mertz, B., Kuczenski, R. S., Larsen, R. T., Hill, A. D., & Reilly, P. J. (2005). Phylogenetic analysis of family 6 glycoside hydrolases. *Biopolymers*, *79*(4), 197–206.

Murzin, A. G., Lesk, A. M., & Chothia, C. (1992). β-Trefoil fold. Patterns of structure and sequence in the Kunitz inhibitors interleukins-1β and 1α and fibroblast growth factors. *J Mol Biol*, *223*(2), 531–543.

Nei, M., & Kumar, S. (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Nikolau, B. J., Ohlrogge, J. B., & Wurtele, E. S. (2003). Plant biotin-containing carboxylases. *Arch Biochem Biophys*, *414*(2), 211–222.

Plowman, K. M. (1972) Enzyme Kinetics. McGraw-Hill, New York. pp. 41–42.

Rodríguez, E., & Gramajo, H. (1999). Genetic and biochemical characterization of the alpha and beta components of a propionyl-CoA carboxylase complex of *Streptomyces coelicolor* A3(2). *Microbiology (Reading, Engl.)*, 145(11), 3109–3119.

Sasaki, Y., & Nagano, Y. (2004). Plant acetyl-CoA carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. *Biosci Biotechnol Biochem*, *68*(6), 1175–1184.

Shatsky, M., Nussinov, R., & Wolfson, H. J. (2004). A method for simultaneous alignment of multiple protein structures. *Proteins: Struct Funct Bioinf*, *56*(1), 143–156.

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., et al. (2012). New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res*, *41*(D1), D490–D498.

Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol*, *24*(8), 1596–1599.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*, *28*(10), 2731–2739.

Toh, H., Kondo, H., & Tanabe, T. (1993). Molecular evolution of biotin-dependent carboxylases. *FEBS Lett*, *215*(3), 687–696.

UniProt Consortium (2009). The universal protein resource (UniProt) in 2010. *Nucleic Acids Res*, *38* (Database), D142–D148.

Waldrop, G. L., Rayment, I., & Holden, H. M. (1994). Three-dimensional structure of the biotin carboxylase subunit of acetyl-CoA carboxylase. *Biochemistry*, *33*(34), 10249–10256.

Zhang, H., Yang, Z., Shen, Y., & Tong, L. (2003). Crystal structure of the carboxyltransferase domain of acetyl-coenzyme A carboxylase. *Science*, *299*(5615), 2064–2067.

CHAPTER 3. ENZYME DATABASES

Carboxylic ester hydrolases

The carboxylic ester hydrolases (CEHs) catalyze the hydrolysis of a carboxylic ester bond to generate a carboxylate and an alcohol (Fig. 1). Because of their importance in pharmaceutical, food, and detergent industries, they are among the most studied enzymes in industry. The active sites of CEHs are Ser/His/Asp catalytic triads, together with oxyanion holes near the catalytic sites.

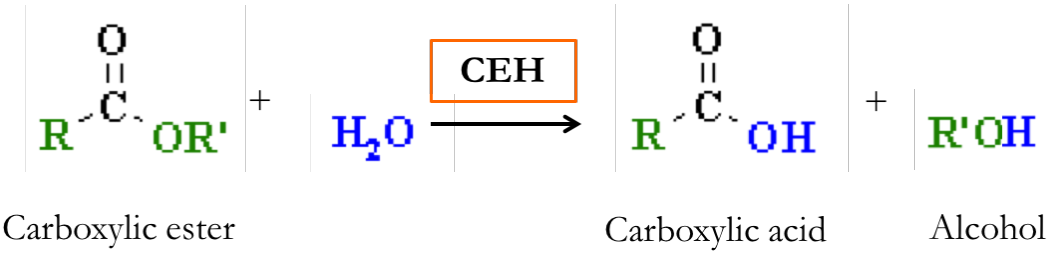


Figure 1. Reactions catalyzed by CEH enzymes.

The widely studied CEHs include carboxylesterases (EC 3.1.1.1) for their important role in the metabolism of a large number of diverse drugs (Satoh and Hosokawa, 1998); triacylglycerol lipases (EC 3.1.1.3) for their triacylglycerol synthesis and secretion for energy storage and release of fatty acids (Lehner and Kuksis, 1996); phospholipase A2's (EC 3.1.1.4) for their production of free fatty acids such as arachidonic acid and the lysoglycerophospholipids, which are the precursors of eicosanoids that play a role in sleep regulation, inflammation, and immune responses (Schaloske and Dennis, 2006); lysophospholipases (EC 3.1.1.5) for their kinetic resolution of racemic mixtures of industrial chemicals (Lo et al., 2003); acetylcholinesterases (EC 3.1.1.7), as they are the targets of nerve agents, insecticides, and therapeutic drugs, in particular the anti-Alzheimer drugs (Silman and Sussman, 2005); butyrylcholinesterases (EC 3.1.1.8) for their involvement in drug hydrolysis to explain drug responses in individuals with diseases (Li et al., 2005); phospholipase A1's (EC 3.1.1.32) for their inhibition of cationic amphiphilic drugs like chlorpromazine, chloroquine, and propran-

olol (Kubo and Hostetler, 1985); cutinases (EC 3.1.1.74) for their hydrolysis of cutin to produce C_{16} and C_{18} fatty acids (Nicolas et al., 1996; Egmond, 2000); and cocaine esterases (EC 3.1.1.84) for their natural role in cocaine metabolism (Larsen et al., 2001; Gao et al. 2009).

Enzyme classification

Enzymes can be classified into groups by different criteria. One widely used method is to classify them based on the chemical reactions that they catalyze, according to the Enzyme Commission (NC-IUBMB, 1992). This way of enzyme classification is presented by a four-digit nomenclature. The four-digit EC number describes enzyme functions in more and more detail. The first digit ranges from one to six, representing oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases, respectively. Enzymes of EC 3 are all hydrolases, and EC 3.1 enzymes are a subgroup of hydrolases that act specifically on ester bonds. EC 3.1.1 represents enzymes hydrolyzing a carboxylic ester bond, to form an alcohol and a carboxylate. An enzyme will be assigned an EC number once its function is known. Taking CEHs as an example: They are assigned EC numbers 3.1.1.X, where X can be any digit between 1 and 97.

Substrate specificities are another way to classify enzymes. CEHs are commonly classified into two groups: esterases and lipases. Esterases catalyze the water-soluble short-chain fatty acids, while lipases prefer the long acyl chain, water-insoluble fatty acids as substrates.

Enzymes can be classified by their primary structures (amino acid sequences) and tertiary structures (three-dimensional structures) as well. The Pfam database (Sonnhammer et al., 1997; Bateman et al., 2004) classifies proteins into domains and families according to their primary structures by multiple sequence alignment. It uses Hidden Markov Model profiles to find the domains in new sequences. Pfam covers about 80% UniProt knowledgebase proteins (Punta et al., 2012). The CATH (Sillitoe et al., 2012) and SCOP databases (Murzin et al., 1995) classify enzymes mainly by their tertiary structures. CATH assigns four hierarchies to

each protein: class, architecture, topology, and homologous superfamilies. The class indicate the secondary structure composition of the proteins: mainly α-helices, mainly β-strands, α and β structures, or a few irregular secondary structures. Then, proteins are classified into more detailed groups within their class, represented by a four-digit CATH number.

CAZy (Lombard et al., 2013) and ThYme (Cantu et al., 2011) are databases that specialize in particular protein groups. The CAZy database is built for carbohydrate-active enzymes, and the ThYme database is for thioester-active enzymes. They classify enzymes into families and clans by their similarities in both primary structures and tertiary structures. They are comprehensive databases that integrate enzyme information about sequences and their producing organisms, active sites, mechanisms, and tertiary structures. They contain links to external databases such as GenBank and the Protein Data Bank. Within each family, amino acid sequences, tertiary structures, and reaction mechanisms are conserved. Within each clan, tertiary structures and mechanisms are well conserved, although their primary structures are completely different from each families in the same clan.

Enzyme databases for CEHs

ESTHER is a database of α/β-hydrolase-fold proteins and their classification (Hotelier et al., 2004), tabulating the sequences, three-dimensional structures, and biochemical and pharmacological information about these proteins. Typical α/β-hydrolase folds are a β-sheet connected by several α-helices. Usually, there are five to eight β-strands on the β-sheet, with the second β-strand antiparallel to the other β-strands (Ollis et al., 1992; Hotelier et al., 2004; Lenfant et al., 2013). The number of families has expanded from 69 in 2004 to 148 in 2013, with over 30,000 manually curated proteins (Lenfant et al., 2013). The ESTHER database takes in all α/β-hydrolase-fold proteins, including proteins other than CEHs such as the peptidases and thioesterases, as long as they have the α/β-hydrolase fold. Some non-catalytic proteins with the same fold are also included. CEHs are a substantial group of enzymes in the

ESTHER database; however, they exist in other protein folds as well, such as six-propeller folds and three-solenoid folds.

The CAZy database (Lombard et al., 2013) contains a group of carbohydrate esterases (CEs) that belong to the CEHs. Since CAZy includes carbohydrases exclusively, only enzymes acting on carbohydrates are found in it.

The Lipase Engineering Database (LED) integrated sequence and tertiary structure information of esterases, lipases, and other related proteins with α/β-hydrolase folds (Fischer and Pleiss, 2003; Fischer et al., 2006; Widmann et al., 2010). The enzymes in LED covered eight EC numbers under 3.1.1.X, with 38 LED superfamilies being assigned to about 25,000 sequences and over 1000 tertiary structures. The cutoff was a BLAST E-value of 10^{-10} in LED, which is lower than those being used in the CAZy and ESTHER databases. This led to a smaller family size and a higher sequence similarity in family members than found in ESTHER families. Annotated multiple sequence alignments for catalytic residues, binding sites and mutation, and phylogenetic trees of each family were available on the database. LED has not been updated since 2010.

MELDB is a database that covers microbial carboxylesterases and triacylglycerol lipases, which are enzymes in EC 3.1.1.1 and EC 3.1.1.3. These two enzyme groups are classified by the primary and tertiary structure similarities. It aimed to find new biocatalysts with unique biochemical properties, and to study directed evolution (Kang et al., 2006). The researchers applied a local alignment algorithm and TribeMCL (a graph clustering algorithm), instead of the common global pairwise alignment. These methods reduced the noise introduced by the global alignment, and they were able to distinguish the outlier sequences successfully, whereas the traditional methods were not. The MELDB classification mainly corresponds to part of the LED database. The MELDB database has not been updated since 2006.

References

Bateman A, Coin L, Durbin R, Finn RD, et al. (2004). The Pfam protein families database. Nucleic Acids Res 32(D), D138–D141.

Cantu DC, Chen Y, Lemons ML, and Reilly PJ (2011). ThYme: a database for thioester-active enzymes. Nucleic Acids Res, 39(Database issue), D342–346.

Egmond MR (2000). *Fusarium solani pisi* cutinase, Biochimie, 82, 1015−1021.

Fischer M, Pleiss J (2003). The Lipase Engineering Database: a navigation and analysis tool for protein families. Nucleic Acids Res, 31(1), 319–321.

Fischer M, Thai QK, Grieb M, Pleiss J (2006). DWARF – a data warehouse system for analyzing protein families. BMC Bioinfo, 7, 495.

Gao D, Narasimhan DL, Macdonald J, Brim R, Ko MC, Landry DW, et al (2009). Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. Mol Pharm, *75*(2), 318–323.

Hotelier T, Renault L, Cousin X, Negre V, Marchot P, Chatonnet A (2004). ESTHER, the database of the α/β-hydrolase fold superfamily of proteins. Nucleic Acids Res, 32(D), D145–D147.

Kang HY, Kim JF, Kim MH, Park SH, Oh TK, and Hur CG (2006). MELDB: A database for microbial esterases and lipases. FEBS Lett, *580*(11), 2736–2740.

Kubo M, Hostetler KY (1985). Mechanism of cationic amphiphilic drug inhibition of purified lysosomal phospholipase A1. Biochemistry, 24(23), 6515–6520.

Larsen NA, Turner JM, Stevens J, Rosser SJ, Basran A, Lerner RA, et al. (2001). Crystal structure of a bacterial cocaine esterase. Nature Struct Biol, 9(1), 17–21.

Lehner R, Kuksis A (1996). Biosynthesis of triacylglycerols. Prog Lipid Res, 35(2), 169–201.

Lenfant N, Hotelier T, Velluet E, Bourne Y, Marchot P, Chatonnet A (2013). ESTHER, the database of the α/β-hydrolase fold superfamily of proteins: tools to explore diversity of functions. Nucleic Acids Res, 41(D1), D423–D429.

Li B, Sedlacek M, Manoharan I, Boopathy R, Duysen EG, Masson P, and Lockridge O (2005). Butyrylcholinesterase, paraoxonase, and albumin esterase, but not carboxylesterase, are present in human plasma. Biochem Pharm, 70(11), 1673–1684.

Lo YC, Lin SC, Shaw JF, and Liaw YC (2003). Crystal structure of *Escherichia coli* thioesterase I/protease I/lysophospholipase L 1: Consensus sequence blocks constitute the catalytic center of SGNH-hydrolases. J Mol Biol, 330, 539–551.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2013). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res, 42(D1), D490–D495.

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol, 247(4), 536–540.

Nicolas A, Egmond M, Verrips CT, de Vlieg J, Longhi S, Cambillau C, and Martinez C (1996). Contribution of cutinase serine 42 side chain to the stabilization of the oxyanion transition state. Biochem, 35(2), 398-410.

Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC–IUBMB) (1992). Enzyme Nomenclature, Academic Press, San Diego. Available at: http://www.chem.qmul.ac.uk/iubmb/enzyme/.

Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, et al. (1992). The α/β hydrolase fold. Protein Eng, 5(3), 197–211.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, et al. (2012). The Pfam protein families database. Nucleic Acids Res, 40(D1), D290–D301.

Satoh T, Hosokawa M (1998). The mammalian carboxylesterases: From molecules to functions. Annu Rev Pharmacool Toxicol, 38, 257–288.

Schaloske RH, Dennis EA (2006). The phospholipase A2 superfamily and its group numbering system. Biochim Biophys Acta–Mol Cell Biol Lett, 1761(11), 1246–1259.

Sillitoe I, et al. 2012. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acid Res, 41(D1), D490–D498.

Silman I and Sussman JL (2005). Acetylcholinesterase: ‘classical’ and ‘non-classical’ functions and pharmacology. Curr opin pharmacol, *5*(3), 293–302.

Sonnhammer EL, Eddy SR, Durbin R (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins: Struct Funct Bioinfor, 28(3), 405–420.

Widmann M, Juhl PB, Pleiss J (2010). Structural classification by the Lipase Engineering Database: a case study of *Candida antarctica* lipase A. BMC Genomics, 11(1), 123.

CHAPTER 4. STRUCTURAL CLASSIFICATION OF CARBOXYLIC ESTER HYDROLASES

Abstract

The carboxylic ester hydrolases (CEHs) are enzymes that hydrolyze an ester bond to form a carboxylic acid and an alcohol. They are one of the enzyme groups that are most explored industrially for their applications in the food, flavor, pharmaceutical, organic synthesis, and detergent industries.

We classified CEHs into families and clans according to their amino acid sequences (primary structures) and three-dimensional structures (tertiary structures). Our work has established the systematic structural classification of the CEHs. Primary structures of family members are similar to each other, and their active sites and reaction mechanisms are conserved. The tertiary structures of members of each clan, which is composed of different families, remain very similar, although amino acid sequences of members of different families are not similar.

CEHs were divided into 127 families by use of BLAST, with 67 families being grouped into seven clans. Multiple sequence alignment and tertiary structures superposition were used, and active sites and reaction mechanisms were analyzed. Python and Shell scripts were implemented to automate the process of comparing CEH primary and tertiary structures.

A comprehensive database, CASTLE (CArboxylic eSTer hydroLasEs), may be constructed to provide the primary and tertiary structures of CEHs. This database would be available at www.castle.enzyme.iastate.edu and will be accessible to the entire biology community.

Introduction

There are two common kinds of carboxylic ester hydrolases (CEHs): esterases and lipases. Esterases hydrolyze water-soluble acyl chains of fatty acids, and lipases hydrolyze

water-insoluble long-chain fatty acids, in each case cleaving an ester bond to form a carboxylic acid and an alcohol. CEHs are one of the enzyme groups that most explored industrially, because of their wide use in the food, flavor, pharmaceutical, organic synthesis, and detergent industries (Hasan et al., 2006).

CEHs are ubiquitous in all life forms: viruses, archaea, bacteria, and eukaryota. Esterases and lipases are two important classes of CEHs. The esterases prefer shorter acyl esters, fewer than ten carbon atoms, than the lipases (Levisson et al., 2009). According to the CATH numbers assigned to CEH tertiary structures (Sillitoe et al., 2013), many of them have α/β-hydrolase folds, which are composed of three α/β/α layers, with the second β-strand being antiparallel to the others in the β-sheet (Ollis et al., 1992; Andreeva et al., 2007). Others may be composed of only α-helices or only β-strands. Some CEH structures have six-propeller folds, which consist of a six-bladed β-sheet with a central axis. Some have four-layer sandwich folds, where several anti-parallel β-strands are arranged in two β-sheets. Three-solenoid folds are also found in CEH structures; they consist of many parallel β-strands arranged into three β-sheets. The outer-membrane CEHs are commonly found in β-barrel folds.

To this point there is no systematic structural classification of CEHs. The CAZy database (Lombard et al., 2013) has classified some of these enzymes, but only the carbohydrate esterases, which catalyze the de-O- or de-N-acylation of substituted saccharides. Other CEHs such as triacylglycerol lipases and acetylcholine esterases are not included in this classification. The ESTHER database (Hotelier et al., 2004; Lenfant et al., 2012) covers part of the CEHs, focusing on the classification of α/β-hydrolase fold structures. It is not limited to CEHs, but includes other enzymes such as peptidases and thioesterases that have this fold. The LED database (Fischer and Pleiss, 2003) classified lipases and esterases by their function, sequences, and crystal structures. The database covered nine EC numbers under EC 3.1.1.X, where X represents digits between 1 and 97. Its founders employed much smaller E-values in their use of BLAST, the Basic Local Alignment Search Tool (Altschul et al., 1997)

to gather primary structures than do the curators of CAZy, implying that the family members in LED were more similar to each other. However, it has not been maintained since 2009.

The research reported in this chapter systematically classifies the CEHs by their primary and tertiary structure similarities. This will cast light on the various ways that CEHs with different primary or tertiary structures catalyze the same reaction. A comprehensive database, CASTLE (CArboxylic eSTer hydroLasEs), will be constructed, to make CEH structural information and their classification fully accessible to the research community over the world. The database will be available at www.castle.enzyme.iastate.edu.

Methods

Potential CEH family identification

To classify the CEH proteins, the query sequences of CEH needed to be gathered first. The EC number (NC-IUBMB, 1992) indicates the enzyme function, with CEHs being found under EC 3.1.1.X, where X represents any number at the fourth position to describe the CEH function in greater detail. As the time of writing, CEHs were classified by 91 EC numbers, from EC 3.1.1.1 to EC 3.1.1.98 and EC 3.1.1.– (unclassified), with seven of them being deleted. All the sequences with evidence at protein level in the UniProt database (UniProt Consortium, 2008) of these EC numbers were collected as query sequences. The criterion of evidence at protein level is to ensure that wet laboratory experiments have been done on these proteins to verify their protein functions as CEHs. This criterion ruled out a large portion of protein sequences in EC 3.1.1.X obtained from whole-genome projects, whose functions are putative because their sequences have only been compared with those with known CEHs, but whose functions have not been verified experimentally.

Query sequences were checked on the Pfam database (Finn et al., 2010) to obtain their catalytic domains only. BLAST was used consecutively to find similar primary structures of these catalytic domains. The up-to-date NR database, which gathers non-redundant protein sequences from various databases such as PDB, PIR, Swiss-Prot, and NCBI RefSeq, was

used to search similar sequences against the query sequences. The threshold E-value in BLAST was set to 0.001. Protein sequences with an E-value lower than 0.001 were regarded as similar enough to the query sequence to be included in one potential family (Cantu et al., 2010). In-house Python and Shell scripts were implemented to automate the process of obtaining catalytic domains of query sequences in Pfam, and performing BLAST consecutively to find potential families. All the scripts were run on the Google cloud platform with Linux Cent OS7 installed.

After each run of BLAST using query sequences, one result file for each resulting outseq file was generated, and sequences within the result file comprise a potential family. The potential families needed to be verified by multiple methods, which will be discussed in the next section.

CEH family verification

Multiple sequence alignment (MSA) and tertiary structure superposition are two main techniques to verify the potential families, with the possibility of merging or splitting them.

A sample of random sequences in each potential family was used to perform the MSA. The alignment is to ensure that these sequences are similar enough, with several positions of amino acid residues conserved along the entire sample. If no amino acid residue is conserved and if clear differences are observed in the MSA result, then the potential family was split into multiple families.

The tertiary structures from each potential family, if available, were superimposed. The monomer of each tertiary structure was extracted and compared. The root mean square deviation (RMSD) of the α-carbon atoms was calculated, together with the P_{avg}, indicating the percentage of atoms that can be compared (Cantu et al., 2010; Chen et al., 2011).

Furthermore, the active sites of the enzyme should remain in the similar position within each family. If active sites were already known in the literature, their positions were checked.

Also, secondary structure elements were compared and analyzed to ensure that each family has almost the same elements.

CEH clan identification

Clans are composed of two or more different families, where their active sites, reaction mechanisms, and tertiary structures are conserved from family to family, although their primary structures are not similar from one family to the next. We used folds defined by CATH (Ollis et al., 2015) to first divide the tertiary structures into separate groups. Tertiary structure representatives from different families were superimposed by MultiProt (Shatsky et al., 2004). RMSD and P_{avg} values were calculated to determine whether they are similar enough to be in a clan. Varying from previous methods to calculate RMSD and P_{avg}, pairwise RMSD and P_{avg} values were calculated for representative structures from each family. This variation is caused by the large number of families with the same fold, which is difficult to visually distinguish in PyMOL (DeLano, 2002) to find potential clans.

Each combination of two representative structures from different families were superimposed by MultiProt, and their pairwise RMSD and P_{avg} values were recorded in matrices, to cluster different families with the same folds assigned by CATH into potential clans. The superposition in MultiProt, along with RMSD and P_{avg} calculations, were implemented by Python scripts. To cluster similar structures into potential clans, the pairwise RMSD matrix were imported into MEGA 6.06 (Tamura et al., 2013), and neighbor joining trees were produced in the form of curved and circular trees. Although MEGA was intended to produce phylogenetic trees for the study of molecular evolution, the pairwise distance matrix used in MEGA is similar enough to be used for the RMSD matrix. Thus, we used MEGA to analyze the RMSD matrix produced by the pairwise structure superposition, in order to cluster the CEH tertiary structures. Potential clans were proposed according to the trees. Then the structures of potential clan members were superimposed and inspected in PyMOL. The proposed classification was tuned until the structures superimposed in PyMOL were in good align-

ment. Interestingly, the pairwise alignments did not perform as well as visual inspection. With visual inspection, similar PDB structures from different families were grouped roughly into potential clans, then their PDBs were superimposed by MultiProt.

CEH clan verification

RMSD and P_{avg} values were calculated for structures within each potential clan, after the structure superposition by MultiProt, and the structures were visually inspected in PyMOL. This is to ensure that the tertiary structures are more similar to others in one clan than to those from other clans. Active sites were checked, if available, to see whether the catalytic residues are in similar positions to act on the substrates, and share the same mechanism in each clan.

Results

The potential 130 families became 127 families after MSA using ClustalW and structure superposition by MultiProt and PyMOL. Among them, 91 families have known PDB structures, and 68 of them were grouped into seven clans. In addition, 36 families have no available tertiary structures (Table 1).

Each clan has its characteristic protein folds. Clan A proteins all have α/β hydrolase folds, in which the second β-strand is antiparallel to the others in the β-sheet (Ollis et al., 1992) (Figure 1). Clan B members have similar folds as those in clan A; however, all their β-strands are parallel to each other and are in the same direction. Clan C enzymes have α/β hydrolase-like tertiary structures as well, but their first β-strands are antiparallel to the others on the same β-sheet, whereas Clan A's second β-strands are antiparallel. The tertiary structures of clan D members are six-propeller folds, where six β-sheet blades share a central axis. Clan E enzymes have three-helix folds, and clan F proteins share three-solenoid folds, which consist of many parallel β-strands gathered into three β-sheets, comprising the solenoid-shape protein fold. Clan G members have 4-layer α/β/β/α sandwich folds. Furthermore, folds

with β-barrels exist in outer-membrane CEHs, and 3-layer sandwich folds exist in CEHs as well. Families with these structures cannot be grouped into clans, because their members have various shapes and cannot be superimposed well, although they share the same fold.

Each family was verified by three methods: MSA, secondary structure analysis, and tertiary structure superpositions. The sequence alignment files from Clustal X can be found at the dissertation supplemental materials at the following URL: (https://www.dropbox.com/sh/x4jhjbv5fgs9jzl/AADRadnpC-EPhQHALfr2mWvOa?dl=0). The conserved amino acid counts for each family are summarized in Table 2, as are RMSD and P_{avg} values obtained by tertiary structure superposition. Clan members of various families can be found in Table 3, where RMSD and P_{avg} values of each clan member are listed, and the protein folds of each clan are summarized. Representative tertiary structures from seven clans are shown in Figure 1. The secondary structures of crystal structures in each family are obtained by secondary structure analysis (Supplementary Information Figure S1). Each clan member has similar secondary structures in their cores, with some extra α-helices or β-strands.

Family and clan numbering

The families were reordered so that all those in the same clan were numbered consecutively. Then families with no PDB structures, which cannot be grouped to clans, were listed below these families.

Family content

Phospholipase A2's (EC 3.1.1.4 and EC 3.1.1.5)

The phospholipase A2 (PLA2) enzymes catalyze the hydrolysis of fatty acids from the *sn*-2 position of glycerophospholipids, to generate free fatty acids (Schaloske and Dennis, 2006) (Figure 2). They are drug targets for inflammatory disease and coronary heart disease (Burke and Dennis, 2008a; Corbett et al., 2010). The PLA2's have five main types: the secreted sPLA2's, cytosolic cPLA2's, calcium-independent iPLA2's, PAF acetyl hydrolase lpPLA2's,

and lysosomal PLA2's. The first four types are EC 3.1.1.4 phospholipases, and the last type are EC 3.1.1.5 lysophospholipases. sPLA2's have histidine (His) and aspartic acid (Asp) residues as a catalytic dyad, with an oxyanion hole near to the active site. The cPLA2's have the same catalytic residues as other CEH families and require Ca^{2+} in the reaction. However, they are regarded as an individual family, because they have a substrate preference of arachidonic acid in the *sn*-2 position of phospholipids (Clark et al., 1991; Ghosh et al., 2006; Burke and Dennis, 2008a). The iPLA2's are enzymes that can catalyze reactions without the presence of Ca^{2+}. They use a serine residue to cleave the *sn*-2 ester bond. Among them, the most characterized enzymes are the iPLA2's, which are regulated through many mechanisms such as ATP binding, calmodulin, caspase cleavage, and possible protein aggregation caused by intervening ankyrin repeats (Burke and Dennis, 2008a,b). The iPLA2's are also important in axon regeneration and wallerian degeneration in nerve injury (Burke and Dennis, 2008a; López-Vales et al., 2008). The PFA acetyl hydrolase/oxidized lipid lpPLA2's can cleave oxidized lipids in the *sn*-2 position, from acetyl up to acyl groups with nine carbon atoms. They use the Ser/His/Asp catalytic triad, instead of the active dyads in all the other PLA2's. These enzymes have anti-inflammatory activity *in vivo*, according to studies of PLA2 from human plasma (Tjoelker et al., 1995; Burke and Dennis, 2008a). PAF lpPLA2's show a positive risk factor in coronary heart disease and are a promising drug target. Lysosomal PLA2's use Ser/His/Asp as the catalytic triad, and they need four cysteine residues in total for catalytic activity (Hiraoka et al., 2002, 2005; Burke and Dennis., 2008b). The PLA2 enzymes exist in families 33 and 35 in clan A, families 61–63 in clan E, and in families 68, 72, 74, 77, 82, 85, 116, 123, and 124 that are not part of clans. Lysophospholipases are in families 13, 20, 33, and 35 in clan A, families 39, 48, and 51 in clan B, family 52 in clan C, and families 90, 104, 123, 124, 126, and 127 that are not part of clans.

Cholinesterases (EC 3.1.1.7 and EC 3.1.1.8)

The cholinesterases have two groups of enzymes: acetylcholinesterases (AChE's, EC 3.1.1.7) and butyrylcholinesterases (BChE's, EC 3.1.1.8) (Figure 3). AChEs hydrolyze acetylcholine to produce choline and acetate. Acetylcholine is a neurotransmitter that carries signals from nerve cells to muscle cells, and the reaction to generate acetylcholine happens very fast. The AChE inhibitors are drug targets for psychotropic diseases such as Alzheimer's (Cummings, 2000; Houghton et al., 2006). There are substantial structural studies on this enzyme group because of their medical importance. Tertiary structures of cholinesterases have a deep (20 Å) "catalytic gorge", and the active sites are located at the bottom of this gorge. The AChE from the electric eel *Torpedo californica* (*tc*AChE), has the active site Ser200/His440/Glu327 (Sussman et al., 1991; Ordentich et al., 1998; Zhang et al., 2002). Human BChE (*hu*BChE) has a structure similar to that of *tc*AChE, and their catalytic triad is Ser198/His438/Glu325 (Vellom et al., 1993; Suárez and Field, 2005). Besides their active sites, AChE's have a peripheral site at the entrance of the narrow gorge, and they are the binding sites for propidium (Barak et al., 1997; Johnson and Moore, 1999) and antibodies (Saxena et al., 1997). AChE sequences are in family 34 in clan A and families 41 and 43 in clan B. Hysteresis of BChEs has been observed in human, rats, and horse types. It was proposed that oscillations occurred when substrates exist in different conformation, interconvertible, and aggregation forms. Although there is no evidence that hysteresis plays a role in BChE functioning, a toxicological or physiological importance for the BChE hysteresis cannot be ruled out (Masson et al., 2005). Kinetic studies of BChE have been conducted. The K_m for substrates decreased as the length of alkyl chain increased, and the longest chain-length substrates have high affinity of BChE enzymes. Molecular modelling revealed that the docking energy decreased as the alkyl chain length increased. The best substrates for rat BChEs were short alkyl homologues (Hrabovska et al., 2006). BchE enzymes are in family 34 of clan A.

Carboxylesterases (EC 3.1.1.1)

Carboxylesterases (CarbE's) are enzymes that hydrolyze carboxylic ester bonds to produce carboxylates and alcohols (Figure 4). CarbE's catalyze the hydrolysis of a substantial number of drugs as substrates, such as cocaine, salicylate, palmitoyl-coenzyme A, and steroids (Satoh and Hosokawa, 1998). The metabolism of heroin and cocaine is the same in human liver CarbE's (Kamendulis et al., 1996). Satoh and Hosokawa (1998) classified mammalian CarbE's into four groups, CES1 to CES4, where subgroups of CES1 are from CES1A to CES1C. Human CarbE's are CES1A1 and CES1A2, as are other mammalian CarbE's, including those from rats, mice, rabbits, and dogs. CES1B's preferentially hydrolyze long-chain acyl-CoA's. In a 2006 study by the same group, the phylogenetic trees expanded from CES1 to CES5, and the CES1 subgroups expanded from CES1A to CES1H. CarbE catalytic amino acid residues are Ser/Glu/His. A comparison of substrate specificities is included in the work as well: CES1's preferentially hydrolyze cocaine, meperidin, and temocapril; CES2's prefer heroin, CPT11, and methylprednisolone 21-hemisuccinate (Satoh and Hosokawa, 2006). Families 6, 9, 11, 15, 17, 27, and 36 in clan A, families 39, 40, 45, 47, and 51 in clan B, family 52 in clan C, and family 110, which is not in any clan, have CarbE enzymes.

Cutinases (EC 3.1.1.74)

Cutinases are named for their hydrolysis of ester bonds in cutin, to produce mainly C_{16} and C_{18} fatty acids, and they can also catalyze the hydrolysis of short- and long-chain triacylglycerols (TAGs) (Egmond and De Vlieg, 2000) (Figure 5). Cutinases bridge the esterases and lipases in terms of their substrate specificities (Martinez et al., 1993). Esterases hydrolyze water-soluble short acyl chains, whereas lipases hydrolyze water-insoluble long chains (Chahiniana and Sarda, 2009). Cutinases have the catalytic mechanism as serine esterases, where they use Ser120, Asp175, and His188 as their catalytic residues, with two nitrogen atoms of Ser42 and Gln121 as the oxyanion hole (Martinez et al., 1993, 1994; Nicolas et al.,

1996). The catalytic residues of bacterial *Thermobifida fusca* cutinase are Ser170, Asp216, and His248, while Tyr100 and Met171 act as the oxyanion hole (Chen et al., 2008). The catalytic serine residue is located at the fifth β-strand on a sharp turn, or the so-called nucleophile elbow (Egmond and De Vlieg, 2000). Homology modeling was performed on a bacterial cutinase from *Thermobifida fusca*, using fungal *Streptomyces exfoliates* lipase as a template, because they were the best match in MSA and they share 63% sequence identity. Potential synergistic effects between two *T. fusca* cutinases were studied as well, and there was no synergism between them, suggesting that further studies about the reason why two genes for *T. fusca* cutinases are needed (Chen et al., 2008). Cutinases are in families 41 and 43 in clan B.

Phospholipase A1's (EC 3.1.1.32)

Phospholipase A1's (PLA1's) catalyze hydrolysis of the *sn-1* ester bond of phospholipids to produce 2-acyl-lysophospholipids. Ser/Asp/His is the catalytic triad of PLA1 (Aoki et al., 2002) (Figure 6). Phosphatidylserine-specific PLA1 (PS-PLA1) is involved in three kinds of reaction to convert PS and 1-acyl-2-lysoPS to 2-acyl-1-lysoPS. PS affects blood coagulation, marker of apoptosis, phagocytosis, and activation of intracellular enzymes. 1-Acyl-2-lysoPS is involved in the activation of mast cells and potentiation of NGF-induced neural cell differentiation. 2-Acyl-1-lysoPS contributes in the growth suppression of T cells, along with the activation of mast cells. PLA2's hydrolyze PS into 1-acyl-2-lysoPS, and PS is hydrolyzed by PS-PLA1 to produce 2-acyl-1-lysoPS (Aoki et al., 2002). Scandella and Kornberg (1971) characterized PLA1 substrate specificities for phase-induced, latent, and purified PLA1 for 1-acyl attack. The substrates are phosphatidylethanolamine (PE), lysoPE (T4-infected cells), lysoPE (osmotic lysis), lysoPE *in vitro*, 1-acyllysoPE, with hydrolysis of 15, 15, 20, 25, and 98%. Their saturated:unsaturated fatty acid ratios are 0.81, 0.29, 0.33, 0.35, and 2.9, respectively. Cationic amphiphilic drugs like chloroquine, chlorpromazine, and propranonol inhibit PLA1 in vitro (Pappu and Hostetler, 1984; Hostetler et al., 1985). Chloroquine competitively

inhibits PLA1 by forming EI_2 complexes. Chlorpromazine and propranonol bind to small unilamellar liposome substrates in a positive and collaborative way with two binding sites: a low-affinity site with high capacity, and a high-affinity low-capacity site (Kubo and Hostetler, 1985). PLA1's exist in family 49 in clan B, and in families 71, 72, 74, and 85 among those enzymes not in clans.

Cocaine esterases (EC 3.1.1.84)

Cocaine esterases (CocE's) are the natural enzymes for treating cocaine overdose and addiction. CocE is the first enzyme in the metabolism of cocaine degradation. The first CocE crystal structure was reported by Larsen and colleagues (2002) (Figure 7). It has three domains: (a) a canonical α/β hydrolase fold; (b) an α-helical domain that is a lid above the active site; and (c) a jelly-roll-like β-domain that interacts with the previous two domains. Their study suggested the substrate recognition is between the highly evolved specificity enzyme pocket and the benzoyl moiety of cocaine. The catalytic triad is Ser117/His287/ Asp259 with Try118 and Try44 of the PDB structure 1JU3 in the oxyanion hole (Larsen et al., 2002). Gao and his co-workers (2009) rationally designed mutants of CocE to improve its thermostability. The computational simulation followed by *in vitro* and *in vivo* experimentation obtained about a 30-fold increase in plama half-life of CocE. The simulation first indentified the weakest domain at a high temperature. Then, it virtually screened the possible mutants through interaction energy calculation, and used the most promising thermostable mutants to test in wet laboratory experimentation. This successful case provides a valuable strategy towards their dramatic implications on CocE therapeutic potentials. CocE's are in family 116.

Triacylglycerol lipases (EC 3.1.1.3)

Triacylglycerol lipases hydrolyze triglycerols to diacylglycerols and carboxylates. Triacylglycerols (triglycerides, TGs) are the main energy storage molecules and fatty acids in most living organisms (Yen et al., 2008) (Figure 8). Two primary sources of fatty acids for

triacylglycerol synthesis are diet and *de novo* synthesis. The tissues that most actively synthesize triglycerols are liver and intestine, where adipose tissue is known for the storage of triacylglycerols and release of fatty acids as albumin-bound complexes in plasma. In liver, brain and other tissues, there are two main places for acyl chain elongation, one in the mitochondria, and the other in the endoplasmic reticulum (Lehner and Kuksis, 1996). The catalytic triad was reported in the X-ray structure of *Mucor miehei* triglyceride lipase as Ser144/His257/Asp203 in PDB 1TGL (Brady et al., 1990). Family 26 in clan A, family 40 clan B, and family 100 in no clan have triacylglycerol lipases.

Comparison with existing databases

Several enzyme databases have classified a partial list of carboxyl ester hydrolases. The ESTHER database (Hotelier et al., 2004; Lenfant et al., 2013) focuses on α/β hydrolase fold-like enzymes. ESTHER has three types of blocks, and 94 rank 1 families and 174 rank 2 families, among which 42 rank 1 families have sequences from our CEH families. These 42 ESTHER families contain 28,349 sequences. Each of the three blocks on ESTHER indicates their sequences come from the common ancestor. Block C overlaps with sequences from our family 34. Block L has sequences from families 49, 51, and 105. Block X has CEH sequences from families 5–7, 10, 11, 14, 16, 22, 23, 26, 45, and 54.

Carbohydrate esterase (CE) families on CAZy contain 16 families, and seven of them overlap with the sequences from our CEH families (Cantarel et al., 2009; Lombard et al., 2013). These seven families include 17,060 sequences, and they have sequences from 17 CEH families: 19, 25, 27, 32, 37, 41, 43, 44, 46, 64, 65, 73, 81, 86, 92, 98, and 119. The Lipase Engineering Database (LED) has three classes, 38 superfamilies, and 112 families of lipases (Fischer and Pleiss, 2003; Fischer et al., 2006; Widmann, 2010). The LED database contains 24,783 sequences with 1117 PDB structures. CEH families are more inclusive than these three databases, because they contain about 480,000 sequences and 2101 PDB structures.

Discussion

Classifying CEH enzymes into families and clans provides valuable insights about them. Several observations may be made about their structural classifications:

1) Some CEHs with the same enzyme function appear in multiple families and clans. For instance, PLA2's are found in 14 families and two clans, suggesting that they have two kinds of tertiary structures. Cholinesterase sequences are from 14 families and three clans.

2) Each clan includes diverse enzymes with various functions. The biggest clan (clan A) includes enzymes with six different EC numbers and more than six enzyme functions. Clan B contains enzymes with five EC numbers.

3) Some families show little experimental work on their enzymes. Fourteen families composed of 11,122 sequences have only one sequence of evidence at protein level for each family, and in addition they have no known tertiary structure. These families have between 13 (CEH 68) to 3091 (CEH 90) sequences. The fourteen families need more attention from researchers, because these unexplored enzymes may have novel substrate specificities that will be useful or important to industrial or medical applications.

4) Another scenario is that the same enzymes occur in multiple families and clans, but they have few studies on them. These enzymes have more than one type of tertiary structures catalyzing the same reaction. They can be good study targets for researchers interested in the structure-function relationships. For instance, arylesterases (EC 3.1.1.2), acylglycerol lipases (EC 3.1.1.23), 3-oxoadipate *enol*-lactonases (EC 3.1.1.24), aminoacyl-tRNA hydrolases (EC 3.1.1.29), and acetylxylan esterases (EC 3.1.1.72) have sequences from five different families, respectively. However, each enzyme has no or only a few tertiary structures, and many families containing them have no tertiary structure of them.

CASTLE database applications

The CASTLE database, if constructed, would provide useful classified CEH-related enzyme information. It would be an essential tool for the scientists working on CEH enzymes.

It reveals the families and clans, where few studies have been done. These unexplored families may hide some valuable CEH enzymes that are suitable for industrial or medical use. CASTLE would also summarize the widely explored families and clans, by listing their existing sequences and tertiary structures, reaction mechanisms, and substrate specificities.

As with the previous success of CAZY and THYME databases, the CASTLE database would aim to help researchers access comprehensive information and resource about CEHs. More conclusions can be deduced from the CASTLE classification. For instance, it is known that within each family, the active sites, tertiary structures, and reaction mechanism remain the same. With existing information about any of the three, family members have the same properties as the experimentally studied enzymes. This will help to clarify a large portion of the sequences from the genome projects, with no experiments on them. CASTLE would also provide insights about the sequence, structure and function relationship, which will help scientists rational design CEHs for desired substrate specificities. Last but not least, the classification of CEH families and clans can be used as a uniformed nomenclature of existing CEHs, which have various names and aliases.

Future work

Although substantial work has been done in this project, further studies can be done in several aspects. For the CASTLE database, more interactive ways to obtain the targeted enzyme data can be developed, compared to the PDB database as a successful case. The sequences and tertiary structures to download can be written into various formats, including csv and Excel spreadsheets. Diverse ways to access the data can be created, such as SQL or drop-down lists. The Python and Shell scripts created for this project will be uploaded to Github, an open source community, for maintenance and improvement purpose. This may include further improvement of existing scripts for better performance and usability, and construction of a new database similar as CASTLE. It also provides an online space for putting together all the scripts and tracking updates for the scripts.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman, DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25, 3389–3402. Available at: http://blast.ncbi.nlm.nih.gov/Blast.cgi/.

Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. (2007). Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res, 36, D419–D425.

Aoki J, Taira A, Takanezawa Y, Kishi Y, Hama K, Kishimoto T, et al. (2002). Serum lysophosphatidic acid is produced through diverse phospholipase pathways. J Biol Chem, 277(50), 48737–48744.

Barak R, Ordentlich A, Barak D, Fischer M, Benschop HP, De Jong LP, and Shafferman A, et al. (1997). Direct determination of the chemical composition of acetylcholinesterase phosphonylation products utilizing electrospray-ionization mass spectrometry. FEBS Lett, 407(3), 347–352.

Brady L, Brzozowski AM, Derewenda ZS, Dodson E (1990). A serine protease triad forms the catalytic centre of a triacylglycerol lipase. Nature, 343, 767–770.

Burke JE, Dennis EA (2008a). Phospholipase A2 structure/function, mechanism, and signaling. J Lipid Res, 50 (Supplement), S237–S242.

Burke, JE, Dennis EA (2008b). Phospholipase A2 biochemistry. Cardiovasc Drug Ther, 23 (1), 49–59.

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res, 37(D), D233–D238.

Cantu DC, Chen Y, Reilly P J (2010) Thioesterases: A new perspective based on their primary and tertiary structures. Protein Sci, 19, 1281–1295.

Chahiniana H, Sarda L (2009). Distinction between esterases and lipases: comparative biochemical properties of sequence-related carboxylesterases. Protein Peptide Lett, 16(10), 1149–1161.

Chen S, Tong X, Woodard RW, Du G, Wu J, Chen J (2008). Identification and characterization of bacterial cutinase. J Biol Chem, 283(38), 25854–25862.

Chen Y, Kelly EE, Masluk RP, Nelson CL, Cantu DC, Reilly PJ. (2011). Structural classification and properties of ketoacyl synthases. Protein Sci, 20(10), 1659–1667.

Clark JD, Lin LL, Kriz RW, Ramesha CS, Sultzman LA, Lin AY, Milona N, Knopf LJ (1991). A novel arachidonic acid-selective cytosolic PLA2 contains a Ca^{2+}-dependent translocation domain with homology to PKC and GAP. Cell, 65, 1043–1051.

Corbett JW, Freeman-Cook KD, Elliott R, Vajdos F, Rajamohan F, Kohls D., et al. (2010). Discovery of small molecule isozyme non-specific inhibitors of mammalian acetyl-CoA carboxylase 1 and 2. Bioorg Med Chem Lett, 20(7), 2383–2388.

Cummings JL (2000). Cholinesterase inhibitors: A new class of psychotropic compounds. Am J Psychiatry, 157(1), 4–15.

DeLano WL (2002) The PyMOL molecular graphics system, DeLano Scientific, Palo Alto, CA, USA, http://www.pymol.org/.

Egmond MR, De Vlieg J (2000). *Fusarium solani pisi* cutinase, Biochimie, 82, 1015–1021.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunesekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Res 38, D211–D222. Available at: http://pfam.sanger.ac.uk/.

Fischer M, Pleiss J (2003). The Lipase Engineering Database: a navigation and analysis tool for protein families. Nucleic Acids Res, 31(1), 319–321.

Fischer M, Thai QK, Grieb M, Pleiss J (2006). DWARF – a data warehouse system for analyzing protein families. *BMC Bioinform*, *7*, 495.

Gao D, Narasimhan DL, Macdonald J, Brim R, Ko MC, Landry DW, et al. (2009). Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. Mol Pharm, *75*(2), 318–323.

Ghosh MD, Tucker E, Burchett SA, Leslie CC (2006). Properties of the group IV phospholipase A2 family. Prog Lipid Res, 45, 487–510.

Hasan F, Shah AA, Hameed A (2006). Industrial applications of microbial lipases. Enzyme Microb Technol, 39(2), 235–251.

Hiraoka M, Abe A, Shayman JA (2002). Cloning and characterization of a lysosomal phospholipase A2 1-O-acylceramide synthase. J Biol Chem, 277(12), 10090–10099.

Hiraoka M, Abe A, Shayman JA (2005). Structure and function of lysosomal phospholipase A2, identification of the catalytic triad and the role of cysteine residues. J Lipid Res, 46(11), 2441–2447.

Hostetler KY, Reasor M, Yazaki PJ (1985). Chloroquine-induced phospholipid fatty liver. Measurement of drug and lipid concentrations in rat liver lysosomes. J Biol Chem, 260(1), 215–219.

Hotelier T, Renault L, Cousin X, Negre V, Marchot P, Chatonnet A (2004). ESTHER, the database of the α/β-hydrolase fold superfamily of proteins. Nucleic Acids Res, 32, D145–D147.

Houghton PJ, Ren Y, Howes MJ (2006). Acetylcholinesterase inhibitors from plants and fungi. Nat Prod Rep, 23(2), 181–199.

Hrabovska A, Debouzy J-C, Froment M-T, Devinsky F, Paulikova I, Masson P (2006). Rat butyrylcholinesterase-catalysed hydrolysis of N-alkyl homologues of benzoylcholine. FEBS J, 273(6), 1185–1197.

Johnson G, Moore SW (1999). The adhesion function on acetylcholinesterase is located at the peripheral anionic site. Biochem Biophys Res Commun, 258(3), 758–762.

Kamendulis LM, Brzezinski MR, Pindel EV, Bosron WF, Dean RA (1996). Metabolism of cocaine and heroin Is catalyzed by the same human liver carboxylesterases. J Pharmacol Exp Ther, 279(2), 713–717.

Kubo M, Hostetler KY (1985). Mechanism of cationic amphiphilic drug inhibition of purified lysosomal phospholipase A1. Biochemistry, 24(23), 6515–6520.

Larsen NA, Turner JM, Stevens J, Rosser SJ, Basran A, Lerner RA, et al. (2002). Crystal structure of a bacterial cocaine esterase. Nature Struct Biol, 9(1), 17–21.

Lehner R, Kuksis A (1996). Biosynthesis of triacylglycerols. Prog Lipid Res, 35(2), 169–201.

Lenfant N, Hotelier T, Velluet E, Bourne Y, Marchot P, Chatonnet A (2012). ESTHER, the database of the α/β-hydrolase fold superfamily of proteins: tools to explore diversity of functions. Nucleic Acids Res, 41(D1), D423–D429.

Levisson M, van der Oost J, and Kengen SWM (2009). Carboxylic ester hydrolases from hyperthermophiles. Extremophiles, 13(4), 567–581.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2013). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res, 42(D1), D490–D495. Available at: http://www.cazy.org/.

López-Vales R, Navarro X, Shimizu T, Baskakis C, Kokotos G, Constantinou-Kokotou, David S et al. (2008). Intracellular phospholipase A2 group IVA and group VIA play important roles in Wallerian degeneration and axon regeneration after peripheral nerve injury. Brain, 131(10), 2620–2631.

Martinez C, de Geus P, Stanssens P, Lauwereys M, Cambillau C (1993). Engineering cysteine mutants to obtain crystallographic phases with a cutinase from *Fusarium solani pisi*. Protein Eng, 6(2), 157–165.

Martinez C, Nicolas A, van Tilbeurgh H, Egloff MP, Cudrey C, Verger R, Cambillau C (1994). Cutinase, a lipolytic enzyme with a preformed oxyanion hole. Biochemistry, 33(1), 83–89.

Masson P, Schopfer LM, Froment M-T, Debouzy J-C, Nachon F, Gillon E et al. (2005). Hysteresis of butyrylcholinesterase in the approach to steady-state kinetics. Chem Biol Interact, 157–158, 143–152.

Nicolas A, Egmond M, Verrips CT, de Vlieg J, Longhi S, Cambillau C, and Martinez C (1996). Contribution of cutinase serine 42 side chain to the stabilization of the oxyanion transition state. Biochemistry, 35(2), 398–410.

Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC–IUBMB) (1992) Enzyme Nomenclature, Academic Press, San Diego, CA. Available at: http://www.chem.qmul.ac.uk/iubmb/enzyme/.

Ollis DL, Cheah E, Cyglerl M, Dijkstra B, Frolow F, Franken SM, et al. (1992). The α/β hydrolase fold. Protein Eng, 5(3), 197–211.

Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken S M, Harel M, Remington SJ, Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees JG, Lehtinen S, Studer RA, Thornton J, Orengo CA. (2015). CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res, 43(D1), D376–D381. Available at: http://www.cathdb.info/.

Ordentlich A, Barak D, Kronman C, Ariel N, Segall Y, Velan B, Shafferman A (1998). Functional characteristics of the oxyanion hole in human acetylcholinesterase. J Biol Chem, 273(31), 19509–19517.

Pappu A, Hostetler KY (1984). Effect of cationic amphiphilic drugs on the hydrolysis of acidic and neutral phospholipids by liver lysosomal phospholipase A. Biochem Pharmacol, 33(10), 1639–1644.

Satoh T, Hosokawa M (1998). The mammalian carboxylesterases: From molecules to functions. Annu Rev Pharmacool Toxicol, 38, 257–288.

Satoh T, Hosokawa M (2006). Structure, function and regulation of carboxylesterases. Chem Biol Interact, 162(3), 195–211.

Saxena A, Maxwell DM, Quinn DM, Radić Z, Taylor P, Doctor BP (1997). Mutant acetylcholinesterases as potential detoxification agents for organophosphate poisoning. Biochem Pharm, 54(2), 269–274.

Scandella CJ, Kornberg A (1971). Membrane-bound phospholipase A1 purified from *Escherichia coli*, Biochemistry, 24, 4447–4456.

Schaloske RH, Dennis EA (2006). The phospholipase A2 superfamily and its group numbering system. Biochim Biophys Acta-Mol Cell Biol Lett, 1761(11), 1246–1259.

Shatsky M, Nussinov R, Wolfson HJ. (2004). A method for simultaneous alignment of multiple protein structures. Proteins Struct Function Bioinf, 56(1), 143–156.

Sillitoe I, Cuff AL, Dessailly BH, Dawson NL, Furnham N, Lee D, Lees JG, Lewis TE, Studer RA, Rentzsch R, Yeats C, Thornton M, Orengo CA (2013). New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic Acid Res, 41(D1), D490–D498.

Suárez D, Field MJ (2005). Molecular dynamics simulations of human butyryl-cholinesterase. Proteins Struct Funct Bioinf, 59(1), 104–117.

Sussman JL, Harel M, Frolow F, Oefner C, Goldman A, Toker L, Silman I (1991). Atomic structure of acetylcholinesterase from *Torpedo californica*: a prototypic acetylcholine-binding protein. Science, 253(5022), 872–879.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol, 30(12), 2725–2729.

Tjoelker LW, Wilder C, Eberhardt C, Stafforini DM, Dietsch G, Schimpf B, Hooper S, Le Trong H, Cousens LS, Zimmerman GA, et al. (1995). Anti-inflammatory properties of a platelet-activating factor acetylhydrolase. Nature, 374, 549–553.

UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36, D190–D195. Available at: http://www.uniprot.org/.

Vellom DC, Radic Z, Li Y, Pickering NA, Camp S, Taylor P (1993). Amino acid residues controlling acetylcholinesterase and butyrylcholinesterase specificity. Biochemistry, 32(1), 12–17.

Widmann M, Juhl PB, Pleiss J (2010). Structural classification by the Lipase Engineering Database: a case study of *Candida antarctica* lipase A. BMC Genomics, 11(1), 123.

Yen CLE, Stone SJ, Koliwad S, Harris C, Farese RV (2008). Thematic review series: Glycerolipids DGAT enzymes and triacylglycerol biosynthesis. J Lipid Res, 49(11), 2283–2301.

Zhang Y, Kua, J, McCammon, JA (2002). Role of the catalytic triad and oxyanion hole in acetylcholinesterase catalysis: an ab initio QM/MM study. J Am Chem Soc, 124(35), 10572–10577.

Table 1. Clans and families of carboxyl ester hydrolases.

Family	Number of sequences	Number of sequences with evidence at protein level	Number of known tertiary structures	Dominant, secondary, tertiary enzyme names	EC numbers
Clan A					
1	1715	14	8	Peptidyl-tRNA hydrolase 2	
2	7996	31	15	6-Phosphogluconolactonase	
3	15097	79	74	Peroxiredoxin, alkyl hydroperoxide reductase	
4	6115	4	19	Carboxymethylenebutenolidase, dienelactone hydrolase	
5	1216	5	1	α/β-Hydrolase, esterase	
6	8649	7	14	Carboxylesterase, α/β-hydrolase	3.1.1.1
7	23483	27	115	α/β-Hydrolase	
8	410	5	2	D-Aminoacyl-tRNA deacylase	
9	6704	17	10	S-Formylglutathione hydrolase	3.1.1.1
10	6241	8	14	Carboxylesterase, α/β-hydrolase	
11	4644	3	2	Esterase, α/β-hydrolase	3.1.1.1
12	4994	53	12	Phospholipase A1	
13	2631	9	5	Monoglyceride lipase, lysophospholipase	3.1.1.5
14	1899	5	3	α/β-Hydrolase domain-containing protein, α/β-hydrolase, 2-hydroxymuconic semialdehyde hydrolase	
15	4038	4	16	Acetylhydrolase, esterase, α/β-hydrolase	3.1.1.1
16	13375	21	38	α/β-Hydrolase, lipase	
17	6084	26	15	2-Hydroxyisoflavanone dehydratase, α/β-hydrolase, lipase	3.1.1.1
18	497	5	10	Lipase	
19	2191	5	14	Acetyl xylan esterase	
20	3629	14	5	Lysophospholipase, caffeoylshikimate esterase	3.1.1.5
21	2593	31	35	Lipase	
22	1181	9	2	Peroxidase, protein phosphatase methylesterase, α/β-hydrolase	
23	496	1	2	Lpx1p, hypothetical protein	

Table 1 continued

24	1437	5	6	Platelet-activating factor acetylhydrolase	
25	730	2	4	Carbohydrate esterase family 15	
26	2432	9	2	Gastric triacylglycerol lipase precursor, lipase member M	
27	23684	60	52	Arylacetamide deacetylase (esterase), neutral cholesterol ester hydrolase, lipase	3.1.1.1
28	1553	11	4	Patatin-like protein	
29	1493	10	1	Galactolipase, phospholipase A1, lipase	
30	3896	2	2	Xylanase, 1,4-β-xylanase	
31	1648	1	0	Pheophytinase, α/β-hydrolase	
32	1359	3	14	Protein notum homolog precursor	
33	10812	32	38	60-kDa Lysophospholipase, cytoplasmic asparaginase I, 1-alkyl-2-acetylglycero-phosphocholine esterase	3.1.1.4, 3.1.1.5
34	24560	158	272	Cholinesterase, butyrylcholinesterase, partial acetylcholinesterase	3.1.1.7, 3.1.1.8
35	1059	10	0	Lysophospholipase, phospholipase A2	3.1.1.4, 3.1.1.5
36	688	2	1	Esterase	3.1.1.1
37	4538	4	19	Acetylxylan esterase, esterase, glycoside hydrolase	
38	869	5	0	Diacylglycerol lipase	
Clan B					
39	5264	4	7	Acyl-CoA thioesterase I, multifunctional acyl-CoA thioesterase I, protease I	3.1.1.5, 3.1.1.1
40	1869	2	13	Esterase, lipase, triacylglycerol lipase	3.1.1.1
41	1004	6	56	Cutinase	3.1.1.7, 3.1.1.74
42	1463	14	10	Acetylhydrolase, lipase	
43	713	2	4	Acetylxylan esterase, cutinase	3.1.1.74
44	2262	6	6	Rhamnogalacturonan acetylesterase, GDSL family lipase	
45	1717	15	43	Methyl esterase	3.1.1.1
46	8906	10	15	Polysaccharide deacetylase	
47	1985	5	4	GDSL family lipase	3.1.1.1
48	3765	5	7	2-Pyrone-4,6-dicarboxylate hydrolase,	3.1.1.5

Table 1 continued

				amidohydrolase	
49	2374	9	18	Lactonizing lipase, α, β-hydrolase	3.1.1.32
50	498	7	7	Lecithin-cholesterol acyltransferase	
51	1537	15	15	Lipase	3.1.1.5, 3.1.1.1
Clan C					
52	5655	28	9	Carboxylesterase, acyl-protein thioesterase, phospholipase	3.1.1.5, 3.1.1.1
53	4262	3	18	Monoacylglycerol lipase, carboxylesterase	
54	1818	4	2	N-Acylhomoserine lactonase, α/β-hydrolase	
55	314	1	12	Lipase	
Clan D					
56	7623	9	9	Regucalcin, gluconolactonase	
57	1853	2	3	Lactonase, gluconolactonase	
58	8401	7	5	6-Phosphogluconolactonase	
59	672	16	7	Serum paraoxonase, arylesterase	
60	3060	26	16	Retinoid isomerohydrolase, carotenoid oxygenase	
Clan E					
61	2624	316	279	Phospholipase A2	3.1.1.4
62	738	12	1	Phospholipase A2	3.1.1.4
63	1693	191	194	Group IID secretory phospholipase A2, Group 10, group IIE phospholipase A2	3.1.1.4
Clan F					
64	4051	12	11	Acyl-CoA thioesterase, pectin esterase	
65	7382	19	12	Pectinesterase 1 precursor	
Clan G					
66	5348	10	15	Lactonase, lactamase	
67	1313	3	5	L-Ascorbate 6-phosphate lactonase, β-lactamase	
Not part of a clan					
68	13	1	0	Phospholipase A2	3.1.1.4
69	5037	4	5	Peptidyl-tRNA hydrolase domain protein, peptide chain release factor 1	
70	438	6	0	Putative peptidyl-tRNA hydrolase	
71	8877	13	19	D-Tyrosyl-tRNA(Tyr) deacylase	3.1.1.4, 3.1.1.32

Table 1 continued

72	11	11	0	Phospholipase A1, phospholipase A2	
73	2790	3	1	Chemotaxis protein CheD	
74	1011	7	7	HRAS-like suppressor, phospholipid-metabolizing enzyme, retinoic acid, receptor responder protein	3.1.1.4, 3.1.1.32
75	15	1	0	Hypothetical protein	
76	1024	2	1	Aldehyde dehydrogenase, lipid A 3-O-deacylase	
77	728	6	0	Group XIIA secretory phospholipase A2 precursor, Group XIIB	3.1.1.4
78	11783	16	43	Peptidyl-tRNA hydrolase	
79	3136	92	4	Peptidyl-tRNA hydrolase ICT1, peptide chain release factor I	
80	293	3	0	Tip1p, Tir2p, Tir4p	
81	1234	3	0	Feruloyl esterase, carbohydrate esterase familyl	
82	1334	1	0	Carboxymethylenebutenolidase, dienelactone hydrolase	3.1.1.4
83	34	1	0	Plasmid partitioning protein phospholipase	
84	4983	4	4	Hypothetical protein, β-lactamase	
85	3296	2	7	Phospholipase A1	3.1.1.4, 3.1.1.32
86	2686	8	0	Esterase	
87	43	1	1	Lipase	
88	335	1	0	Phospholipase A	
89	683	1	0	Amidohydrolase	
90	3091	1	0	Lysophospholipase L2	3.1.1.5
91	7012	12	0	GDSL esterase, lipase	
92	12511	8	4	Chemotaxis-specific methylesterase	
93	261	4	0	Chlorophyllase	
94	6298	11	0	GDSL esterase, lipase	
95	2245	3	2	Dihydroorotase, amidohydrolase, metallo-dependent hydrolase	
96	84	1	0	Ldh1p	
97	812	4	1	Rrt2p, iphthamide biosynthesis protein 7	
98	30281	33	163	β-Lactamase, D-alanyl-D-alanine carboxypeptidase penicillin-binding protein	
99	284	1	0	Triacylglycerol lipase	

Table 1 continued

100	7425	17	0	Monoacylglycerol lipase, α/β-hydrolase	
101	6175	11	0	α/β-Hydrolase domain-containing, α/β-hydrolase	
102	2535	1	0	Hydrolase, α/β-hydrolase	
103	688	2	0	Say1p, lipase, thioesterase, α/β-hydrolase fold protein	
104	395	10	9	Hemagglutinin-esterase	3.1.1.5
105	521	3	9	Holyurethanase, hemolysin E	
106	514	3	0	Triglyceride lipase-cholesterol esterase, lysosomal acid lipase, cholesteryl esterase	
107	1828	7	0	Patatin-like phospholipase domain	
108	212	1	0	Triglyceride lipase ATG15	
109	8009	3	0	Hydrolase, proteinase	
110	386	4	1	Senescence-associated carboxylesterase	3.1.1.1
111	192	2	0	Yeh2p, Yeh1p	
112	3312	3	1	Feruloyl esterase, tannase, feruloyl esterase	
113	2953	2	0	Sialate O-acetylesterase precursor	
114	332	2	0	Acyloxyacyl hydrolase	
115	944	13	5	Phospholipase B-like 2	
116	4270	6	20	Serine esterase hydrolase, peptidase	3.1.1.4, 3.1.1.84
117	55	1	0	EstP	
118	799	1	0	Lipase	
119	2137	75	107	Bifunctional xylanase-xylan deacetylase, 1,4-β-xylanase	
120	439	2	0	Lipase	
121	2378	30	0	Phospholipase, hypothetical protein	
122	465	3	0	Cytochrome C1, D-(-)-3-hydroxybutyrate oligomer hydrolase	
123	2129	12	3	Cytosolic phospholipase A2	3.1.1.4, 3.1.1.5
124	2358	17	0	Calcium-independent phospholipase A2, patatin	3.1.1.4, 3.1.1.5
125	1666	5	0	Tgl4p, patatin-like phospholipase	

Table 1 continued

126	8073	11	0	Patatin-like phospholipase domain containing 7, lysophospholipase NTE1, cyclic nucleotide-binding protein	3.1.1.5
127	862	13	0	Phospholipase B1	3.1.1.5

Table 2. Tertiary structural similarity within families.

Family	MSA * count[a]	MSA : count[b]	RMSD	P_{avg}
1	10	9	1.16	96.04
2	3	8	1.39	96.17
3	6	7	1.17	93.34
4	4	2	0.57	98.67
5[c]	0	0	NA	NA
6	0	0	1.71	91.45
7	0	0	1.13	83.59
8	11	4	1.48	96.42
9	10	15	0.40	94.53
10	0	0	1.10	97.01
11	2	2	0.08	100.00
12	2	7	1.28	91.38
13	3	2	0.58	95.90
14	6	13	1.13	95.22
15	3	0	0.40	98.97
16	0	1	1.24	90.63
17	3	1	1.40	91.96
18	10	8	0.63	99.44
19	8	4	0.72	98.57
20	5	7	0.58	95.90
21	4	1	1.43	93.89
22	4	5	2.47	53.07
23	3	3	0.47	100.00
24	0	1	0.41	99.79
25	8	2	0.50	99.72
26	3	4	0.87	90.45
27	3	1	1.27	89.85
28	0	2	0.53	100.00
29[c]	9	8	NA	NA
30	4	2	1.24	92.53
31[c]	4	2	NA	NA

Table 2 continued

32	1	0	0.62	97.63
33	2	9	1.26	54.93
34	8	7	0.98	98.73
35[c]	0	0	NA	NA
36[c]	3	8	NA	NA
37	2	1	1.43	65.42
38[c]	6	5	NA	NA
39	0	0	0.76	98.62
40	3	5	0.3	99.91
41	11	8	0.42	98.80
42	3	1	0.32	99.43
43	5	4	0.38	100.00
44	3	2	0.52	91.36
45	4	12	0.08	100.00
46	0	1	0.08	100.00
47	2	5	0.61	99.44
48	3	4	1.10	97.01
49	2	6	1.04	85.35
50	1	7	0.53	99.91
51	0	0	0.49	98.11
52	3	4	1.15	95.08
53	3	1	0.92	94.13
54	3	3	0.09	100.00
55	5	9	0.43	97.88
56	4	3	0.57	99.62
57	3	4	0.26	98.44
58	1	1	0.86	98.63
59	10	9	0.68	97.00
60	0	0	0.81	97.76
61	11	1	1.33	96.85
62[c]	14	6	NA	NA
63	13	2	0.92	97.32
64	4	3	0.51	95.74
65	4	2	0.76	96.80
66	5	5	0.09	100.00

Table 2 continued

67	1	0	0.46	90.30
68[c]	10	3	NA	NA
69	9	6	1.64	73.69
70[c]	7	9	NA	NA
71	11	10	0.6	98.67
72[c]	8	6	NA	NA
73[c]	4	2	NA	NA
74	2	11	1.17	84.35
75[c]	3	5	NA	NA
76[c]	9	2	NA	NA
77[c]	27	14	NA	NA
78	12	13	1.38	93.98
79	6	7	1.25	100.00
80[c]	4	8	NA	NA
81[c]	3	2	NA	NA
82[c]	5	2	NA	NA
83[c]	1	1	NA	NA
84	2	2	1.13	83.59
85	0	1	0.40	94.53
86[c]	4	2	NA	NA
87[c]	7	9	NA	NA
88[c]	6	3	NA	NA
89	15	6	NA	NA
90[c]	1	1	NA	NA
91	3	2	NA	NA
92	7	4	1.38	93.43
93[c]	5	9	NA	NA
94[c]	3	3	NA	NA
95	0	0	1.45	92.18
96[c]	15	11	NA	NA
97[c]	10	6	NA	NA
98	2	1	0.87	90.45
99[c]	6	8	NA	NA
100[c]	3	0	NA	NA
101[c]	0	1	NA	NA

Table 2 continued

102[c]	3	10	NA	NA
103[c]	6	3	NA	NA
104	51	32	1.27	92.78
105	20	33	0.77	97.79
106[c]	4	5	NA	NA
107[c]	0	3	NA	NA
108[c]	8	7	NA	NA
109[c]	2	4	NA	NA
110[c]	4	5	NA	NA
111[c]	9	11	NA	NA
112[c]	2	7	NA	NA
113[c]	11	8	NA	NA
114[c]	55	43	NA	NA
115	0	0	0.76	63.81
116	4	0	0.84	95.28
117[c]	7	13	NA	NA
118[c]	0	0	NA	NA
119	1	3	1.32	85.78
120[c]	27	34	NA	NA
121[c]	1	4	NA	NA
122[c]	1	1	NA	NA
123	8	8	1.00	99.60
124[c]	4	3	NA	NA
125[c]	4	5	NA	NA
126[c]	5	3	NA	NA
127[c]	4	5	NA	NA

[a] Total conservation of amino acid residues over multiple sequence alignment.

[b] Total conservation of chemically similar amino acid residues over multiple sequence alignment.

[c] Zero or one known tertiary structure in this family.

Table 3. Tertiary structural similarity within clans.

Clan	Number of families	CEH families	RMSD (Å)	P_{avg} (%)	Fold
A	37	1–38	2.46	25.51	α,β-Hydrolase, 2nd β-strand antiparallel
B	13	39–52	2.47	54.39	α,β-Hydrolase, all β-strands parallel
C	4	53–56	2.47	54.39	α,β-Hydrolase, 1st β-strand antiparallel
D	6	57–62	2.21	70.94	6-Propellor
E	4	63–66	1.62	52.43	3-α-Helix
F	2	67–68	1.39	90.54	3-Solenoid
G	2	69–70	2.24	70.61	4-Layer sandwich

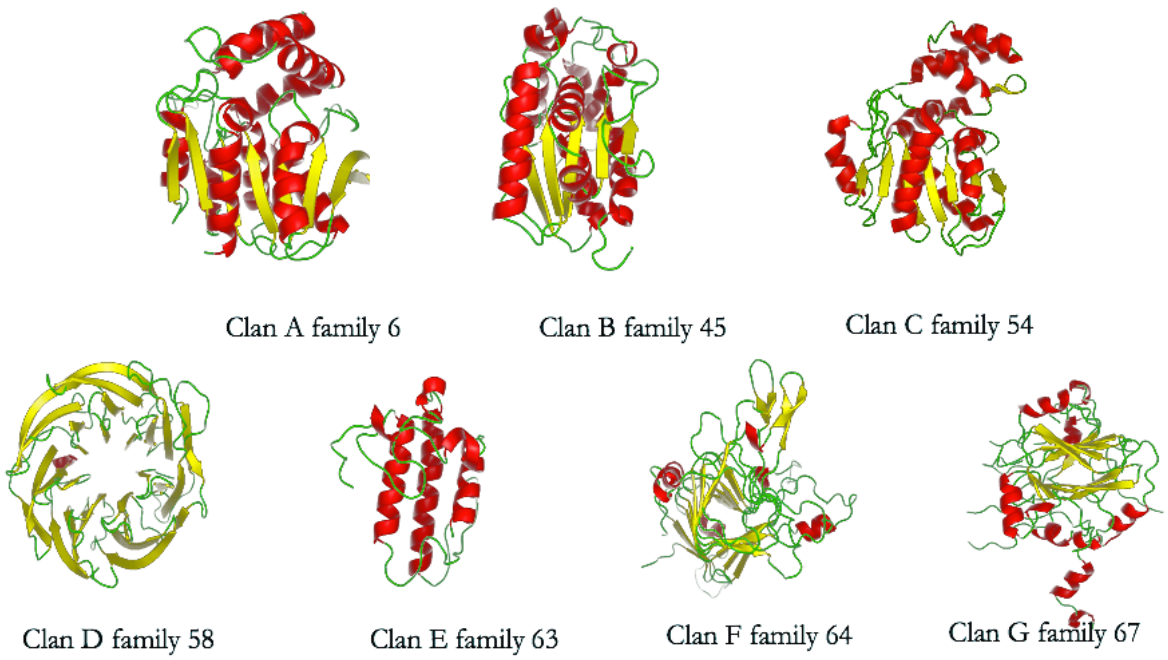


Figure 1. Tertiary structures from seven clans. They are colored by secondary structures. Red indicates α-helices, and yellow indicates β-strands. PDB IDs for each clan above are: clan A: 1VA4, clan B: 4PSD, clan C: 3BF8, clan D: 3FGB, clan E: 1C1J, clan F: 1QJV, and clan G: 2WYL.

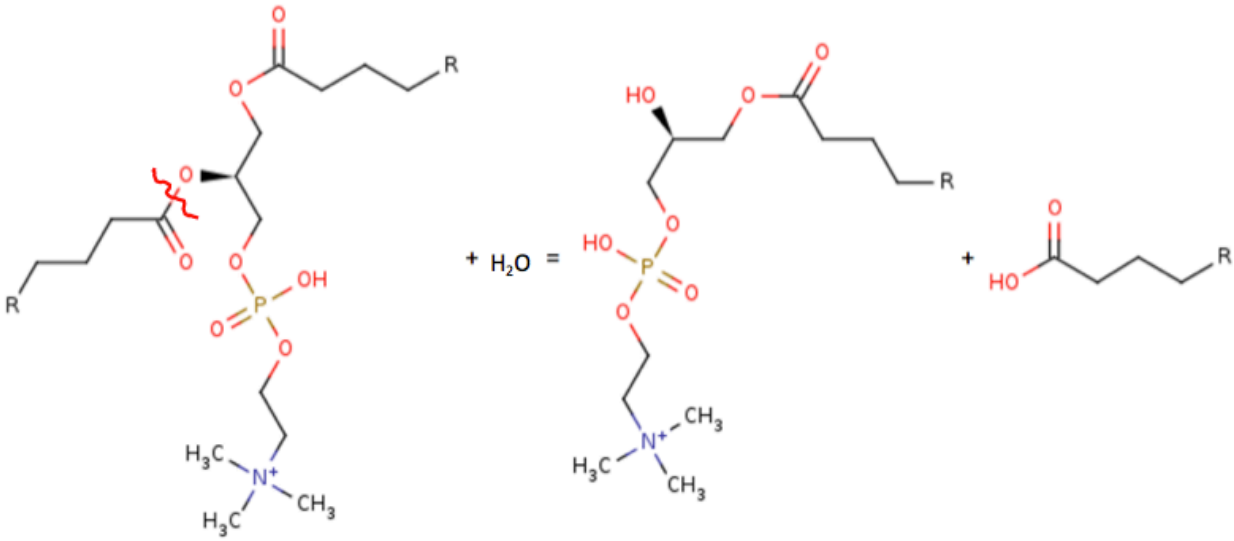


Figure 2. Reaction catalyzed by phospholipase A2 (EC 3.1.1.4). Red curve indicates the bond to be hydrolyzed.

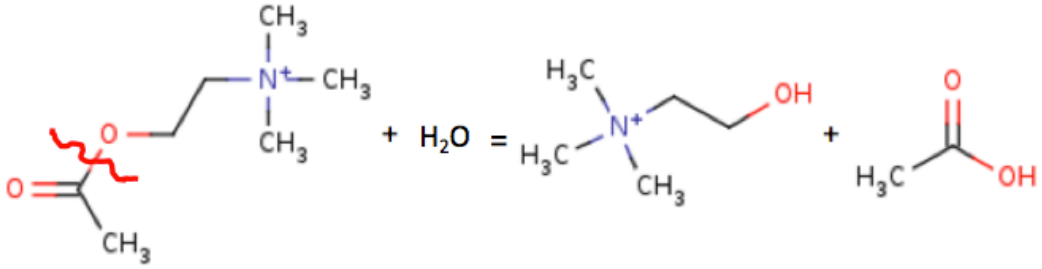


Figure 3. Reaction catalyzed by acetylcholinesterase (EC 3.1.1.7). Red curve indicates the bond to be hydrolyzed.

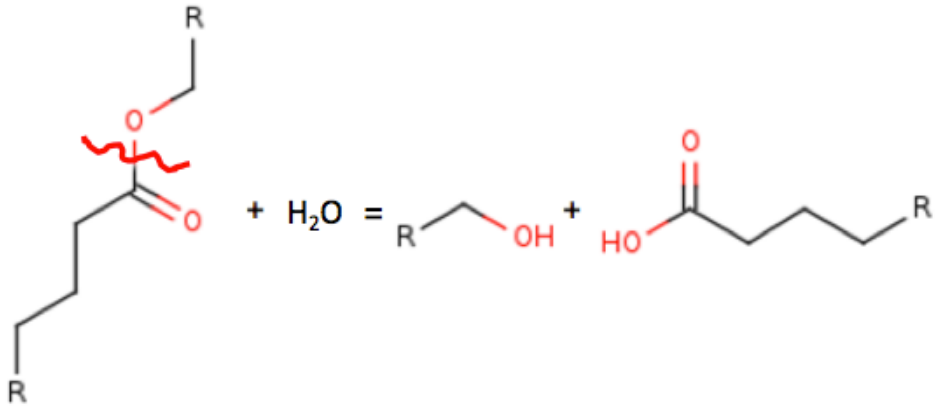


Figure 4. Reaction catalyzed by carboxylesterase (EC 3.1.1.1). Red curve indicates the bond to be hydrolyzed.

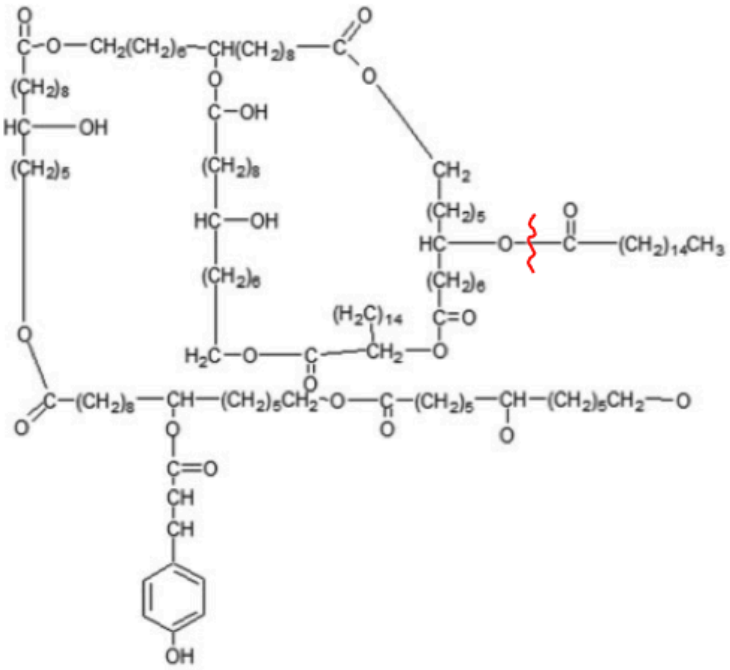


Figure 5. Cutin structure. Red curve indicates the bond to be hydrolyzed by cutinase (EC 3.1.1.74). Figure adapted from "Identification and characterization of bacterial cutinase" by Chen S et al., 2008, J Biol Chem, 283(38), 25855.

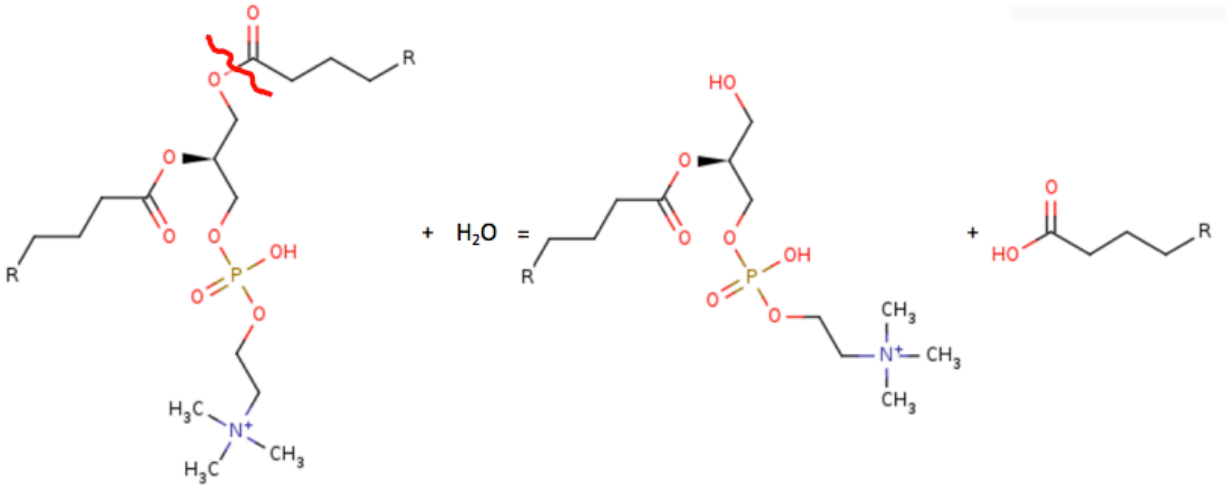


Figure 6. Reaction catalyzed by phospholipase A1 (EC 3.1.1.32). Red curve indicates the bond to be hydrolyzed.

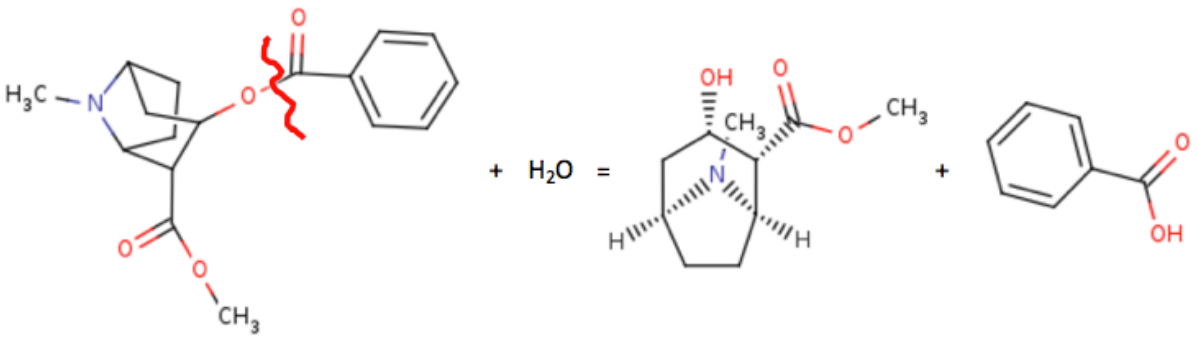


Figure 7. Reaction catalyzed by cocaine esterase (EC 3.1.1.84). Red curve indicates the bond to be hydrolyzed.

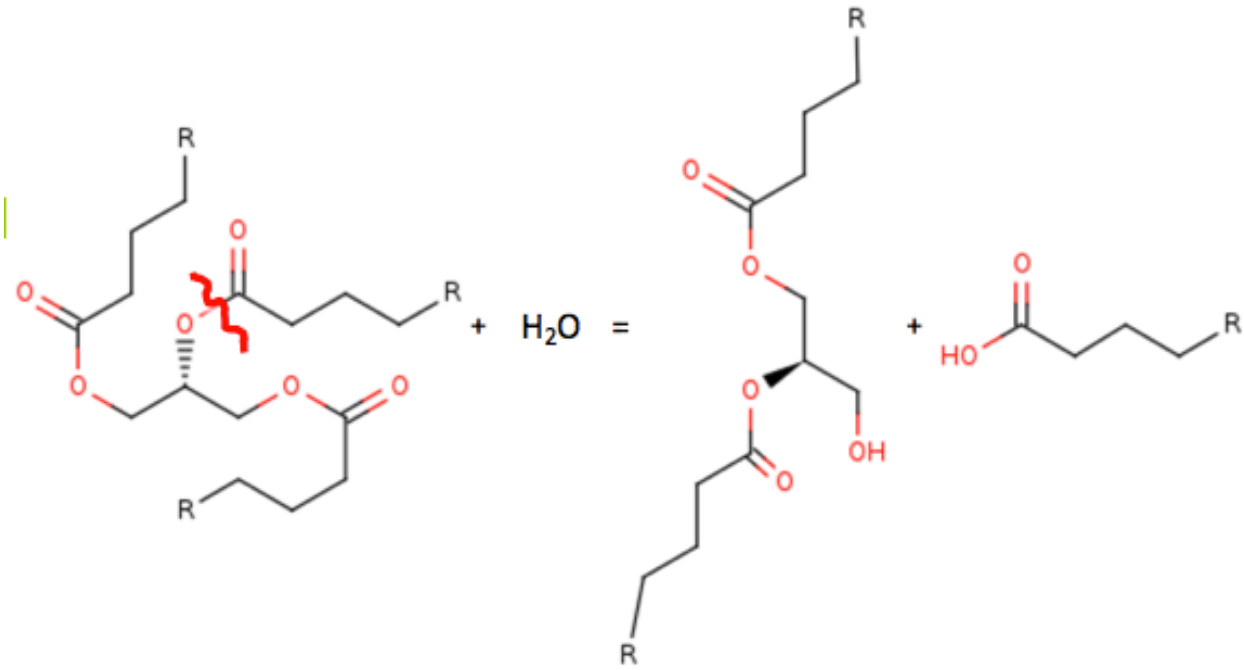


Figure 8. Reaction catalyzed by triacylglycerol lipase (EC 3.1.1.3). Red curve indicates the bond to be hydrolyzed.

APPENDIX. LEARNING PROTEIN CRYSTALLIZATION CONDITIONS: ANALYSIS, OPTIMIZATION, AND APPLICATION

Introduction

Protein crystallization is affected by many factors, including chemical parameters like pH, precipitant, buffer, additives, and ligands, physical parameters like temperature, pressure, time, gravity field, and electric field, and protein parameters like purity, concentration, and mutation (Giegé, 2013). It is common for researchers to test hundreds or thousands of combinations of conditions to find the appropriate crystallization condition for one protein, and this remains a challenging and time-consuming process. Finding the conditions to crystallize proteins is a bottleneck for scientists to obtain protein tertiary structures, given the substantial gap between the number of protein sequences available from genome projects and the number of solved protein tertiary structures (Wooh et al., 2003) (Zhu et al., 2006) (Newstead et al., 2009) (Zucker et al., 2010) (Parker & Newstead, 2012).

By studying protein crystallization conditions statistically, we hope to discern the relationship among these factors and protein properties. We aim to reduce and optimize the crystallization condition sets, in order to decrease the number of trials that researchers need to do before crystallizing a protein. This will save both time and resources for researchers, especially in the small to medium laboratory scale.

The greatly increasing number of tertiary structures available on the Protein Data Bank (PDB) (Bernstein et al., 1977), (Berman et al., 2000) provides a great resource to extend information about protein tertiary structures.

The detailed objectives of this project were as follows:

(a) Parse the crystallization condition data on the PDB from plain text format to tabular format, to make it easier for further analysis.

(b) Analyze the relationships among protein crystallization conditions, protein structures, and other protein properties using statistical methods, based on the protein tertiary structure information available on the PDB.

(c) Optimize protein crystallization conditions of particular kind of proteins, to best represent the independent condition sets that need to be explored, in order to improve the success rate of crystallization.

Protein tertiary structure and X-ray crystallography

Protein tertiary structures, or three-dimensional structures, provide important clues about enzyme functions, such as their active sites, substrate specificities, and reaction mechanisms. Tertiary structures with different precision can give various kinds of biological information. For example, the most precise structures, of 1.0 Å resolution, can yield clues about reaction mechanism. Structures of 1.5 Å resolution can be used to guide site-directed mutagenesis to reveal sequence-structure-function relationships, and to study active sites and binding pockets. Structures of 3.5 Å resolution, although low in resolution, can still provide enough information to study enzyme functions (Eswar et al., 2006).

Among the various experimental methods used to determine protein tertiary structures, X-ray crystallography is the most widely used, followed by nuclear magnetic resonance (NMR) and electron microscopy (EM). X-Ray crystallography contributed 84,406 (89.4%) of the 94,415 protein structures available on the PDB, while NMR and EM yielded 9,232 (9.8%) and 560 (0.6%) structures, respectively (statistics obtained in August 2014). The main research objects of this project deal with the crystallization conditions of proteins solved by X-ray crystallography, yet comparing protein structures determined by X-ray, NMR, and EM will occur as well.

Protein crystallization is the critical step before a sample is submitted for an X-ray crystallographic study. The protein solution first needs to be purified in high concentration. Then the protein solution is brought into a supersaturated state, and hopefully crystals will start to

form. The success of this effort depends on many factors, including protein purity and concentration, precipitant, buffer type and concentrations, ligands, pH, temperature, pressure, time, magnetic field, and electric field in some rare cases.

Vapor diffusion methods, such as the hanging drop and sitting drop techniques, are the most popular methods to obtain protein crystals for X-ray crystallography. Figure 1 shows the hanging drop method (Drenth, 2007). The hanging drop on the top of the sealed container usually has a lower precipitant concentration than that in the bottom reservoir. Precipitants bind with water and compete with the protein for it, increasing the protein concentration in the hanging drop as water migrates to the bottom reservoir until its vapor pressure becomes equal in the hanging drop and the bottom reservoir. Eventually the protein concentration in the drop will become supersaturated. If all the other parameters like pH, buffer, and temperature are appropriate, the protein will start to crystallize.

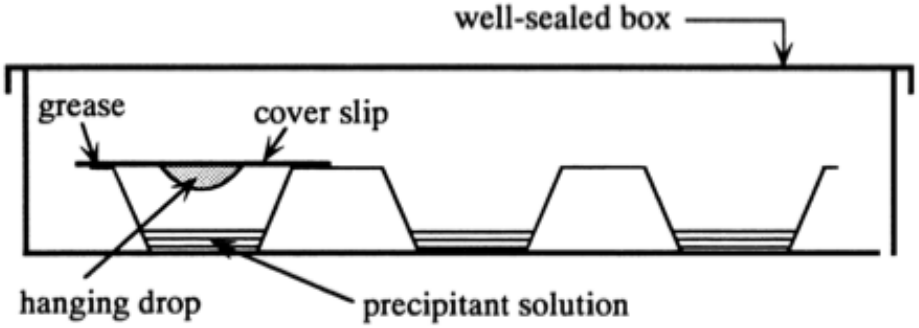


Figure 1. The hanging drop method for protein crystallization. The grease ensures the sealed environment. Figure adapted from *Principles of Protein X-ray Crystallography*. Springer, New York (Drenth, 2007).

Batch crystallization is the technique that mixes protein and reagents directly to create a supersaturated solution, and the mixture is covered by oil to keep it isolated from the environment. When the system reaches equilibrium, crystals should start to grow. This method is suitable for automation and miniaturization, and it is known as the microbatch technique. Dialysis is another technique used to change the concentration of precipitants. It is suitable to grow large crystals, but it is hard to miniaturize (Rupp, 2010).

Previous studies on protein crystallization conditions

Giegé (2013) summarized the parameters of crystallization conditions extensively in his review about protein crystallization history to the present day. Precipitants, buffers, pH, temperature, ligands, additives, and detergents for membrane proteins are the main parameters for protein crystallization. Pressure, temperature, time, gravity field, and electric field also affect crystallization. Last but not least, the protein itself can be regarded as the most significant parameter. This includes the purity, concentration, mutation, truncation, and deletion of the protein (Dale et al., 2003).

Precipitants are compounds that bind water and compete with proteins for it, so that the proteins have more difficulty in accessing water, and the protein concentration is considered to be higher than those in the same amount of water without precipitant. Polyethylene glycol (PEG) (Figure 2) is a widely used polymer precipitant. Salts, organic molecules, and ionic liquids, such as ammonium sulfate and 2-methyl-2,4-pentanediol (MPD), are also used as precipitants. Buffer is employed to maintain a certain pH of the protein solution, and to provide a specific local charge distribution of the protein. The latter contributes significantly in the intermolecular interaction leading to crystal formation (Rupp, 2010). Additives include everything else that promotes crystallization, and they are added when there is protein aggregation during crystallization, small-sized crystals, or weak diffraction in X-ray crystallography.

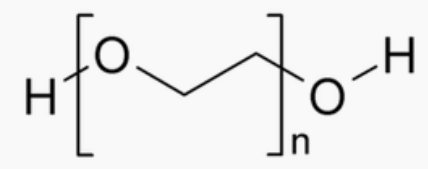


Figure 2. Chemical structure of PEG. This polymer has molecular weights from 300 g/mol to 10,000,000 g/mol. PEG used as precipitant usually has mean molecular weights from 300 to 10,000 g/mol. For example, PEG 400 refers to PEG that has a 400 g/mol mean molecular weight.

Continual efforts have been devoted to protein crystallization since the first discovery of hemoglobin protein crystals in blood in 1840 by Hunefeld (Giegé, 2013). Several studies have emerged since the 1990s to try to rationalize protein crystallization conditions, although these various parameters mentioned above make it impossible to exhaust all the combinations of crystallization conditions (Luft et al., 2011). Jancarik and Kim (1991) developed a screening method, sparse matrix sampling, based on published crystallization conditions. This sampling method has three major variables: pH and buffer materials, precipitants, and additives. Statistically, it uses the Carter and Carter (1979) incomplete factorial method to decrease the number of screening conditions. Fifty conditions were proposed to effectively cover the wide range of pH, precipitants, and additives. Fifteen previously crystallized proteins are used as test data. Crystals were obtained successfully from all of the proteins by using at least one of these 50 conditions. The Jancarik and Kim screening method is widely used and remains popular, as it assumes no *a priori* knowledge for the protein to be crystallized, and it can be applied for proteins with limited information on their properties. Yet this is occasionally not as effective as it is in other circumstances: when protein properties are known, it does not incorporate the information into the crystallization conditions.

Some popular crystallization condition kits were also commercialized and available from several companies, such as Qiagen, Molecular Dimensions, and Hampton Research. The Joint Center for Structural Genomics (JCSG) tested 480 commercially available conditions for the *Thermotoga maritima* proteome, found the redundancy in these commercial conditions, and minimized them to the 67 most effective ones (Page et al., 2003). A systematic study of pH, anion and cation-testing (PACT) screening conditions with PEG were also developed and tested (Newman et al., 2005). Their research results were converted into the JCSG+ and PACT commercial kits.

The correlation between protein isoelectric point (pI) and pH of the protein crystallization solution was investigated (Kantardjieff & Rupp, 2004). A total of 9,596 unique protein crystals from the PDB were studied, and a significant relationship ($R^2 = 0.62$), although not a dir-

ect correlation, was found between pI and the difference between pH and pI (pH – pI). The pI values are calculated using the pK_a values of Bjellqvist et al. (1993). Based on this relationship, a prototype pH range calculator (CrysPred) was developed. This server-based tool aims to optimize the efficiency of initial crystallization screening conditions, with a predicted saving of material of 30 to 50%.

Membrane proteins, of rising interest due to their relevance with human diseases and medicines, are studied individually for their crystallization conditions (Newstead et al., 2009), (Parker and Newstead, 2012). Compared to soluble proteins, membrane proteins need detergents to isolate and solubilize them, which adds an additional parameter to the crystallization process. Membrane proteins were classified by their functions into eight groups, such as channels, transporters, and receptors. The variables, including detergents, precipitants, buffers, pH range, and salts, are visualized against the number of successful crystallizations. Their analysis led to the commercial screening kits MemGold (2009) and MemGold2 (2012), designed for membrane proteins.

The Biological Macromolecule Crystallization Database (BMCD) (Gilliland et al., 1996), (Tung & Gallagher, 2009) is a manually curated database that provides detailed information about crystallization since 1988. The crystallization details, collected from the literature, are listed as macromolecular concentration, pH, temperature, and growth time, while crystallization solutions are recorded as reagent type, concentration, and dimension. Several crystallization studies have been based on the data obtained from the BMCD. Cluster analysis of crystallization parameters, including pH, temperature, molecular weight, macromolecular concentration, precipitant type, and crystallization methods, has been conducted based on BMCD data (Samudzi et al., 1992), (Farr et al., 1998), and XtalBase (Meining, 2006), and is a web-based tool to generate new condition sets for crystallization experiments.

On the other hand, due to the exponential growth in available tertiary structures, information on crystallization conditions leading to these structures expands rapidly. The PDB has become the comprehensive resource when researchers need information about crystallization

conditions, as the BMCD has been updated less frequently as of late. However, all the crystallization conditions in the PDB are recorded in plain text but in varying formats. For instance, ammonium sulfate has more than 30 spelling alternatives (Peat et al., 2005). The variation in chemical names and dimensions has hampered the further data analysis of crystallization conditions. Therefore, Newman and colleagues (2014) constructed a standard dictionary to map chemical names to their aliases, together with their common classes and dimensions, and proposed a rule for standard nomenclature used in macromolecular crystallization.

Glycoproteins

Glycoproteins are proteins that have sugars, or glycans, covalently attached to protein side chains. About half of all human proteins may be glycosylated (Apweiler et al., 1999). There are two common types of glycosylation: N-glycosylation and O-glycosylation. N-glycosylation often occurs where the protein has the specific sequence Asn–X–Ser/Thr/Cys–X, where X is any amino acid except proline. O-glycosylation often exists in an area of the protein with large numbers of serine, threonine, and proline residues (Nettleship, 2012).

In terms of structural biology, various glycans that increase the heterogeneity of the protein surface can affect the protein structure. Furthermore, the overall glycan mass can be from 1% to 80% of the glycoprotein total mass (Varki et al., 2009**).** The variation of glycan type and number may in some cases hamper protein crystallization. On the other hand, glycan presence can sometimes benefit crystallization when it allows intermolecular contact.

Statistical methods of data mining

The PDB stores plenty of information in each PDB file, such as the protein name, protein function by EC number (Webb, 1992), structure classification by CATH (Sillitoe et al., 2012) and SCOP database (Murzin et al., 1995), as well as methods and conditions to obtain crystals and the three-dimensional coordinates of each atom. With PDB entries growing exponentially, we have more structural information available than ever before. Such large databases

usually contain hidden knowledge, and they can be further investigated by statistical methods, such as by so-called data mining. Data mining involves using knowledge from statistics, database management, computer science, and machine learning (Fayyad et al., 1996).

Supervised learning and unsupervised learning are two common algorithm types in machine learning methods. Supervised learning uses data with predefined output as a training dataset, to generate functions to map input data with known output data. Then it uses generated functions to predict the output value for a new input dataset. Widely used supervised learning algorithms are support vector machine (SVM) (Cortes & Vapnik, 1995), k-nearest neighbor (KNN) (Altman, 1992), and neural network (NN) (Hagan et al., 1996).

Unsupervised learning tries to find hidden patterns in previously undefined data. Unlike supervised learning, unsupervised learning does not have predefined output data, so it uses computation to find the groups of similar input data on its own. This kind of grouping is called clustering, where hierarchical clustering (Sibson, 1973) and centroid-based clustering (Lloyd, 1982) are two common clustering methods.

Previous studies on biological problems using data mining

Researchers have successfully applied supervised and unsupervised learning techniques to study various biological problems. Protein structure classification (Krishnaraj and Reddy, 2010) and transmembrane protein topology prediction (Jones, 2007) are two data mining examples related to protein structures. These two examples will be summarized, focusing on the way to pre-process the dataset, the methods used to perform data mining, and the evaluation of the results inferred from the biological data, in the following paragraphs.

Protein structure classification sorts protein tertiary structures into corresponding protein folds. This can be approached using various supervised learning methods, such as SVM, KNN, NN, and boosting. The accuracy and efficiency of these methods are compared in the work of Krishnaraj and Reddy (2010). The dataset is from the SCOP database (Murzin et al.,

1995): a training set of 311 proteins with no more than 35% sequence identity from 27 representative SCOP fold are selected, and the test data are 383 proteins from the same 27 folds with no more than 40% sequence identity, excluding the training set. The parameters, or features, of the datasets are: amino acid composition, predicted secondary structure, hydrophobicity, polarity, normalized van der Waals volume, and polarizability. For each protein, these six parameters are extracted. Then SVM, KNN, NN, and boosting methods are applied to classify the protein structures. The measurement to evaluate the classification is the standard Q percentage accuracy. Q_i equals the number of correct predicted proteins in fold i divided by total number of test proteins in number n, and Q equals the weighted average of individual fold accuracy Q_i. Using the evaluation criteria, the boosting method performs better than other supervised methods to solve the protein structure classification problem.

Predicting transmembrane protein topology is another problem to which the machine learning technique has been applied. Jones (2007) used the neural network methods in MEMSAT3 to predict the secondary structure of transmembrane proteins, and this method has better accuracy (80%) than other methods that have 62–72% accuracy. MEMSAT3 combines the existing MEMSAT2 methods with the sequence conservation information using neural network methods. As for the dataset, it uses 184 proteins where their topology is known experimentally. Each of the 184 proteins is used as a separate training set to allow the cross-validation. To evaluate the results, three criteria are applied: The number of transmembrane helices, the topology, and the location of the transmembrane regions were all correctly predicted. The results were plotted to compare MEMSAT3 and other four methods, including MEMSAT2. The data set was also divided into subsets by organism and single- or multiple-spanning proteins to show their accuracy in subsets using different methods. The evaluation is also illustrated in plots such as false-positive rate on identifying globular proteins and false-negative rate on transmembrane proteins.

Methods

Crystallization data acquisition

Protein crystallization data and information about their corresponding protein properties were obtained from the PDB. Protein structures determined by X-ray crystallography were of research interest. If more than one PDB entry with exactly the same sequence exists, only one of the entries was included in the data set. Customized reports of crystallization details and various protein properties were downloaded from the PDB. The protein properties range from structure resolution and source organism, to structure type from the CATH protein structure classification database (Sillitoe et al., 2012). Table 1 contains a full list of the protein properties in our data set.

Table 1. Protein attributes obtained from the PDB.

Structure summary	PDB ID	Chain ID	Structure Title	Release Date
	Resolution	Structure MW	Residue Count	Atom Site Count
	Ligand ID	Ligand MW	Ligand Formula	Ligand Name
Biological Details	Plasmid	Source	Taxonomy ID	Biological Process
	Cellular Component	Molecular Function	EC No	Expression Host
Domain details	CATH ID	Cath Description	SCOP ID	Scop Fold
	Pfam ID	Pfam Description		
crystallization	Exp. Method	Crystallization Method	Crystal Growth Procedure	
	Temp (K)	pH Value		

The crystallization details to experimentally obtain the protein structure, such as precipitant and buffer type and their concentrations, are recorded in plain text in the crystal growth procedure in the PDB. Not every PDB entry has detailed crystal growth procedures, thus entries of fewer than seven characters in this field were discarded, because they had no or a limited amount of information, like the pH or temperature, while these values can be found separately elsewhere in the record.

Data preprocessing

In order to recognize and separate different reagents into precipitant, buffer, salt, additive,

and detergent categories and their concentrations, the raw PDB data were processed by a series of in-house Python scripts, described as follows. A comprehensive list (plist) of precipitants, buffers, salts, additives, and other reagents was created to guide the text separation. It was summarized based on the common types of known reagents, and every entry in the list became the keyword in the following search. Based on the standard names and their alias summarized by Newman and her colleagues (2014), a name list was created to map various names from the PDB into standard chemical names of our data set. The keywords of the plist are formatted into standard name from the name list. Next, the text of crystal growth procedure downloaded from the PDB was first divided into pieces by commas, and then each piece was searched by the keywords and their aliases in the name list. If any keyword or alias was found, the text before the name was checked for the reagent concentration and its dimension, and its standard name instead of the alias was recorded. In this way, the results of the crystallization conditions were formatted and written into a new CSV (comma-separated values) file. If multiple chemicals exist in one reagent class, for example, if two chemicals are used as additives, they are written in one entry into the result file, but will be separated later in the data analysis. Finally, the formatted data were combined with protein properties from the PDB, and they were used for further analysis using Python and the data mining software WEKA (Witten et al., 2011).

Preliminary results

Crystallization condition data format

The plain text of the materials added to further the crystallization of PDB entries were classified into precipitants, buffers, salts, additives, and detergents, if available. Their concentrations and dimensions are also recorded, if provided, into a CSV file, with some exceptions to be finished in future work. This unified format helps the further analysis of data mining of the crystallization conditions.

CATH groups and crystallization

The CATH database classifies protein structures (Sillitoe et al., 2012) into four hierarchies: class, architecture, topology, and homologous superfamilies. Class is the top hierarchy, divided based on secondary structure. Class 1 holds mainly α-helices, class 2 has mainly β-strands, class 3 contains α and β structures, and class 4 includes a few irregular secondary structures. Within each class, their architectures are then classified based on the arrangement similarity of secondary structure on three-dimensional space. Five, twenty, fourteen, and one architectural groups exist within the four classes, respectively. In the topology levels, the connectivity between secondary structures is considered. Finally, homologous superfamilies group proteins according to similar structures, sequences, or functions.

The number of PDB protein structures determined by X-ray and NMR is listed in the forty second-level CATH groups. The total number of structures with CATH classification found by X-ray and NMR are about 16,400 and 2,800, respectively. The three-layer (α-β-α) sandwich (CATH 3.40) is the most common structure determined by X-ray, followed by the orthogonal bundle (CATH 1.10) and the two-layer sandwich (CATH 3.30). The latter two CATH groups are also the most popular groups determined by NMR. Within the top ten popular CATH groups for each method, X-ray and NMR share eight of them. On the other hand, tertiary structures of α-β barrel (CATH 3.20) and α-β complex (CATH 3.90) proteins are mainly determined by X-ray crystallography, whereas structures of proteins of irregular secondary structure (CATH 4.10) are determined mainly by NMR.

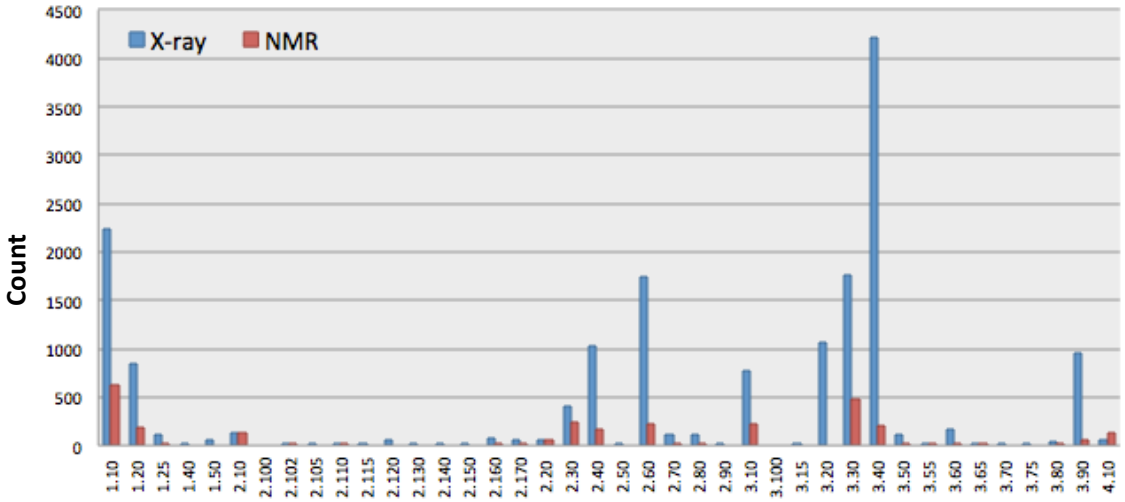


Figure 3. The number of PDB protein structures in second-level CATH groups determined by X-ray (in blue) and to NMR (in red).

The molecular weights of PDB structures by CATH groups are shown in Figure 4. The average molecular weight of those determined by X-ray covers a wide range, from 20 kDa to 160 kDa. Samples subjected to X-ray have higher molecular weights than those subjected to

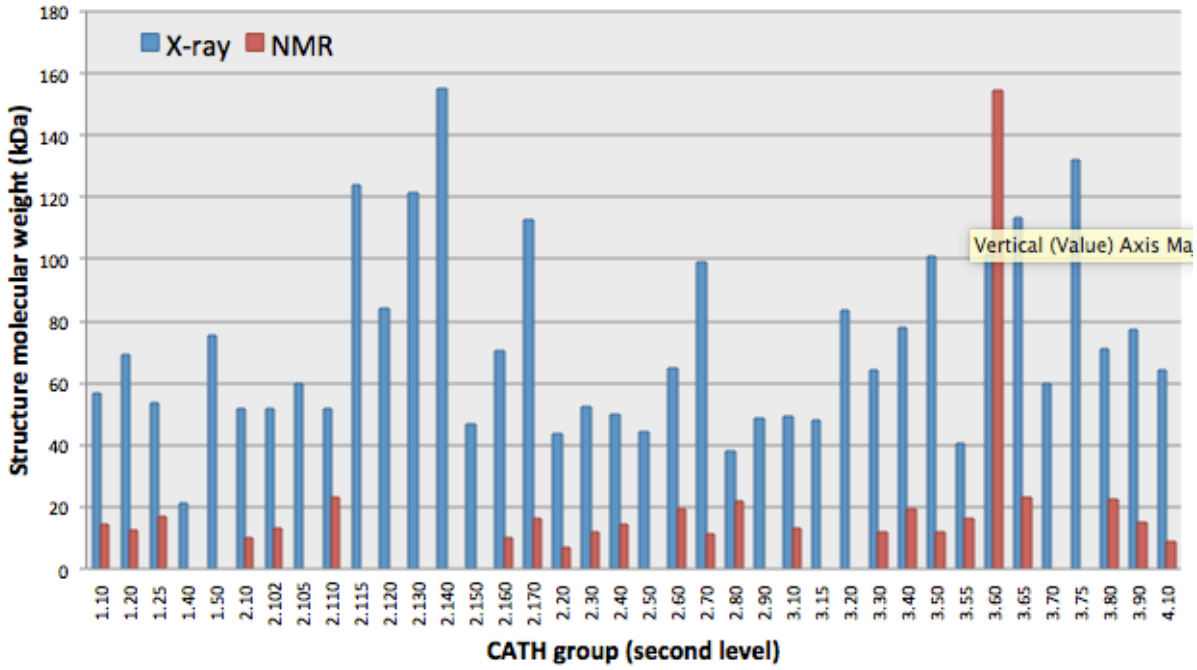


Figure 4. The average structure molecular weight (kDa) in second-level CATH groups determined by X-ray (in blue) and NMR (in red).

NMR in most cases, as the latter is known to be limited to solving protein structures less than 50 kDa. The only exception in this case is CATH 3.60, where two proteins, 2KU1 and 2KU2, were solved by TROSY-NMR (Religa et al., 2010) and overcame this limit. Since these two proteins are the only ones solved by NMR in CATH 3.60, an exceptional average molecular weight results (Figure 4). Other groups determined by NMR have average molecular weights between 10 kDa and 20 kDa.

The resolutions of various CATH groups are shown below in Figure 5. Generally, the median resolutions (red dashes) of second-level CATH groups are around 2 Å. The groups of

α-solenoids (CATH 1.40) and three-layer sandwiches (CATH 2.102) have resolutions around 1.5 Å. CATH groups with more outliers (blue crosses) such as CATH 1.10, CATH 2.40, and CATH 3.40, overlap with the groups with higher number of X-ray crystal structures in Figure 4, which indicates their wide range of resolution.

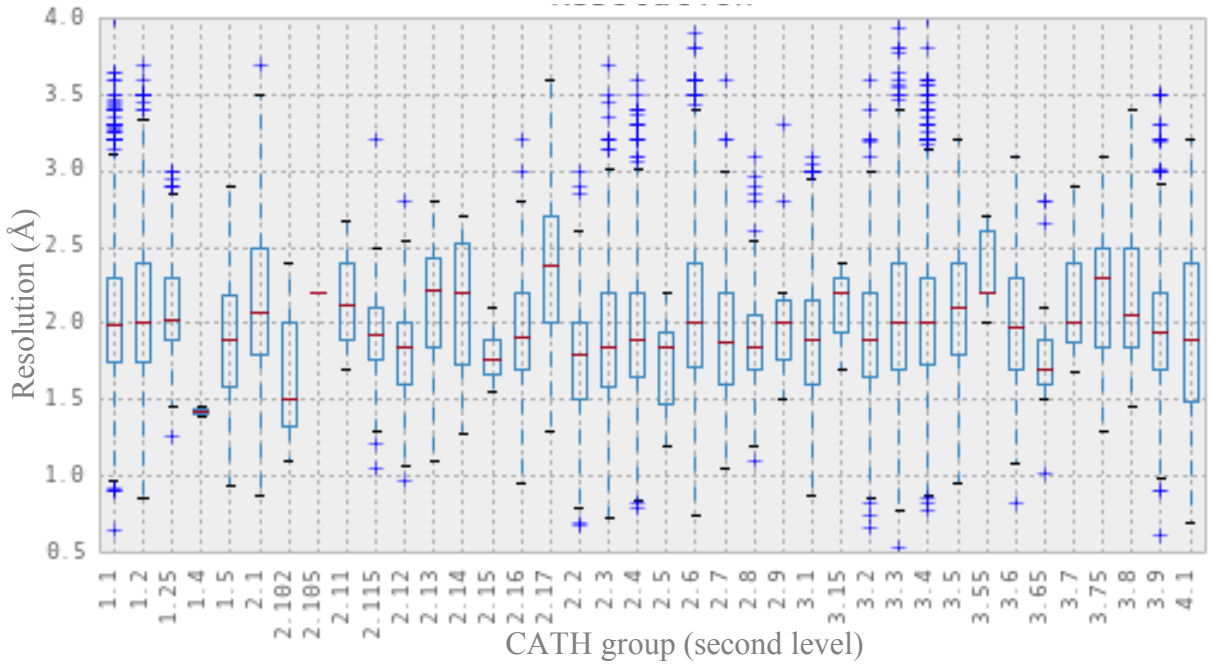


Figure 5. The boxplot of resolution (Å) in second-level CATH groups.

Prediction of CATH groups by crystallization conditions

This is based on the hypothesis that crystallization conditions differ in crystallizing various tertiary structures, classified by CATH groups. Thus, computation can classify successfully crystallized proteins into CATH groups by their crystallization conditions and protein properties such as molecular weight. If tertiary structures in different CATH groups can be correlated with various crystallization conditions, we can infer that proteins in different CATH groups prefer diverse crystallization conditions.

All proteins with available CATH group classification from the PDB, totaling 15,428 PDB entries, were downloaded. The crystallization conditions were parsed into precipitant, buffer, temperature, and pH values using Python scripts. The information was imported to the data mining software WEKA (Witten et al., 2011) for further analysis. Supervised machine

learning was performed, as we use the forty second-level and four first-level CATH groups, as the labels. Several methods were implemented and evaluated by tenfold cross-validation. This validation method divided the dataset into ten subsets, with nine sets used to train the model, and the tenth subset used to test the correctness of the model, for ten iterations.

Since many machine learning algorithms exist, algorithms from each kind that are suitable for the datasets are implemented and tested. Decision trees are constructed by dividing an attribute into branches for each possible value, and dividing each branches recursively, until all instances in the nodes have the same classification. Rule methods examine each class in turn and try to cover all the instances in it, and this is reversed to the top-down methods of decision tree. Function classifiers are those that can be written naturally in mathematical equations. Bayesian classifiers are those use Bayes theorem explicitly to solve problems. Ensemble methods use multiple models to train the data independently and combine them in some way to improve the prediction (Witten et al., 2011).

Table 2. The methods used to predict CATH groups and cross-validation results. Five attributes include precipitant, buffer, temperature, pH, and second-level CATH group ID.

Methods	Correctly classified instances, %	Incorrectly classified instances, %	Kappa statistic	ROC area
Decision tree - J48	26.5	73.5	0.0084	0.521
Bayes - Naïve Bayes	25.0	75.0	0.0014	0.509
Lazy - IBK (kNN)	18.0	82.0	0.0279	0.530
Rule - OneR	26.6	73.4	0.0133	0.506

As shown in Table 2, the crystallization conditions successfully predicted second-level CATH groups < 30% of the time by each of the four different methods. The kappa statistics correct the overall success rate by deducting the success rate occurring by chance, as shown below (Witten et al., 2011):

$$\text{Kappa} = (\text{observed accuracy} - \text{expected accuracy})/(1 - \text{expected accuracy})$$

The observed accuracy is the correctly predicted instances for all classes divided by the total number of instances. The expected accuracy is the accuracy that any random predictor

can occur by chance. Kappa statistics over 0.75 are considered excellent, 0.4 to 0.75 as fair to good, and lower than 0.4 as slight (Fleiss, 1981). The ROC curves, or receiver operating characteristic curves, are plotted by the true positive rate against the false positive rate, as shown below. The ROC area refers to the area under ROC curves, where one is a perfect classification, and 0.5 means a random guess.

$$\text{True positive rate} = TP/(TP + FN). \quad \text{False positive rate} = FP/(FP + TN)$$

where TP is true positive, FN is false negative, FP is false positive, and TN is true negative.

Table 3. The methods used to predict CATH groups and the cross-validation results. Six attributes include precipitant, buffer, temperature, pH, structure molecular weight, and first-level CATH group ID.

Methods	Correctly classified instances, %	Incorrectly classified instances, %	Kappa statistic	ROC area
Ensemble method - Random Forest	56.2	43.8	0.1094	0.596
Decision tree - J48	57.3	42.7	0.0284	0.561
Lazy - IBK (KNN)	46.3	53.7	0.0771	0.542
Bayes - Naïve Bayes	55.7	44.3	0.0018	0.536
Bayes - Bayes NET	57.2	42.8	0.0541	0.585

Table 3 shows the prediction of first-level CATH groups by crystallization conditions using various methods. Using five different methods, the average correctly classified instance rate is around 54.5%, with low kappa statistics and ROC areas. The increase in the number of correctly classified instances from second-level to first-level CATH groups is mainly because the decrease of group numbers from forty to four. This indicates that CATH groups, of either first-level or second-level, are difficult to predict successfully based on crystallization conditions. This also implies that the overall shape of the protein structure and the arrangement of secondary structure might not be a main factor that affects crystallization conditions.

Resolution

Structure resolution is a measurement indicating the quality of data obtained from X-ray crystallography. The higher the resolution is, indicated by low Å values, the better the crystal

data will be. When the crystals are highly ordered in an identical way, the X-ray diffraction pattern of all the proteins in the crystal will be the same, yielding a high-resolution structure around 1 Å with the electron density map showing each atom clearly. When the proteins are in slightly different locations in crystals because of local flexibility or movement, the diffraction pattern will not be so detailed, so that only the contours of protein chains will be detected and the atomic structure will be inferred.

The protein resolutions of structures determined by X-ray crystallography are shown in Figure 4. Structures near a resolution of 2.0 Å have the largest population, while the resolution range is mainly from 0.75 Å to 3.5 Å. Structures of resolution less than 1.5 Å usually are clear enough to see the proteins at atomic level, and the structures have almost no error from the electron density maps. Such structures are around 15% of the PDB. Proteins with resolution greater than 3.0 Å, which are about 5% of the PDB, can have correctly determined secondary structure elements, although their side-chain structure may have many errors. Structures between these two extremes are the main part of the PDB, with their structures near 2.0 Å resolution being most commonly found.

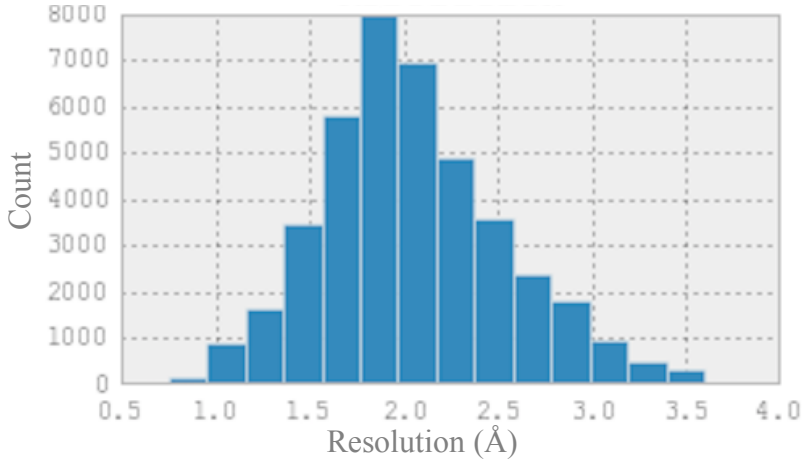


Figure 6. Resolution distribution and the counts of protein crystals in the PDB.

pH values

The pH value during crystallization is a constantly reported variable. The effect of pH on crystallization is likely to be the local charge distribution (Rupp 2010). Figure 7 shows the

histogram of pH values in the PDB protein crystals, which ranges from around pH 3 to pH 11. Interestingly, the most prevalent pH ranges are 6–6.5 and 7.5–8, where the range in between these two bins (around 7) is less populated. Diagrams of PH values with other variables are shown in Figure 8.

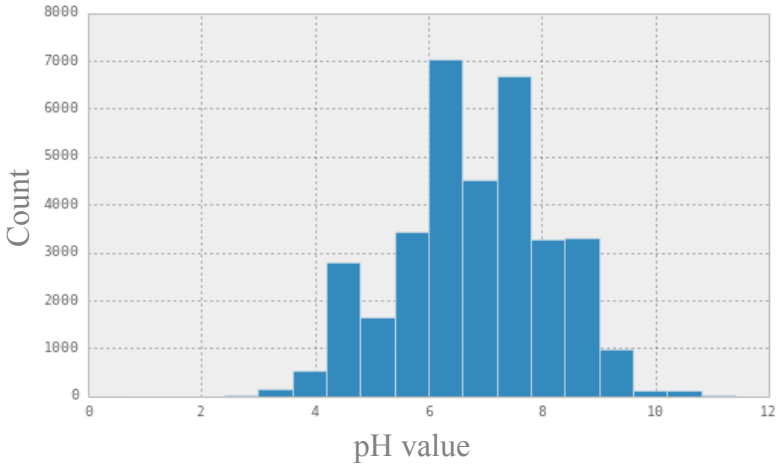


Figure 7. pH value distribution and the counts of protein crystals in the PDB.

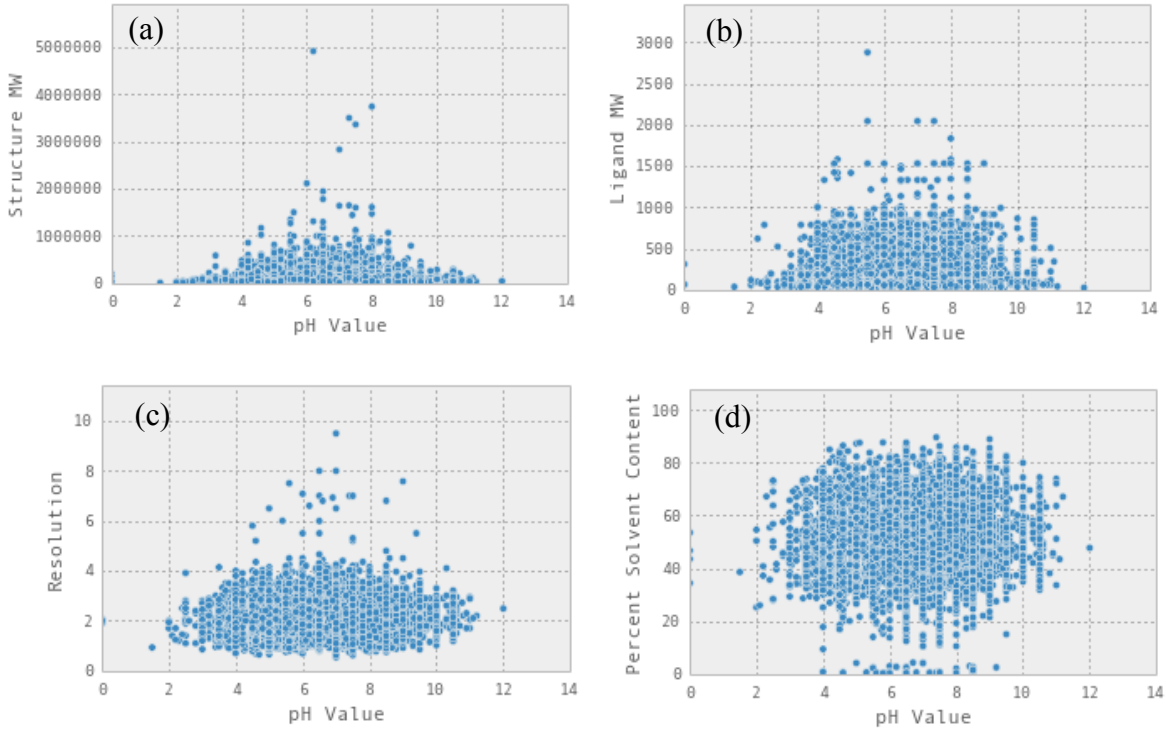


Figure 8. Scatter plots of pH values and (a) structure molecular weight (Da) (b) ligand molecular weight (Da) (c) resolution (Å) (d) percent solvent content (%).

Temperature

Temperature during crystallization is another constantly reported variable. Figure 9 is the histogram of temperature of protein crystals in the PDB from 250K to 340K. This excludes temperature outliers that are lower than 100 K (less than 1% of the total counts). These outliers may be caused by manual confusion of Kelvin with degrees Celsius and degrees Fahrenheit. Temperatures around 275 K, 290 K, and 300 K are most popular, mainly because experiments are conducted in these cold-room or room temperature. Figure 10 contains the diagrams of temperature versus other variables.

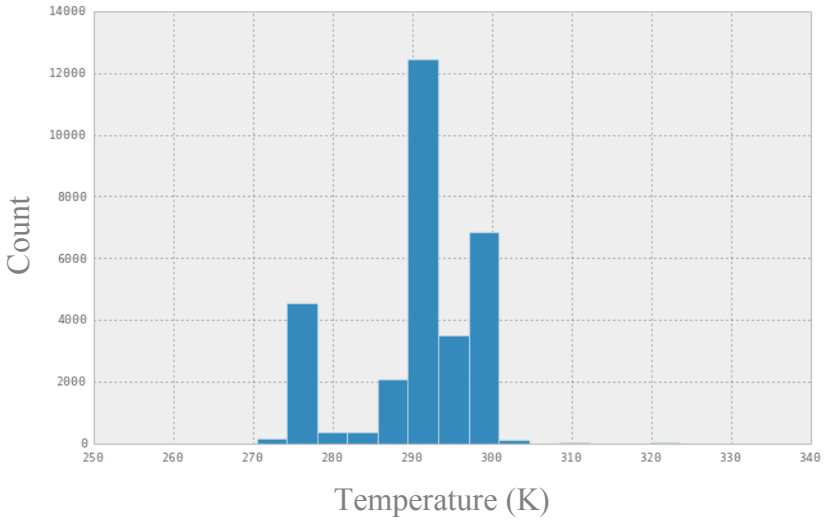


Figure 9. Temperature (K) distribution and the counts of protein crystals in the PDB.

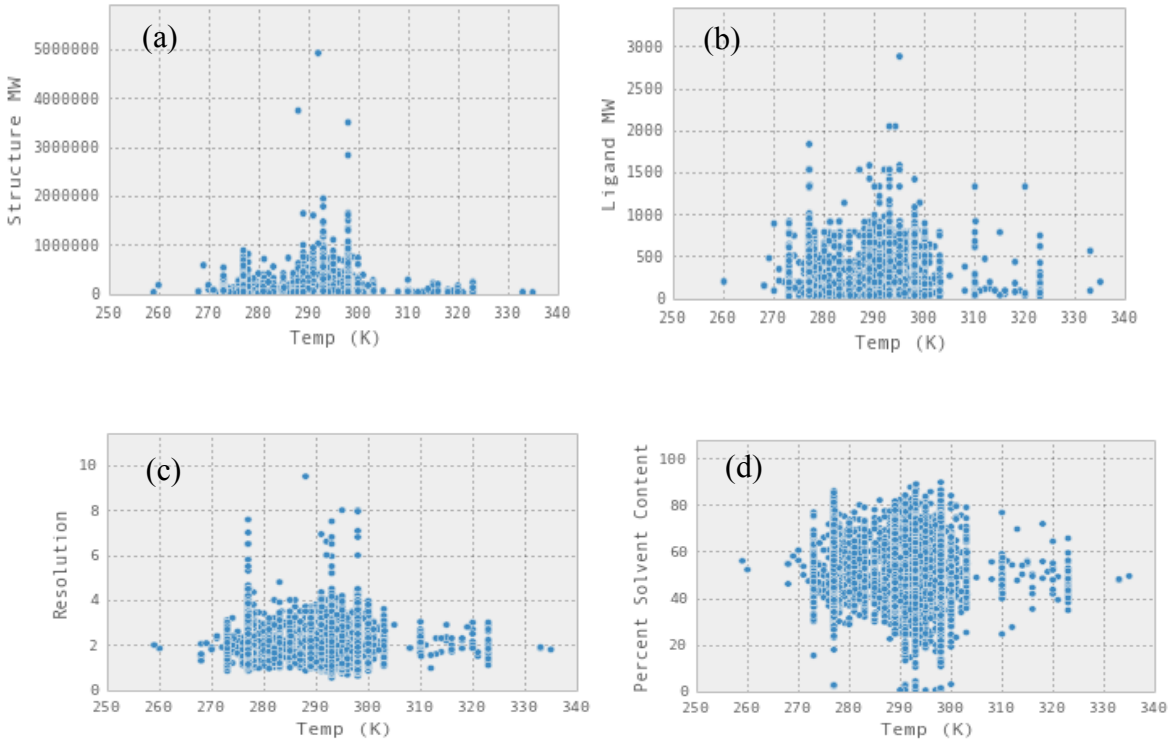


Figure 10. Scatter plots of temperature (K) and (a) structure molecular weight (Da); (b) ligand molecular weight (Da); (c) resolution (Å); (d) percent solvent content (%)**.**

Percent solvent content

The solvent content is the crystal volume occupied by solvent. A general trend was found that crystals with less solvent content generally have higher-resolution structures, or smaller resolution values (Figure 11).

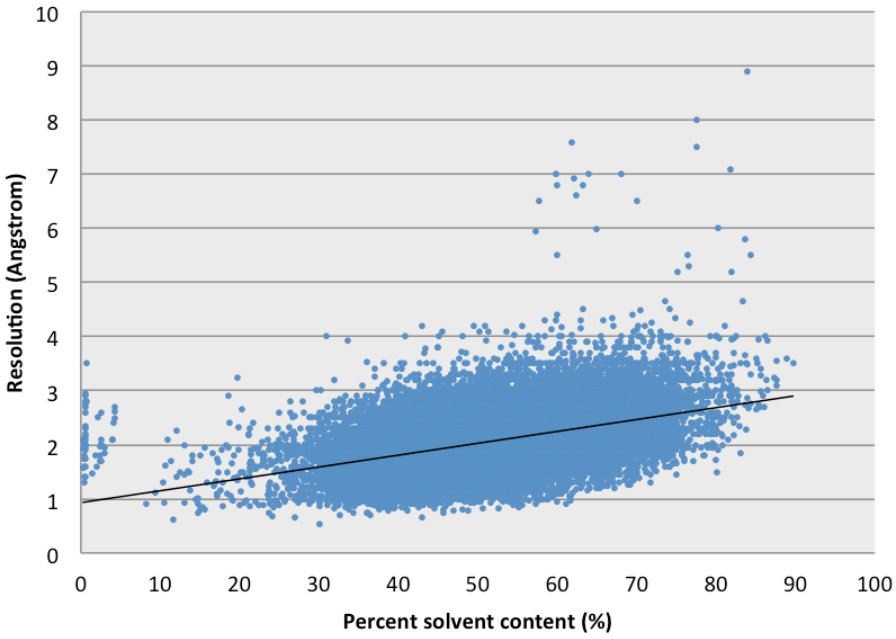


Figure 11. The relationship of percent solvent content and resolution. The linear regression yields $y = (0.0219 \pm 0.0002)\,x + (0.931 \pm 0.012)$, where the ranges are the standard errors, y is the resolution (Å), and x is the % solvent content. The 95% confidence level intervals of the slope and intercept are 0.0214 to 0.0224 and 0.906 to 0.955, respectively.

Besides some outliers of less than 10% solvent content that have lower resolution than expected and a few outliers with exceptionally low resolution (>5 Å), the general trend is that the higher the percent solvent content, the lower the resolution.

Crystallization of glycoproteins

Glycoproteins are of interest because their glycans can be on the protein surface, which may affect crystallization. The numbers of crystal conditions, including precipitants, buffers, salts, and additives, necessary to crystallize glycoprotein crystal structures in the PDB is compared to those necessary to crystallize other protein crystal structures in the PDB in Figure 12. The total number may indicate the difficulty of protein crystallization, as more reagents do not need to be added if fewer reagents already give crystals of high quality. As shown in Figure 6, glycoproteins need about the same number of conditions as other proteins. Three conditions are most commonly required for successful crystallization, followed by two

and four conditions. This probably indicates that the difficulty of crystallizing glycoproteins is the same as that of crystallizing proteins in general.

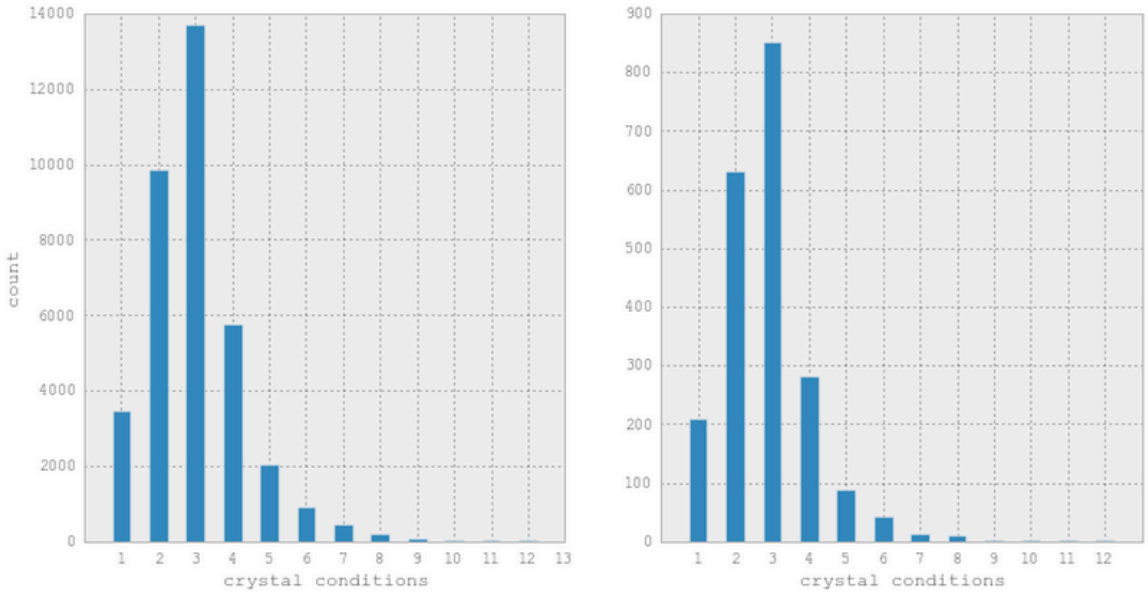


Figure 12. The number of crystallization conditions and successful PDB counts of all other proteins (left) and glycoproteins (right) determined by X-ray crystallography.

The commonly used precipitants are summarized by the order of their popularity in Table 4 of glycoproteins and other proteins. In general, the top ten precipitant types are similar, where pentaerythritol propoxylate is used in glycoproteins but not in the other proteins. PEG, the most popular precipitant, appears in 80.6% of glycoprotein crystallization cocktails, whereas the number is 74.7% in other proteins.

Table 4. The most popular precipitants used in crystalizing glycoprotein (left) and other protein (right), with their percentage of the total precipitant counts.

Glycoprotein		**Other proteins**	
Precipitant type	%	Precipitant type	%
PEG	80.6	PEG	74.7
Ammonium sulfate	11.4	Ammonium sulfate	13.5
MPD	1.9	MPD	3.4
Jeffamine	0.8	Sodium chloride	3.4
Sodium chloride	2.8	Lithium sulfate	1.5
Lithium sulfate	1.0	Sodium phosphate	1.2
Propanol	0.4	Propanol	0.9
Sodium phosphate	0.4	Jeffamine	0.4
Pentaerythritol propoxylate	0.2	Tacsimate	0.3
Tacsimate	0.2	Hexanediol	0.2

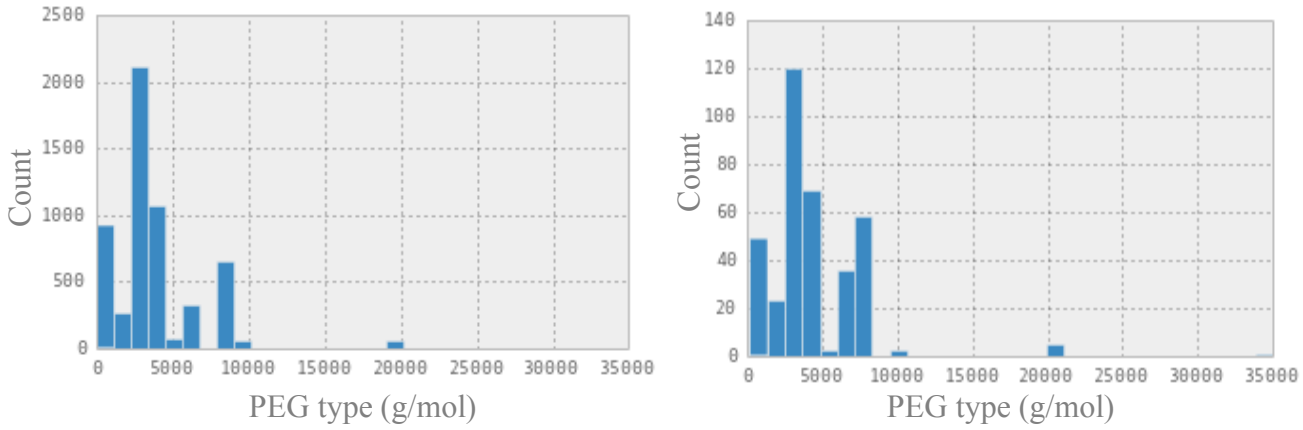


Figure 13. The different types of PEG used for crystallization, indicated by their molecular weights (g/mol) for all other proteins (left) and for glycoproteins (right).

As the most-used precipitant, PEGs of various molecular weights are summarized in Figure 13 of glycoproteins and for other proteins. The frequencies of PEG types, ranging from 200 to 35,000 g/mol, used in crystallization conditions are almost the same for the two groups. From the diagram, PEG 3000 to PEG 4000 are most prevalent, followed by PEG 400 and PEG 8000.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am Statis*, *46*(3), 175–185.

Apweiler, R., Hermjakob, H., & Sharon, N. (1999). On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta*, *1473*(1), 4–8.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res*, 28(1), 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., et al. (1977). The Protein Data Bank. *FEBS J*, *80*(2), 319–324.

Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J. C., Frutiger, S. and Hochstrasser, D. (1993). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, 14, 1023–1031.

Carter, C. W., & Carter, C. W. (1979). Protein crystallization using incomplete factorial experiments. *J Biol Chem*, *254*(23), 12219–12223.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach Learn*, *20*(3), 273–297.

Dale, G. E., Oefner, C., & D'Arcy, A. (2003). The protein as a variable in protein crystallization. *J Struct Biol,* 142(1), 88–97.

Drenth, J. (2007). *Principles of protein X-ray crystallography.* New York: Springer, 2007.

Eswar, N., Webb, B., Marti Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., et al. (2006). Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics, 5.6. 1–5.6. 30.

Farr, R. G., Jr, Perryman, A. L., & Samudzi, C. T. (1998). Re-clustering the database for crystallization of macromolecules. *J Cryst Growth*, *183*(4), 653–668.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edition (pp. 38–46). Wiley, New York.

Giegé, R. (2013). A historical perspective on protein crystallization from 1840 to the present day. *FEBS J*, *280*(24), 6456–6497.

Gilliland, G. L., Tung, M., & Ladner, J. (1996). The Biological Macromolecule Crystallization Database and NASA Protein Crystal Growth Archive. *J Res Natl Inst Stand Technol*, *101*(3), 309–320.

Hagan, M. T., Demuth, H. B., & Beale, M. H. (1996). *Neural Network Design* (Vol. 1). PWS Publishing, Boston.

Jancarik and Kim (1991). Sparse matrix sampling: a screening method for crystallization of proteins. *J Appl Cryst*, 24, 409–411.

Jones, D. T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, 23(5), 538–544.

Lloyd, S. (1982). Least squares quantization in PCM. Information theory. *IEEE Trans*, *28*(2), 129–137.

Luft, J. R., Snell, E. H., & DeTitta, G. T. (2011). Lessons from high-throughput protein crystallization screening: 10 years of practical experience. *Expert Opin Drug Discov*, *6*(5), 465–480.

Krishnaraj, Y., & Reddy, C. K. (2008). Boosting methods for protein fold recognition: an empirical comparison. In Bioinformatics and Biomedicine, 2008. *BIBM'08. IEEE International Conference on* (pp. 393-396). IEEE.

Meining, W. (2006). XtalBase–A comprehensive data management system for macromolecular crystallography. *J Appl Crystallogr*, *39*(5), 759–766.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, *247*(4), 536–540.

Nettleship, J. E. (2012). *Structural Biology of Glycoproteins*. INTECH Open Access Publisher, Rijeka, Croatia.

Newman, J., Peat, T. S., & Savage, G. P. (2014). What's in a name? Moving towards a limited vocabulary for macromolecular crystallisation. *Aust J Chem*, *67*(12), 1813–1817.

Newstead, S., Ferrandon, S., & Iwata, S. (2009). Rationalizing α-helical membrane protein crystallization. *Prot Sci*, *17*(3), 466–472.

Page, R., Grzechnik, S. K., Canaves, J. M., Spraggon, G., Kreusch, A., Kuhn, P., et al. (2003). Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the Thermotoga maritima proteome. *Acta Crystallogr D Biol Crystallogr*, 59(6), 1028–1037.

Parker, J. L., & Newstead, S. (2012). Current trends in α-helical membrane protein crystallization: An update. *Prot Sci*, *21*(9), 1358–1365.

Peat, T. S., Christopher, J. A., & Newman, J. (2005). Tapping the Protein Data Bank for crystallization information. *Acta Crystallogr Sect D: Biol Crystallogr*, *61*(12), 1662–1669.

Religa, T. L., Sprangers, R., & Kay, L. E. (2010). Dynamic regulation of archaeal proteasome gate opening as studied by TROSY NMR. *Science*, 328(5974), 98–102.

Rupp, B. (2010). *Biomolecular Crystallography*. Garland Science. New York.

Samudzi, C. T., Fivash, M. J., & Rosenberg, J. M. (1992). Cluster analysis of the biological macromolecule crystallization database. *J Cryst Growth*, *123*(1), 47–58.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *Comp J*, *16*(1), 30–34.

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., et al. (2012). New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res*, *41*(D1), D490–D498.

Tung, M., & Gallagher, D. T. (2009). The Biomolecular Crystallization Database, Version 4: Expanded content and new features. *Acta Cryst*. D65, 18–23.

Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., & Sharon, N. (2009). *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Webb, E. C. (1992). *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes (No. Ed. 6).* Academic Press, San Diego.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. Burlington, MA.

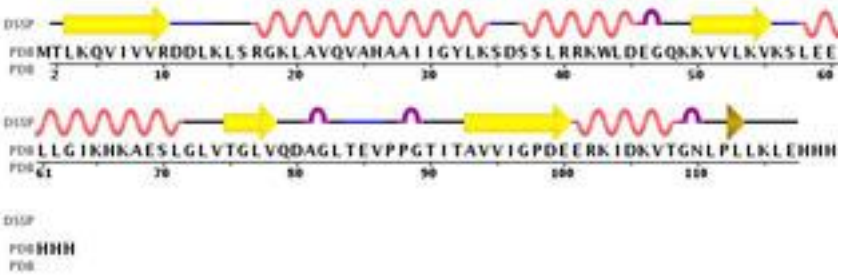
Wooh, J. W., Kidd, R. D., Martin, J. L., & Kobe, B. (2003). Comparison of three commercial sparse-matrix crystallization screens. *Acta Crystallogr Sect D, Biol Crystallogr*, *59*(4), 769–772.

Zhu, D. W., Garneau, A., Mazumdar, M., Zhou, M., Xu, G. J., & Lin, S. X. (2006). Attempts to rationalize protein crystallization using relative crystallizability. *J Struct Biol*, *154*(3), 297–302.

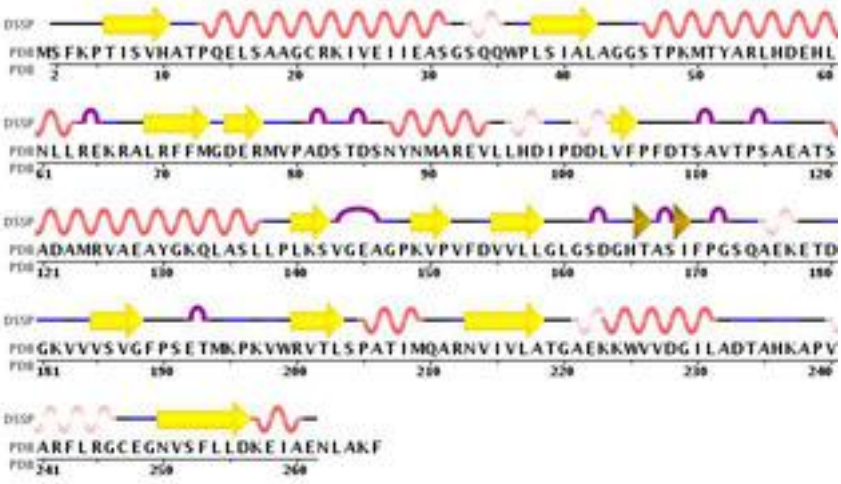
Zucker, F. H., Stewart, C., dela Rosa, J., Kim, J., Zhang, L., Xiao, L., et al. (2010). Prediction of protein crystallization outcome using a hybrid method. *J Struct Biol*, *171*(1), 64–73.

SUPPLEMENTAL MATERIALS. CEH SECONDARY STRUCTURE DIAGRAMS

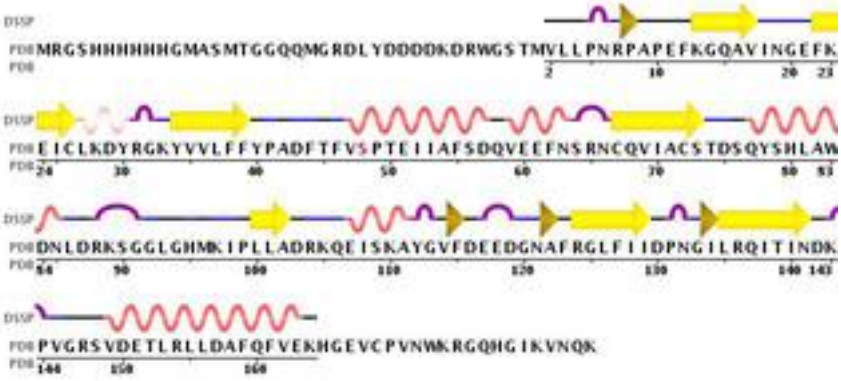
CEH1 3ERJ



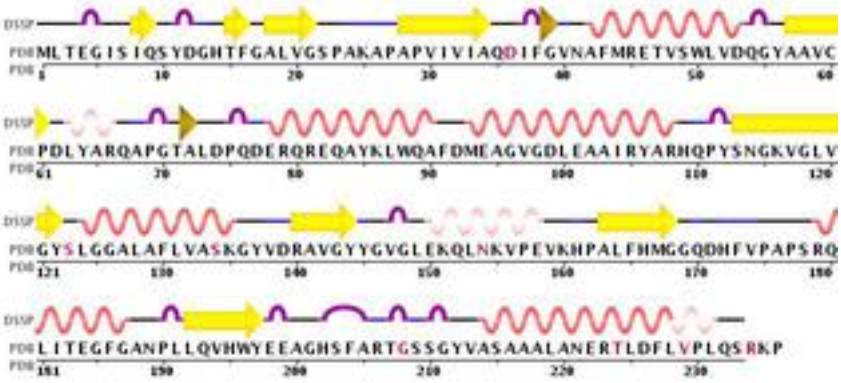
CEH2 3EB9



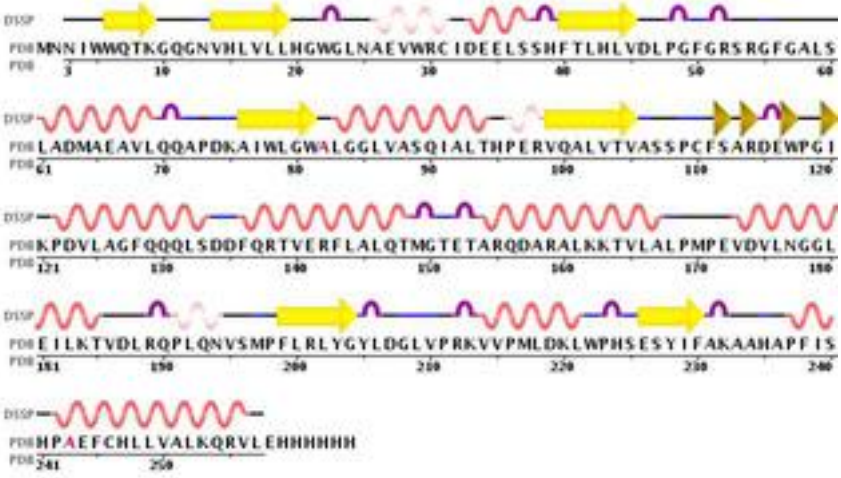
CEH3 3ZL5



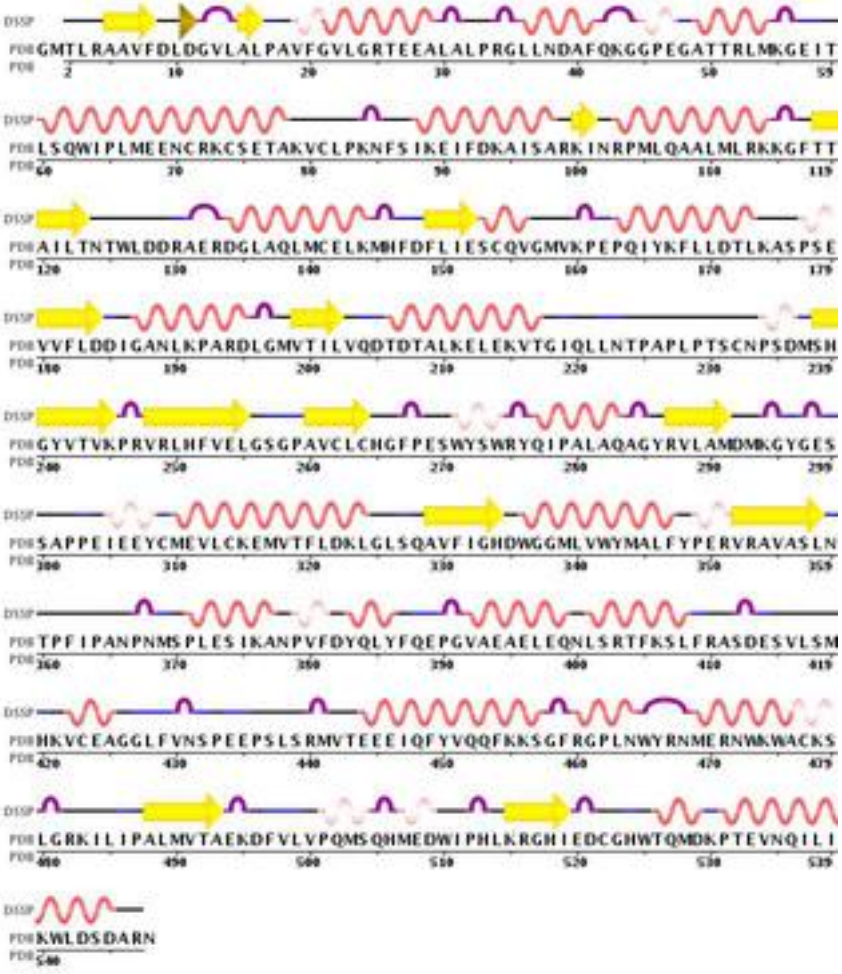
CEH4 1ZJ5



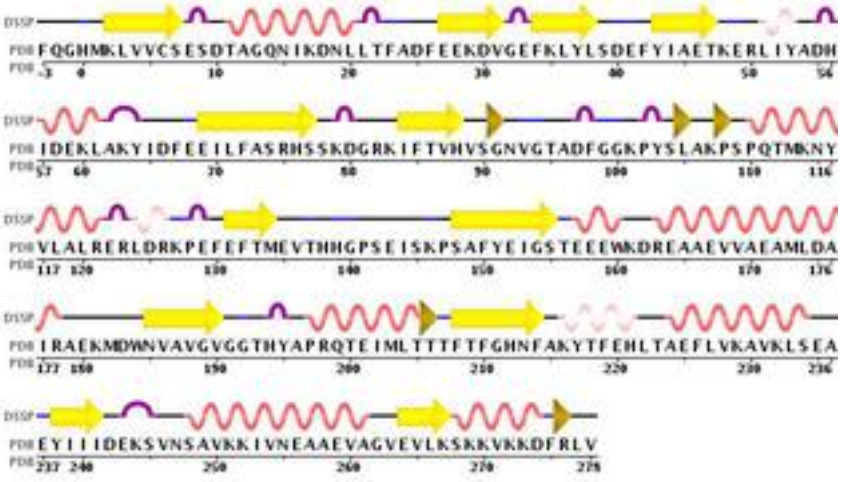
CEH6 4ETW



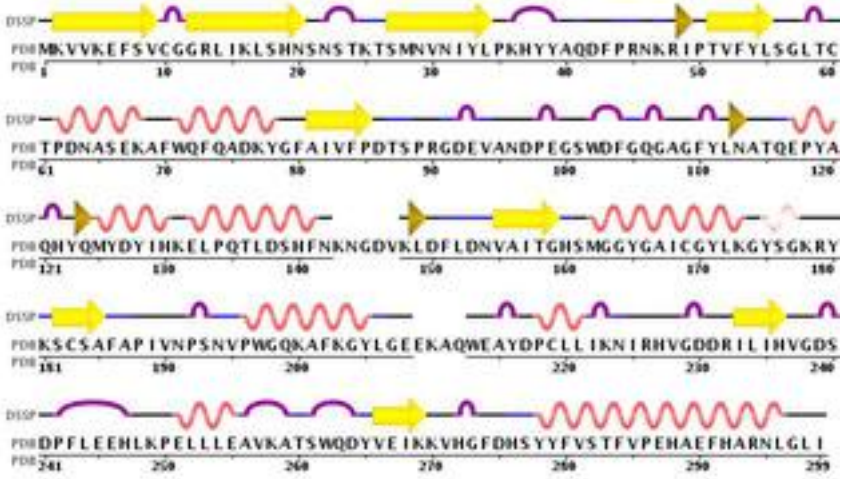
CEH7 5ALM



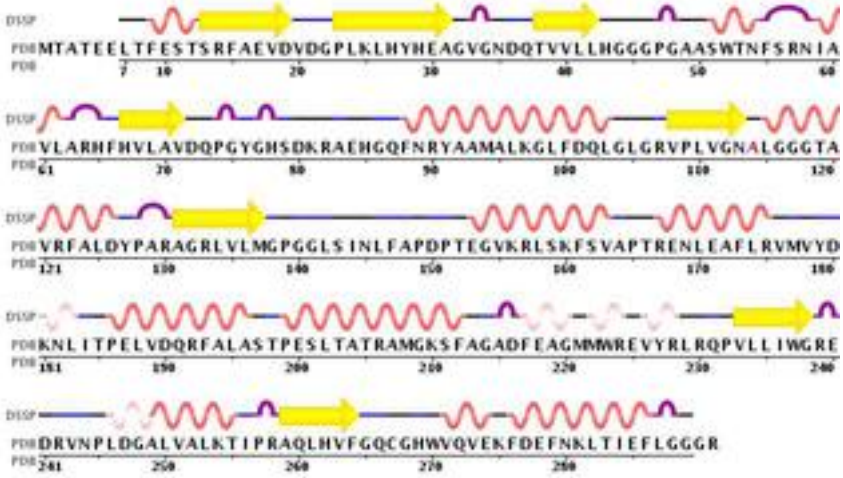
CEH8 1YQE



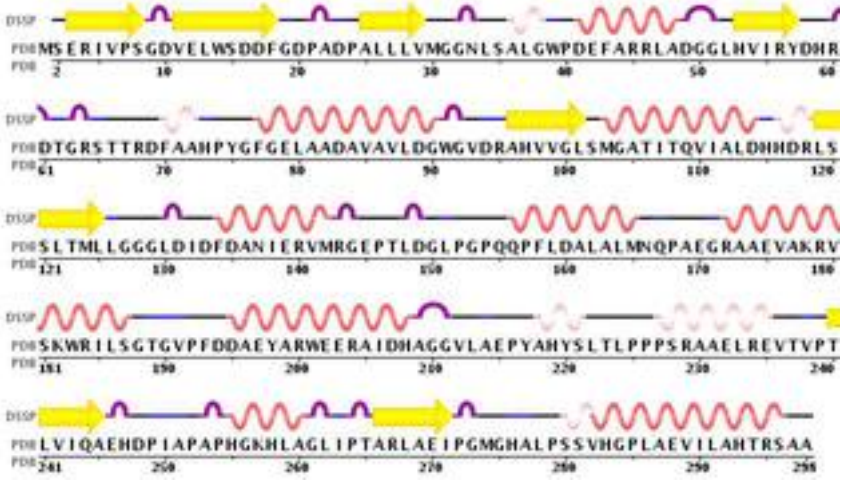
CEH9 1PV1



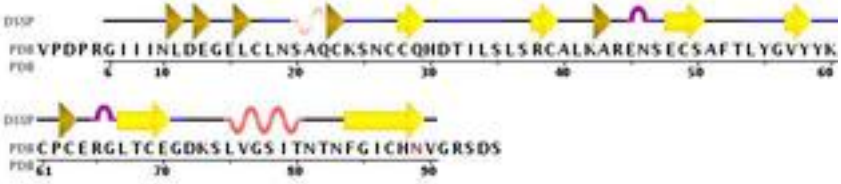
CEH10 2WUE



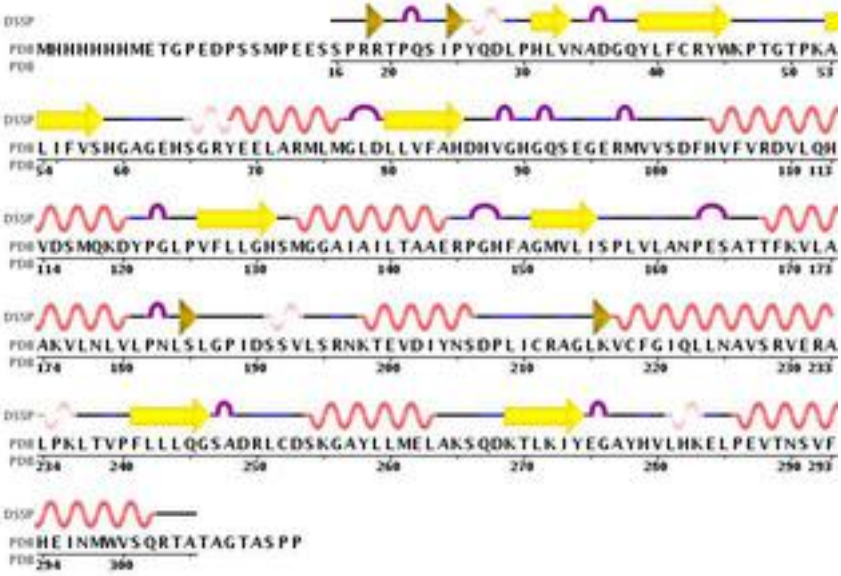
CEH11 1O0Z



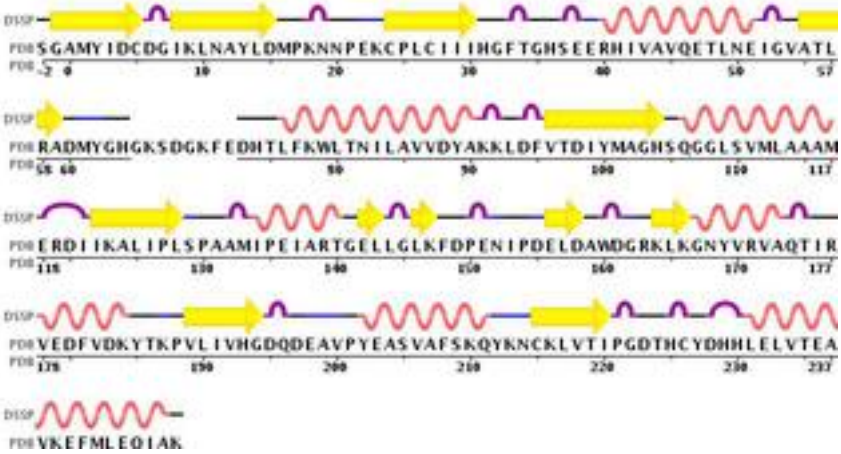
CEH12 1LPB



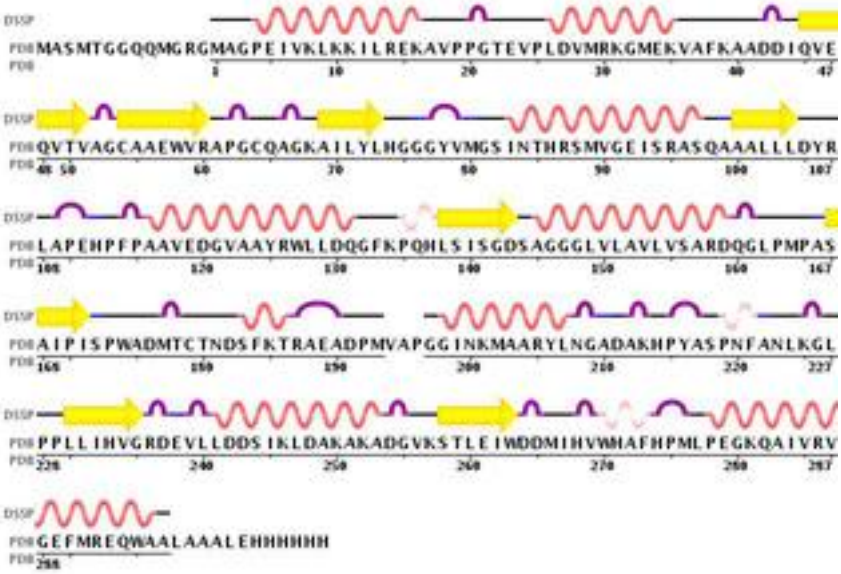
CEH13 4UUQ



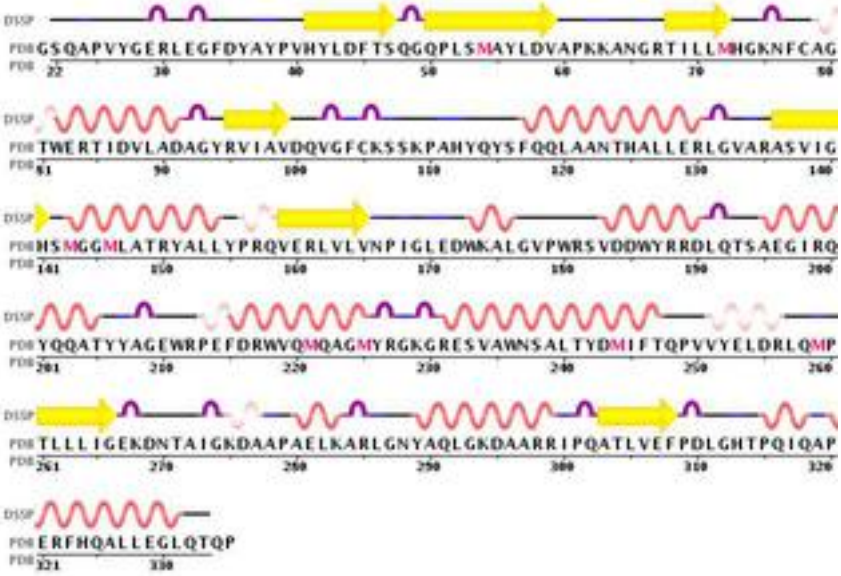
CEH14 2WTM



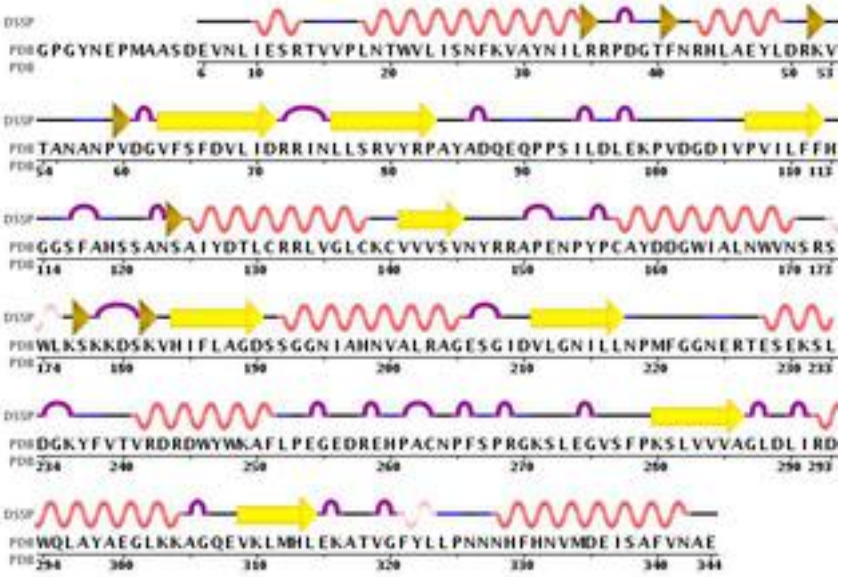
CEH153L1J



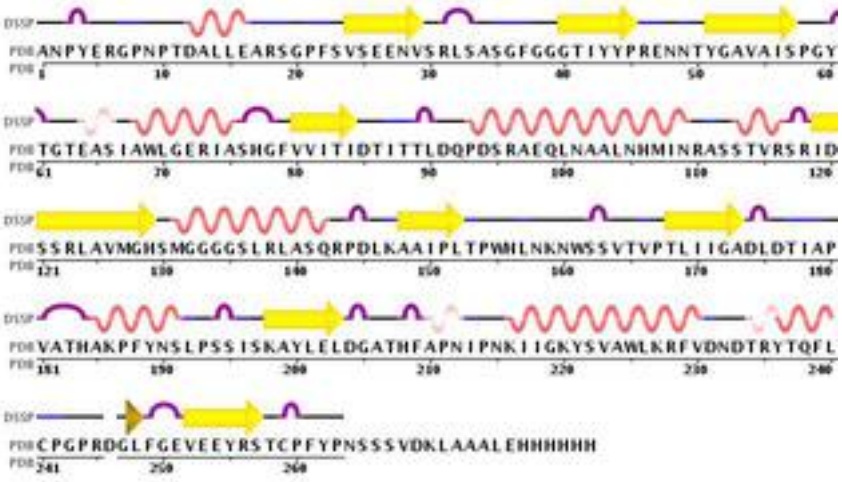
CEH164F0J



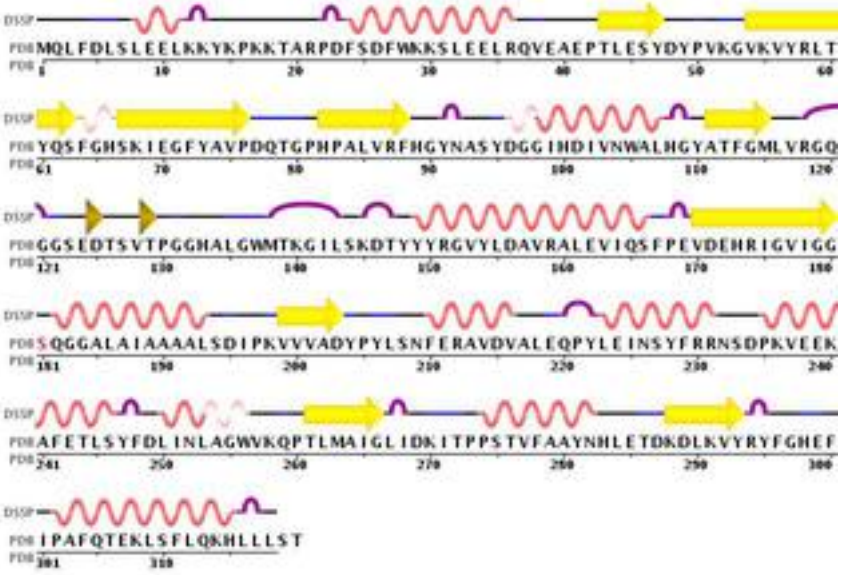
CEH17 2ZSI



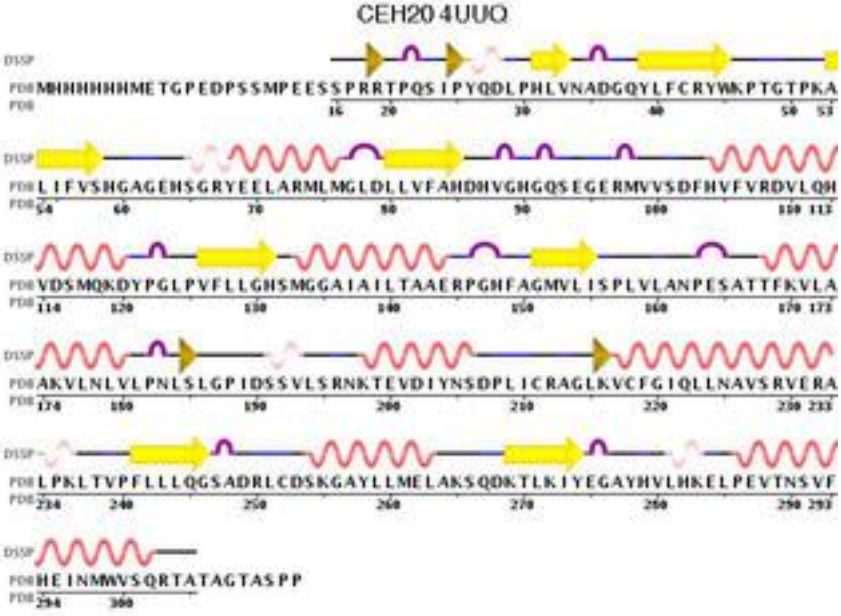
CEH18 4CG1



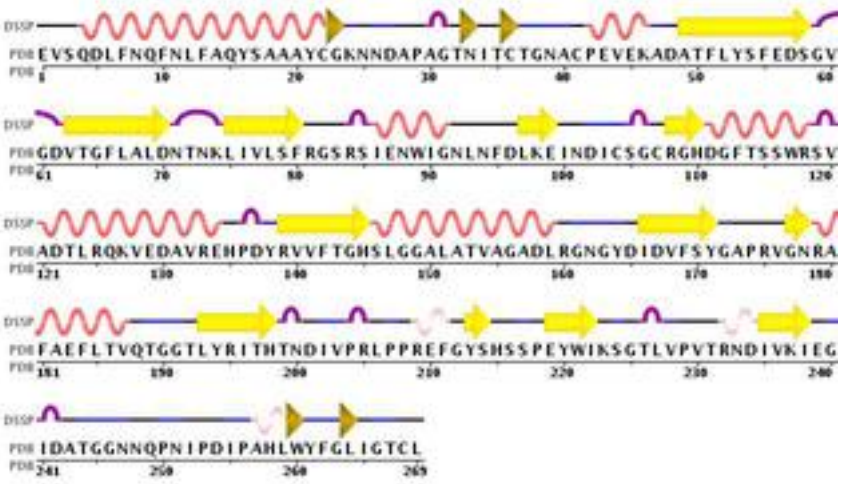
CEH193FYU



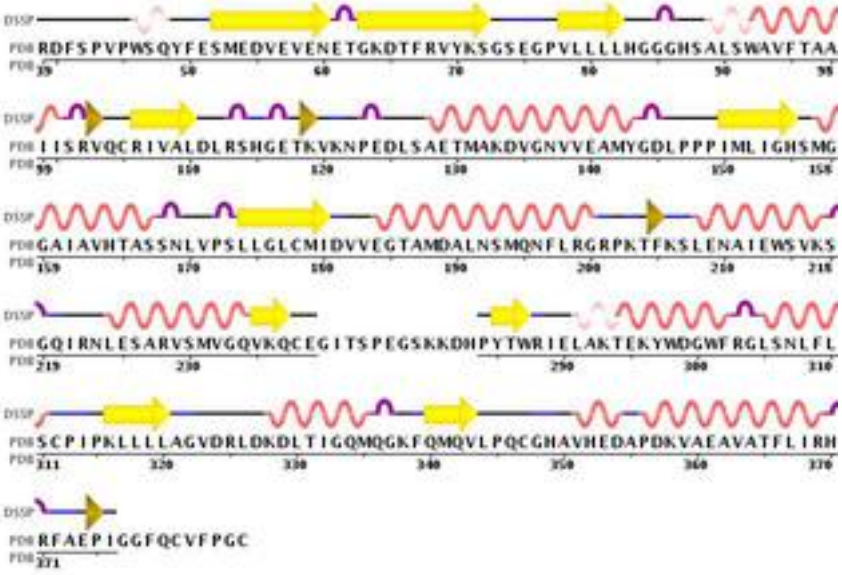
CEH204UUQ



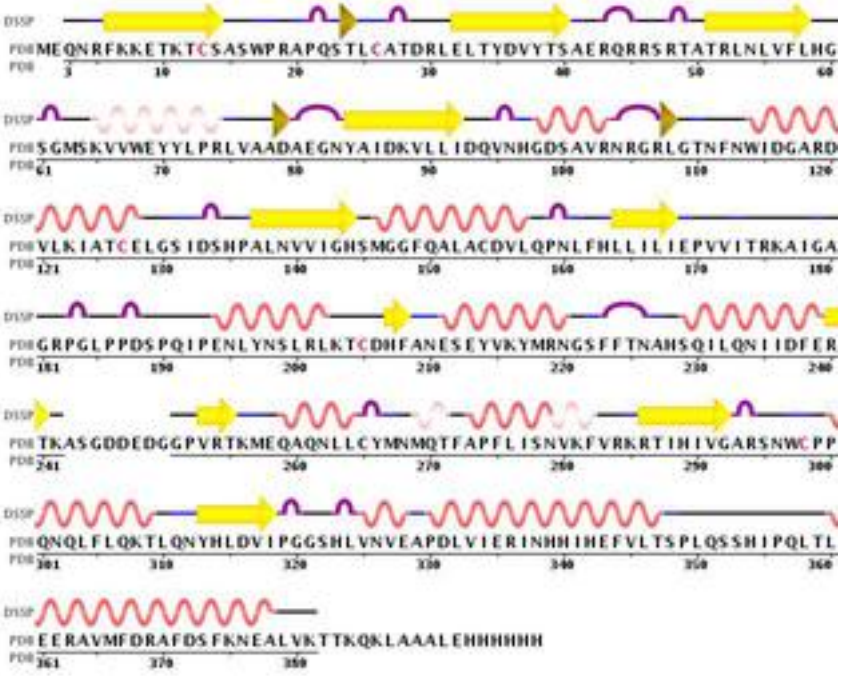
CEH21 4GBG



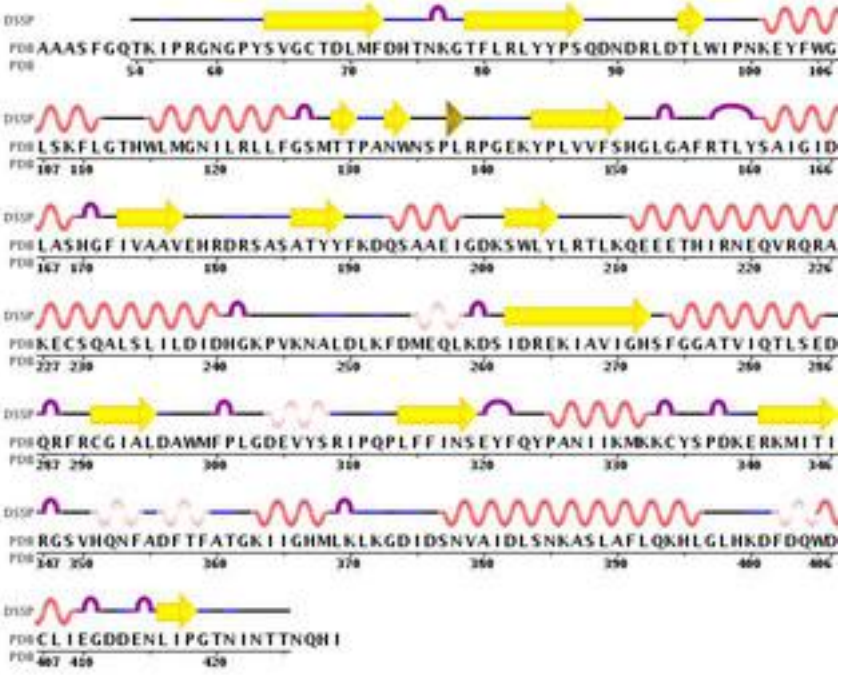
CEH22 3C5V



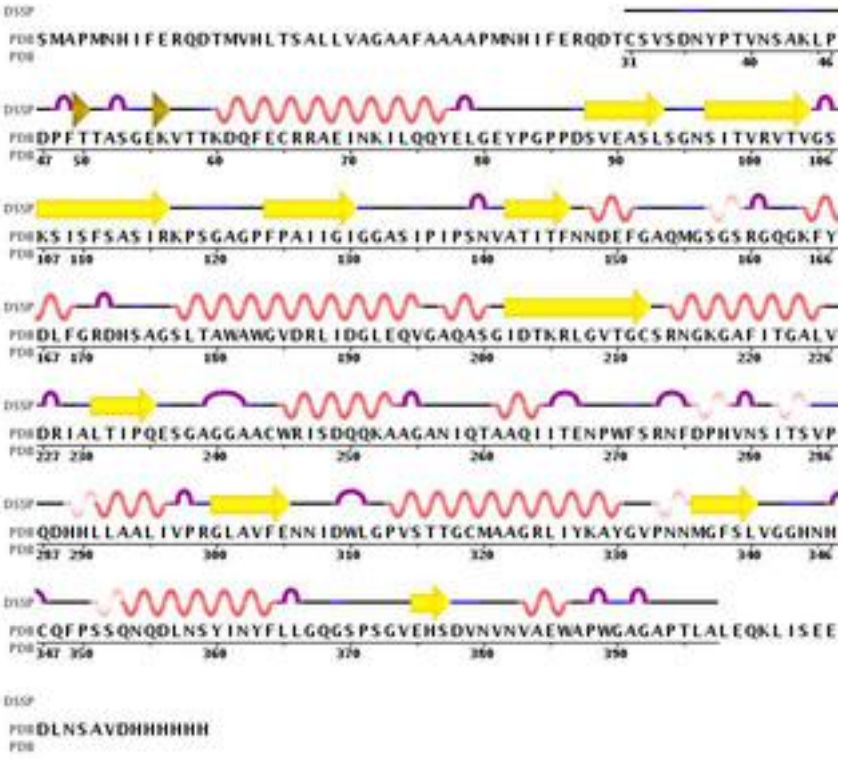
CEH23 2Y6U



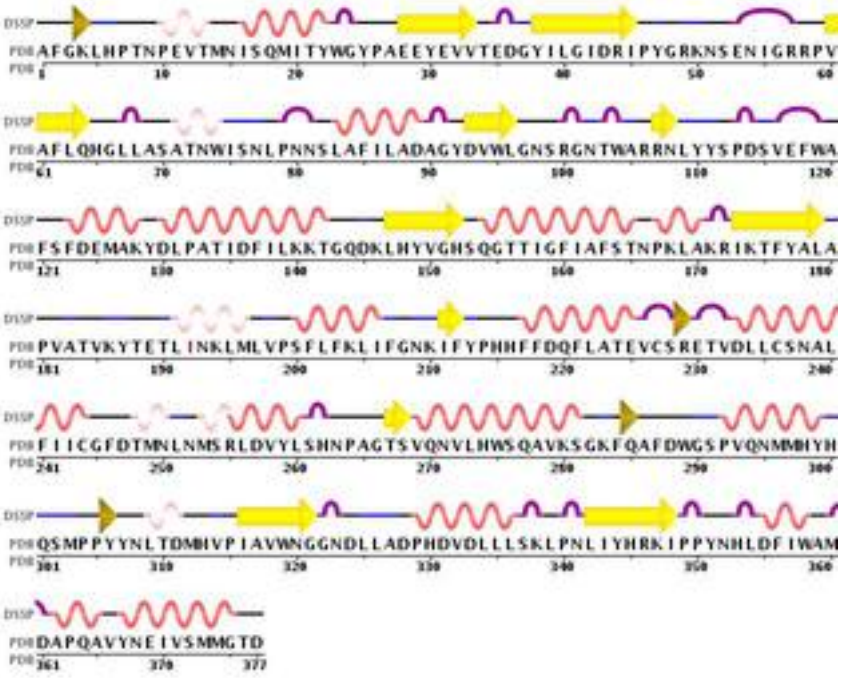
CEH24 3D59



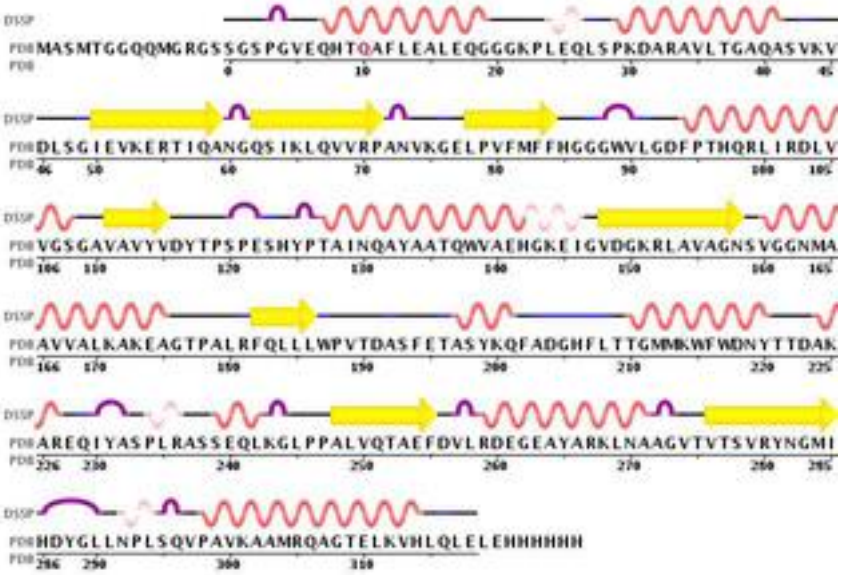
CEH25 4G4G



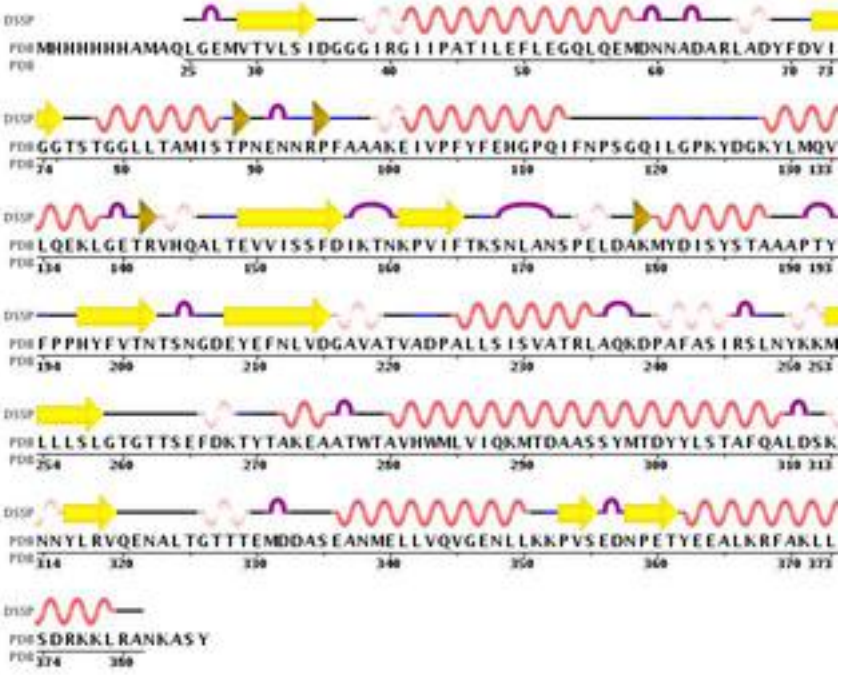
CEH26 1K8Q



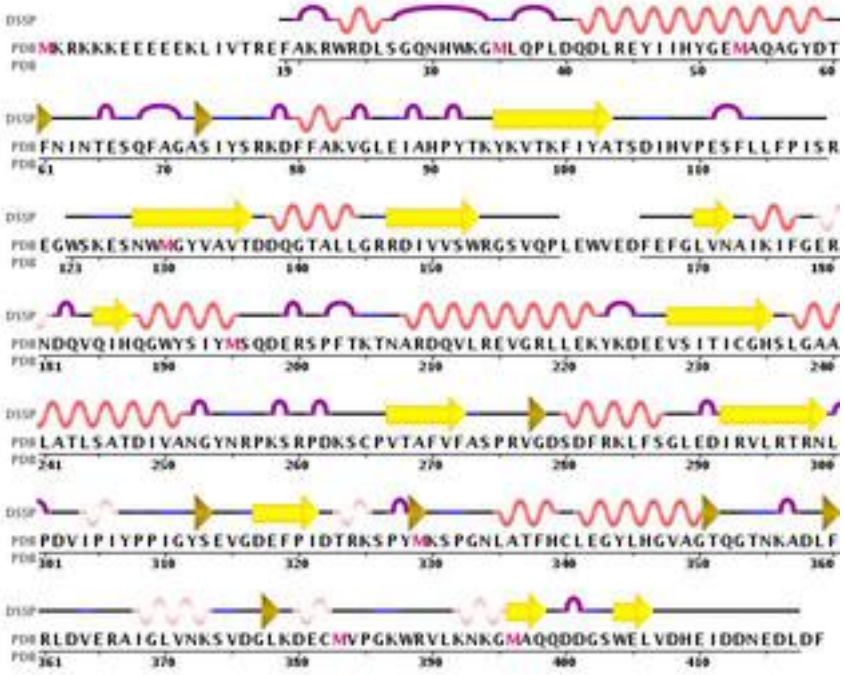
CEH27 4OB8



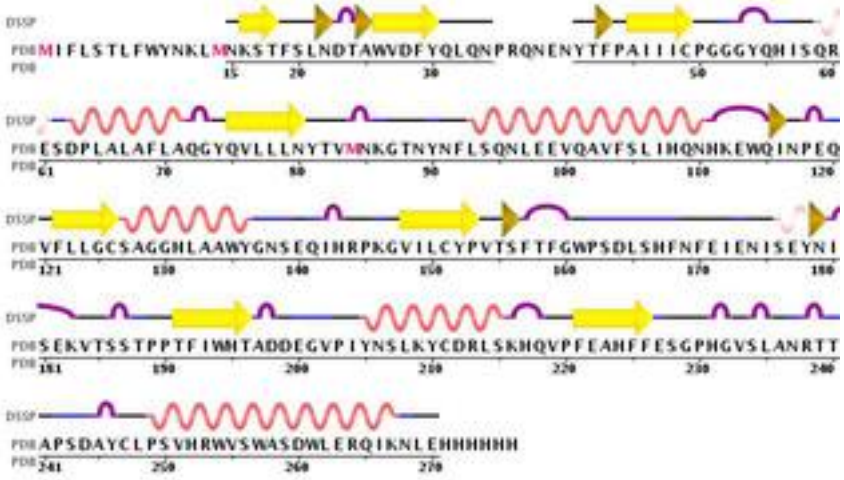
CEH28 4PKB



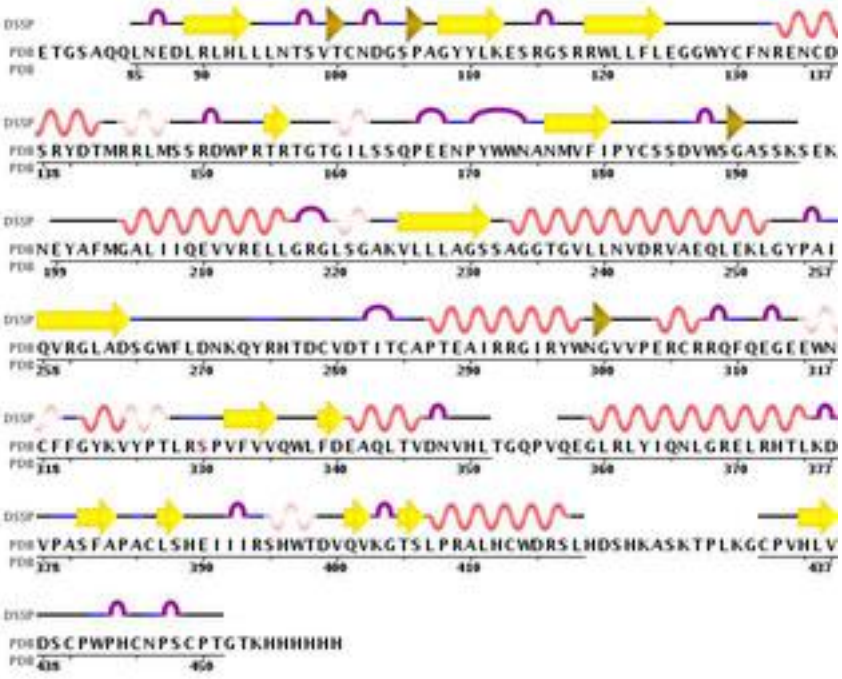
CEH29 2YLJ



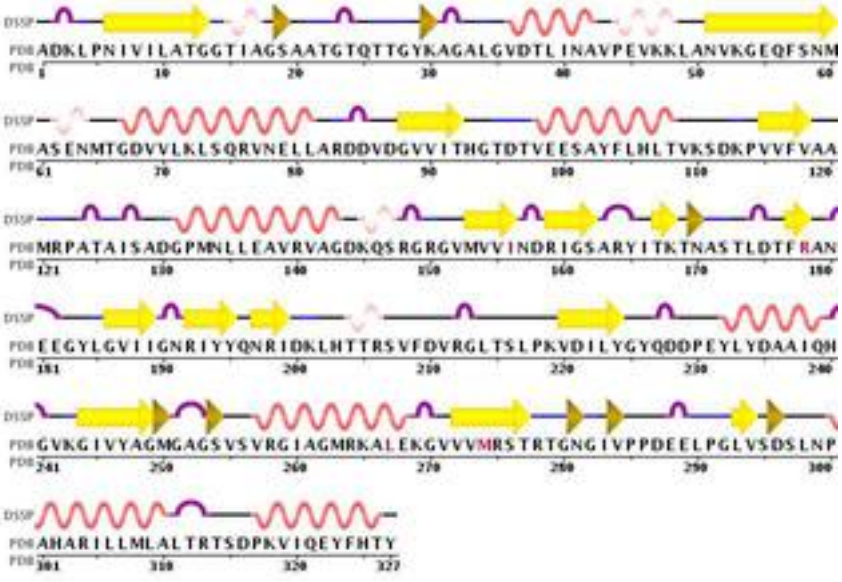
CEH30 3HXK

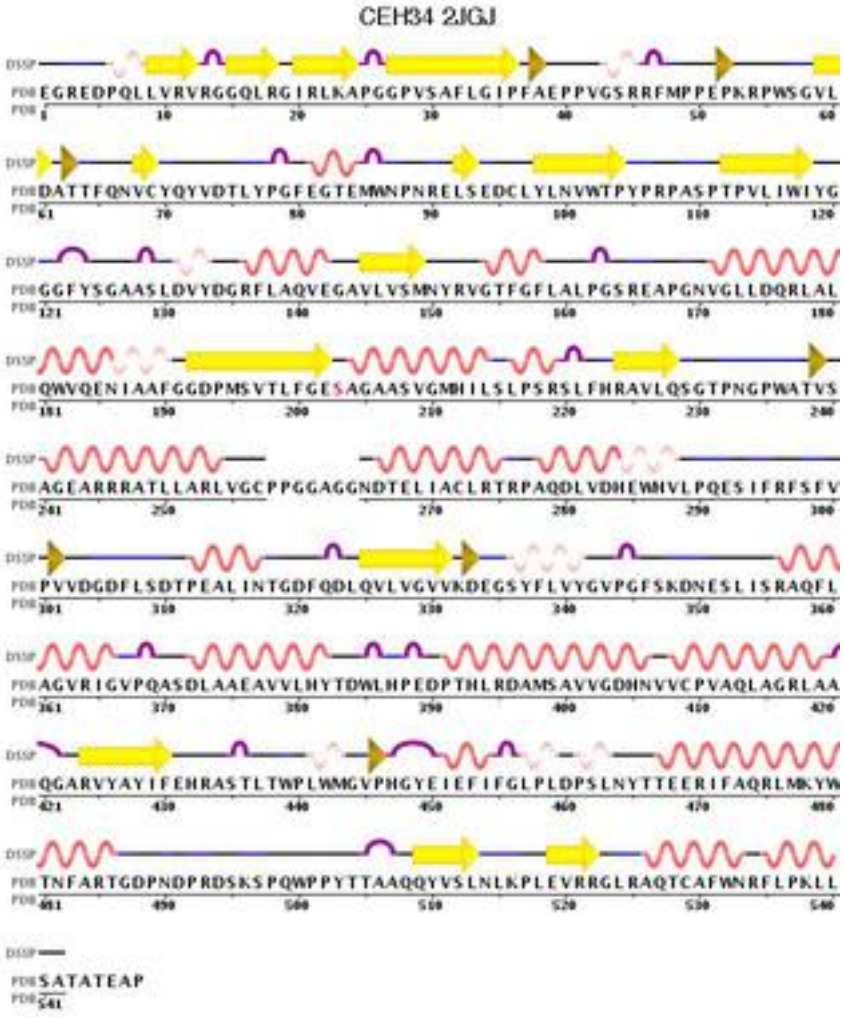


CEH32 4UYU

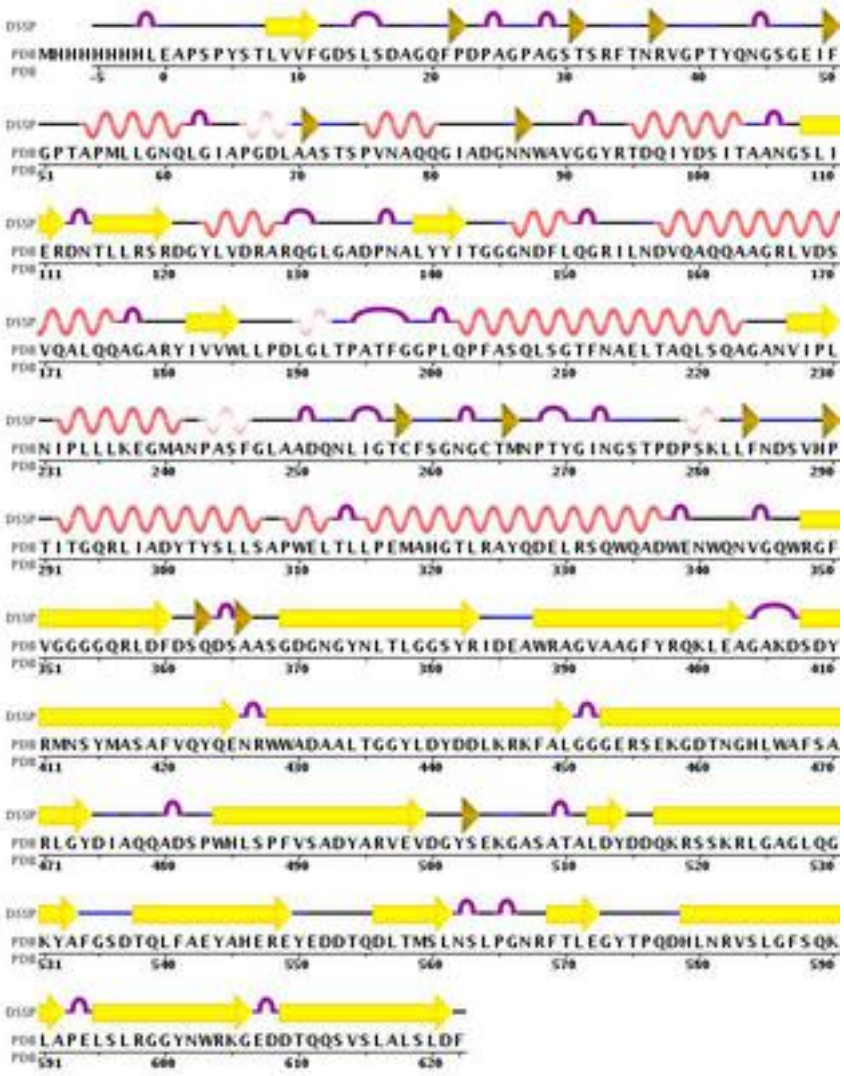


CEH33 1HG0

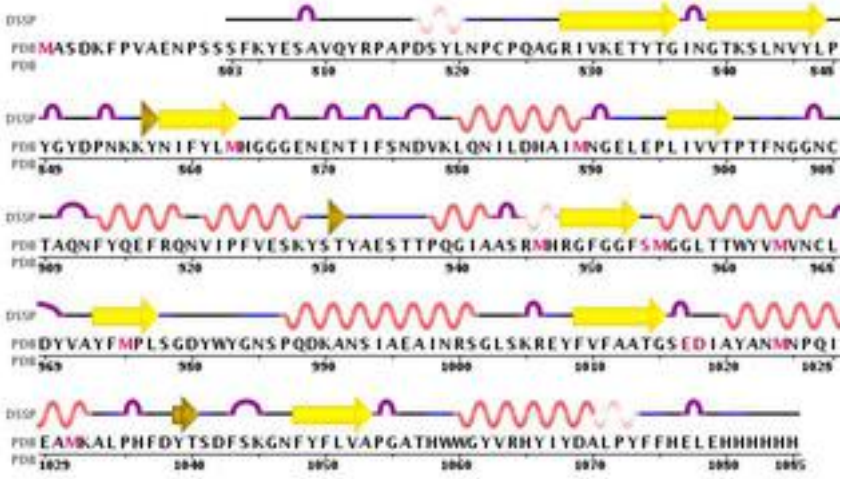



CEH34 2JGJ
DSSP
PDB EGREDPQLLVRVRGGQLRGIRLKAPGGPVSAFLGIPFAEPPVGSRRFMPPEPKRPWSGVL
PDB 1 10 20 30 40 50 60
DSSP
PDB DATTFQNVCYQYVDTLYPGFEGTEMWNPNRELSEDCLYLNVWTPYPRPASPTPVLIWIYG
PDB 61 70 80 90 100 110 120
DSSP
PDB GGFYSGAASLDVYDGRFLAQVEGAVLVSMNYRVGTFGFLALPGSREAPGNVGLLDQRLAL
PDB 121 130 140 150 160 170 180
DSSP
PDB QWVQENIAAFGGDPMSVTLFGESAGAASVGMHILSLPSRSLFHRAVLQSGTPNGPWATVS
PDB 181 190 200 210 220 230 240
DSSP
PDB AGEARRRATLLARLVGCPPGGAGGNDTELIACLRTRPAQDLVDHEWHVLPQESIFRFSFV
PDB 241 250 270 280 290 300
DSSP
PDB PVVDGDFLSDTPEALINTGDFQDLQVLVGVVKDEGSYFLVYGVPGFSKDNESLISRAQFL
PDB 301 310 320 330 340 350 360
DSSP
PDB AGVRIGVPQASDLAAEAVVLHYTDWLHPEDPTHLRDAMSAVVGDHNVVCPVAQLAGRLAA
PDB 361 370 380 390 400 410 420
DSSP
PDB QGARVYAYIFEHRASTLTWPLWMGVPHGYEIEFIFGLPLDPSLNYTTEERIFAQRLMKYW
PDB 421 430 440 450 460 470 480
DSSP
PDB TNFARTGDPNDPRDSKSPQWPPYTTAAQQYVSLNLKPLEVRRGLRAQTCAFWNRFLPKLL
PDB 481 490 500 510 520 530 540
DSSP
PDB SATATEAP
PDB 541

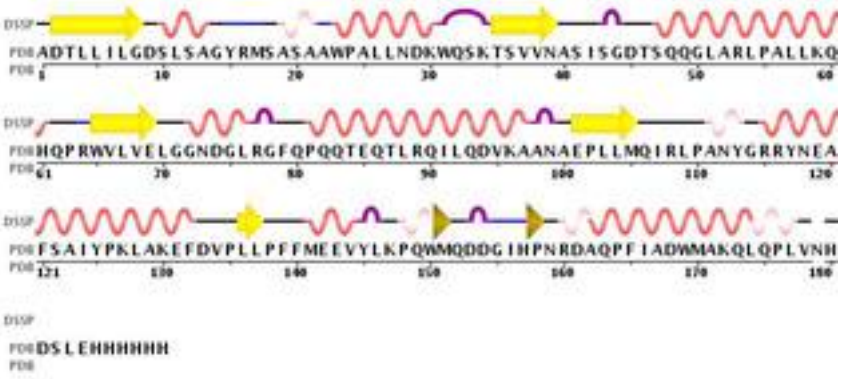
CEH36 3KVN



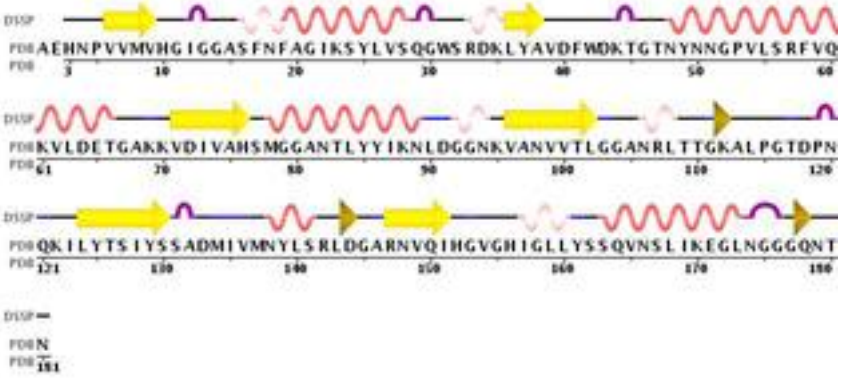
CEH37 3ZI7



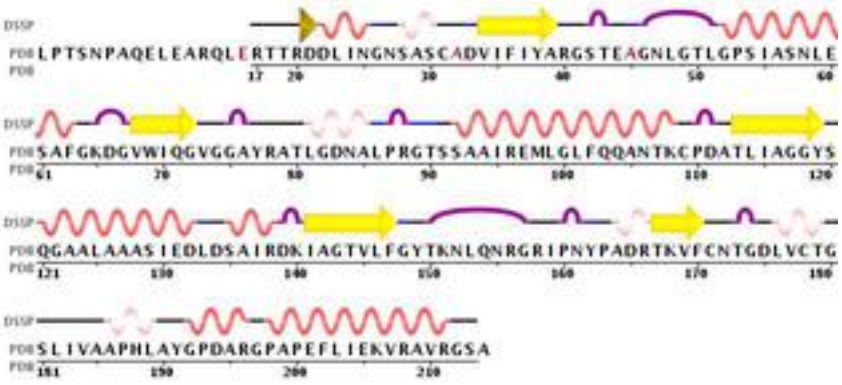
CEH39 1U8U



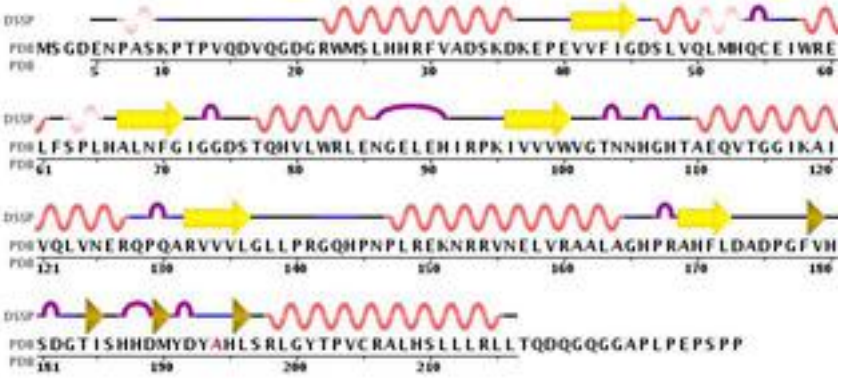
CEH40 1R50



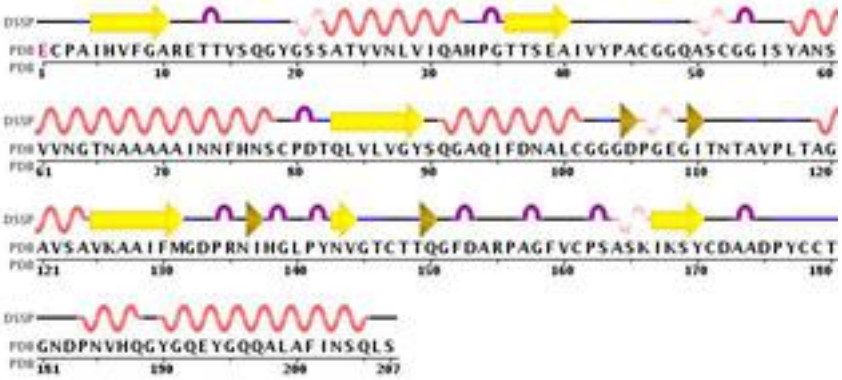
CEH41 1XZG



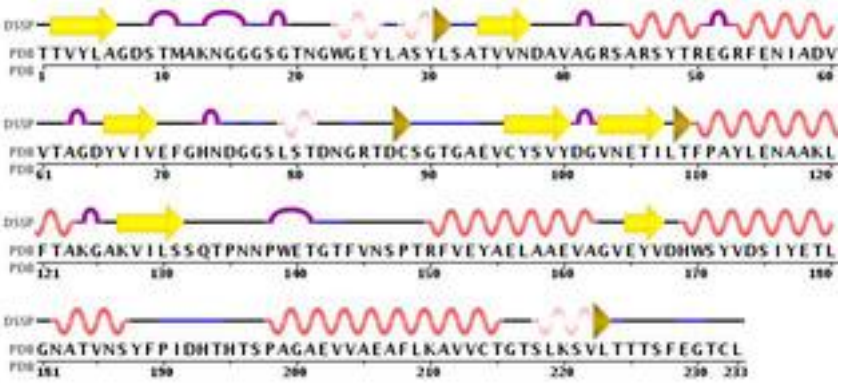
CEH42 1BWQ



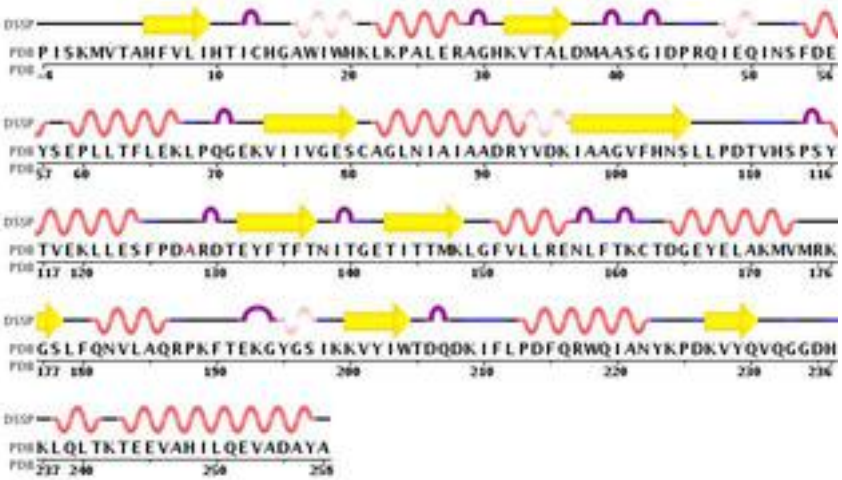
CEH43 1OOZ



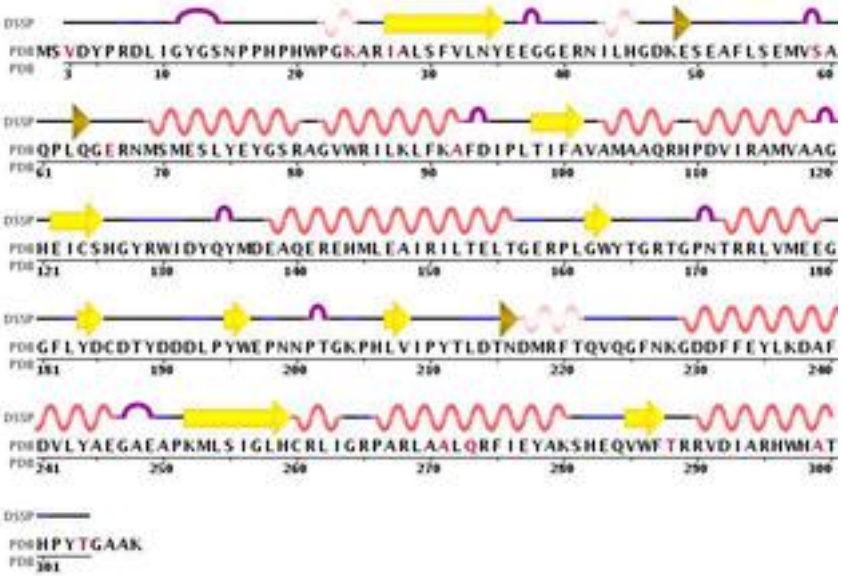
CEH44 1DEO



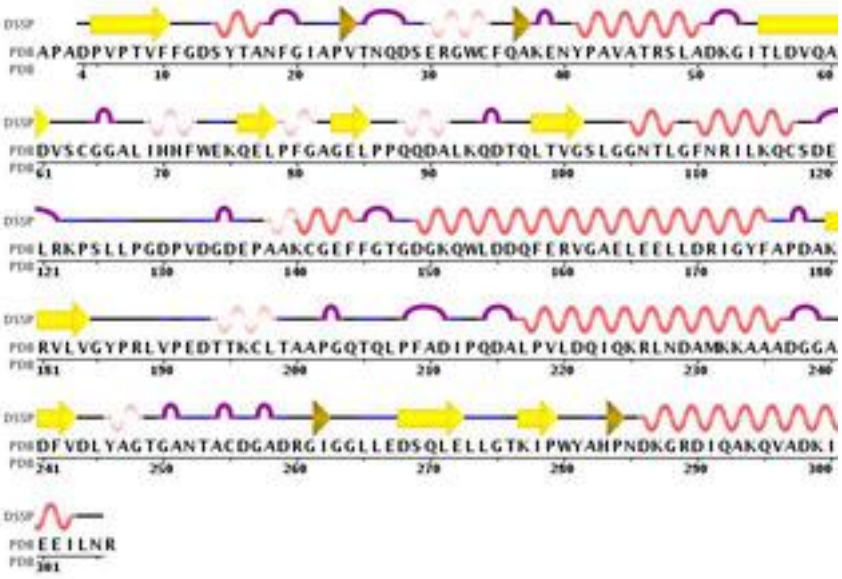
CEH45 1EB8



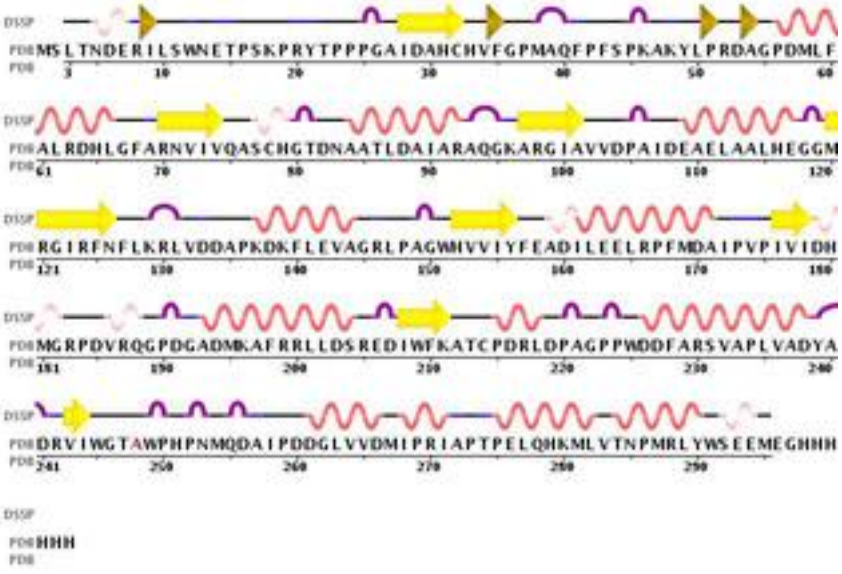
CEH46 3CL6



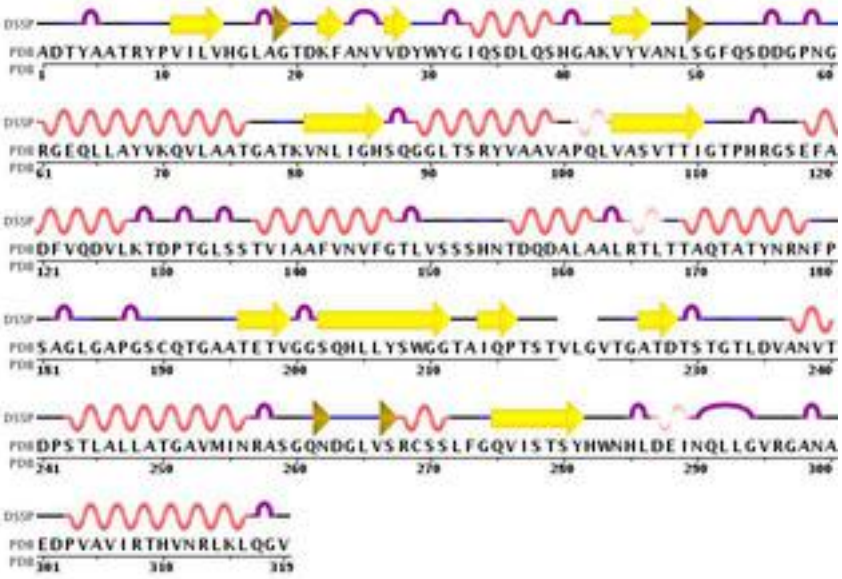
CEH47 1ESD



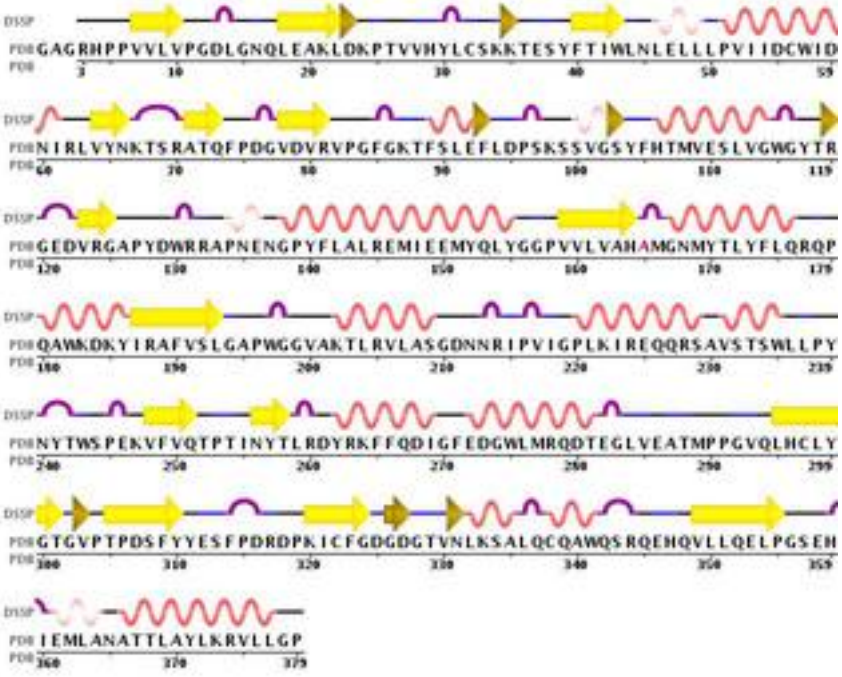
CEH48 4DI9



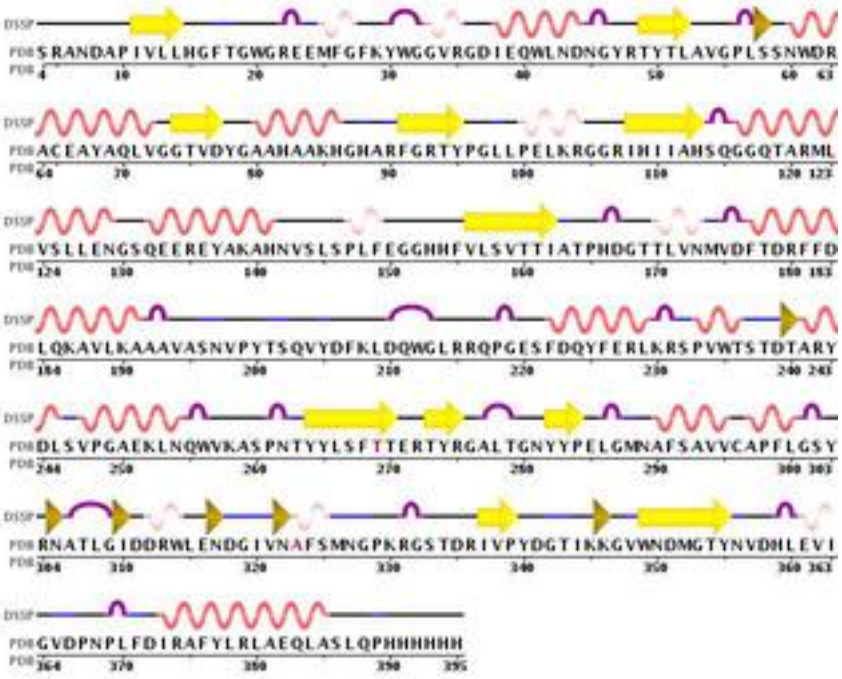
CEH49 1CVL



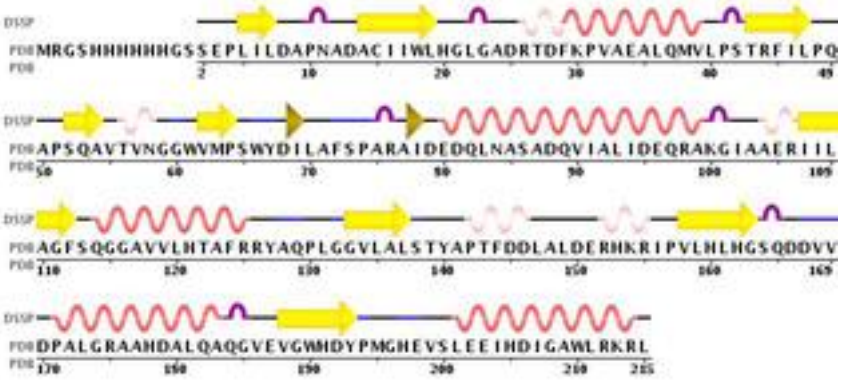
CEH50 4X92



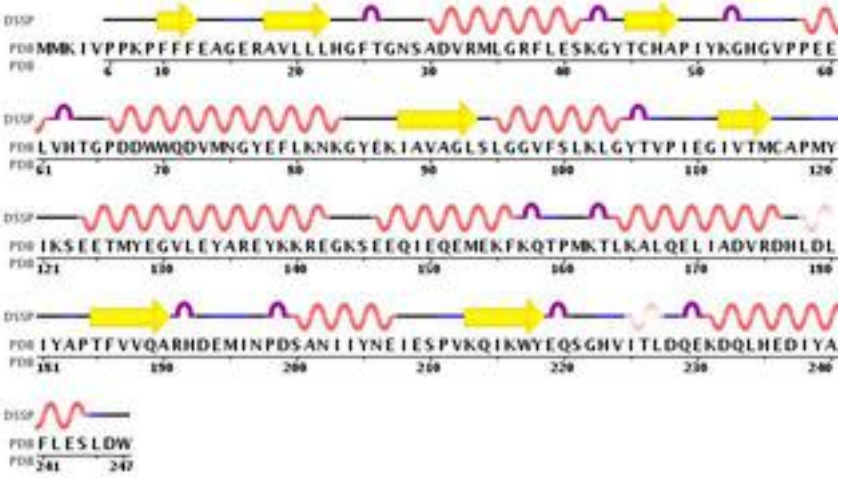
CEH51 4X71



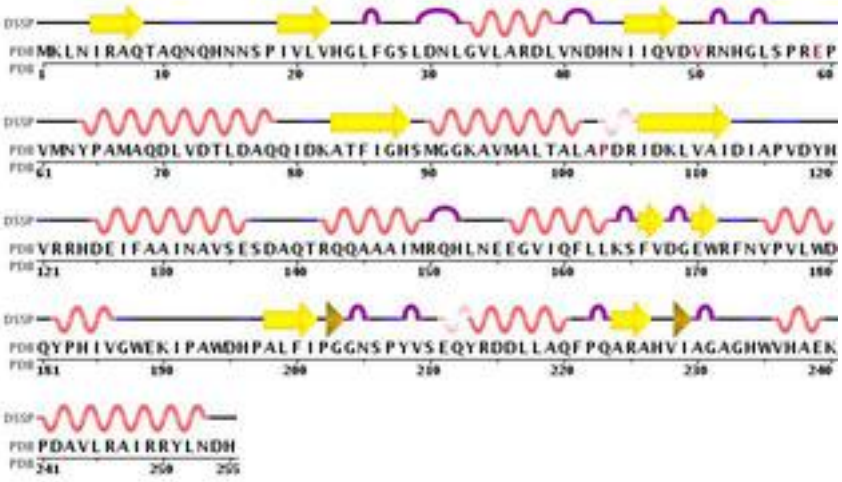
CEH52 3CN9



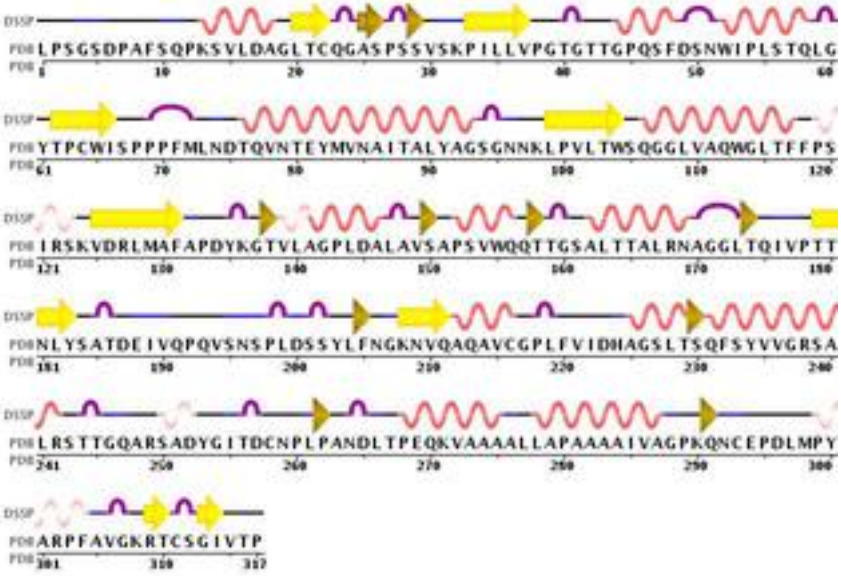
CEH53 1TQH



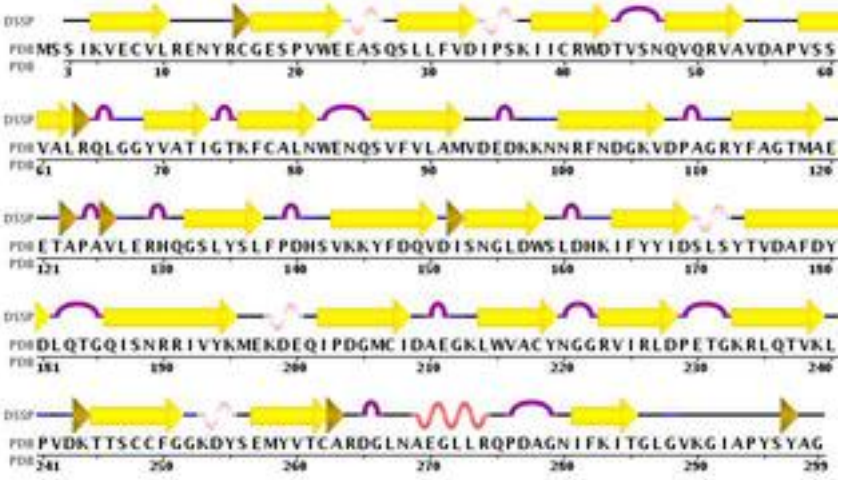
CEH54 3BF8



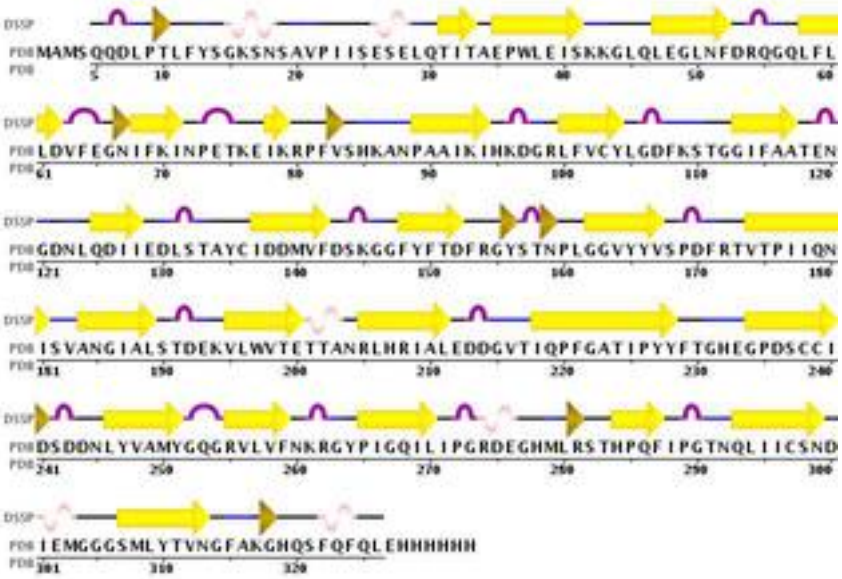
CEH55 1LBS



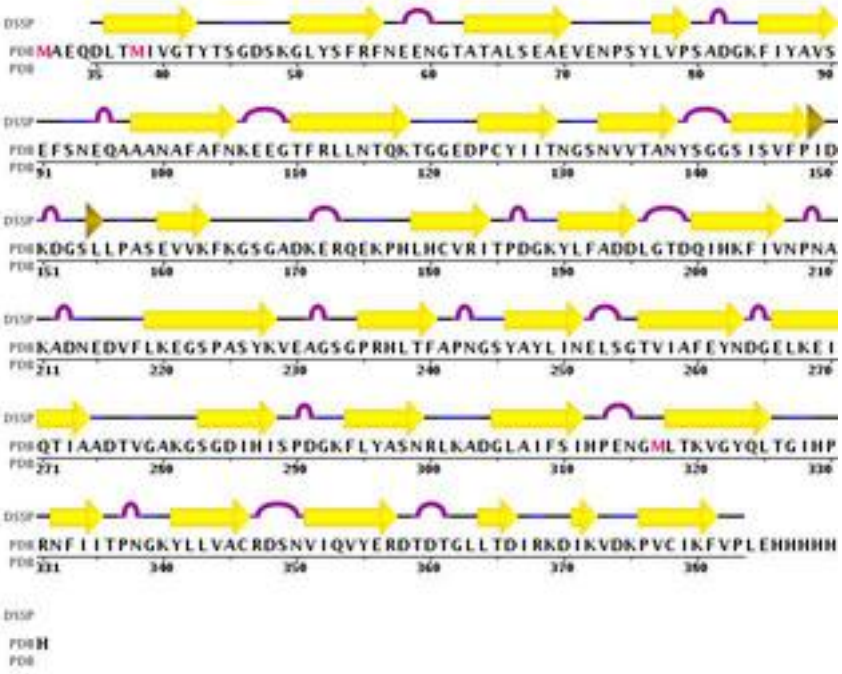
CEH56 4GN9

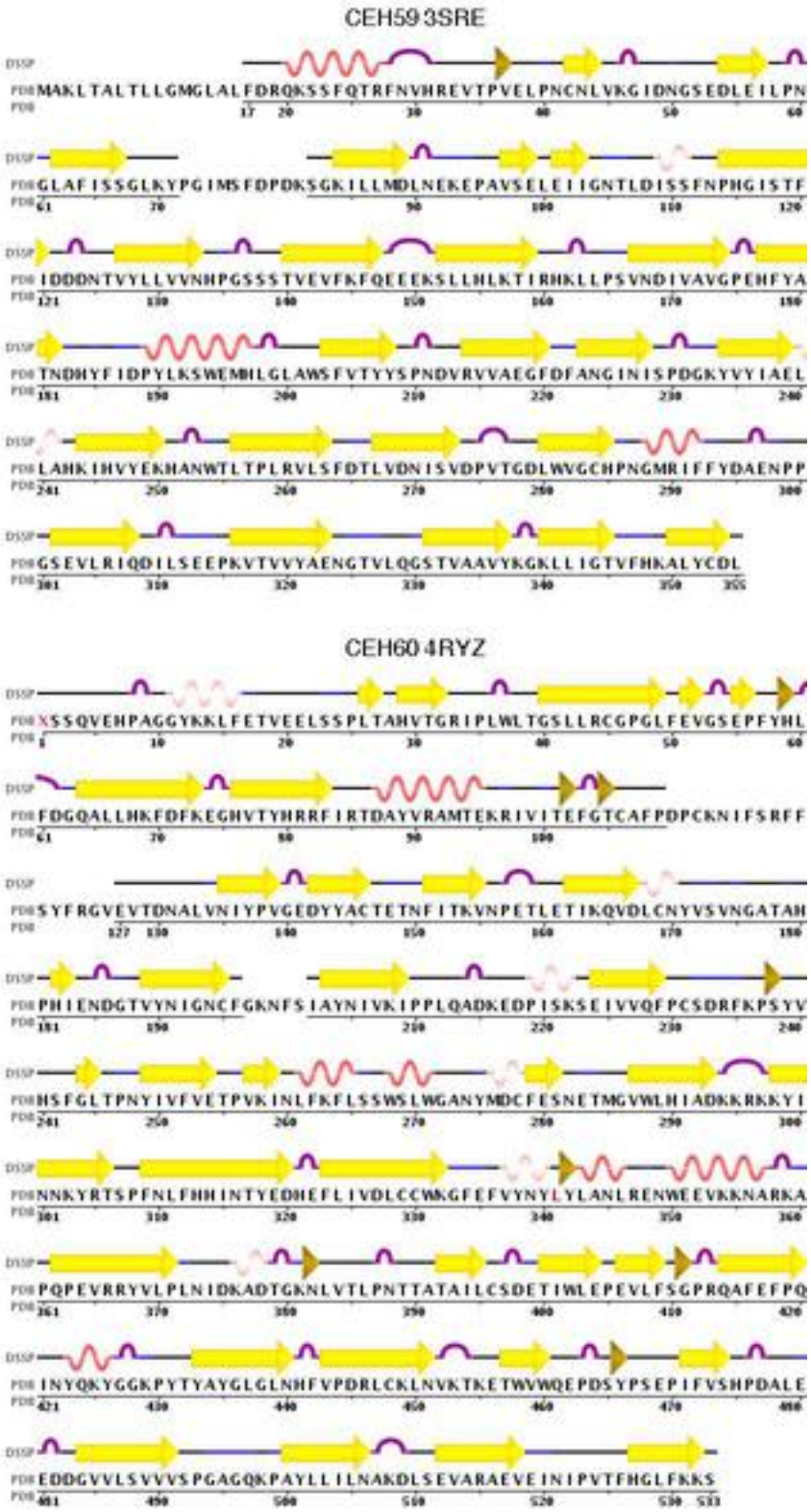


CEH57 2DG1

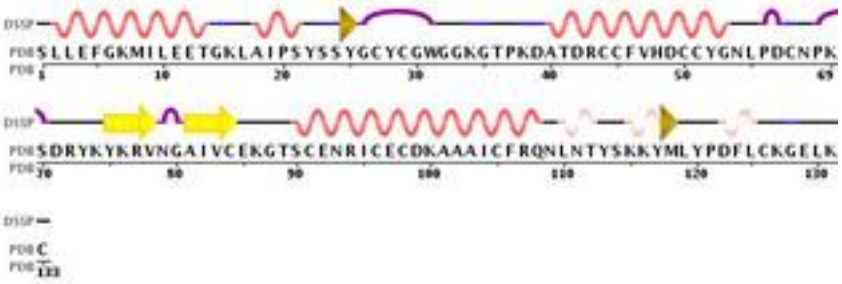


CEH58 3FGB

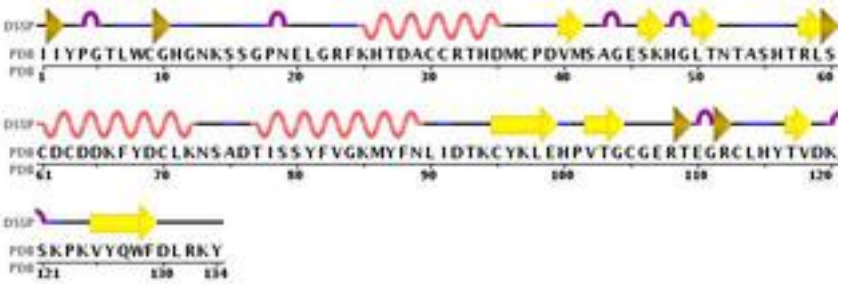



CEH59 3SRE
CEH60 4RYZ

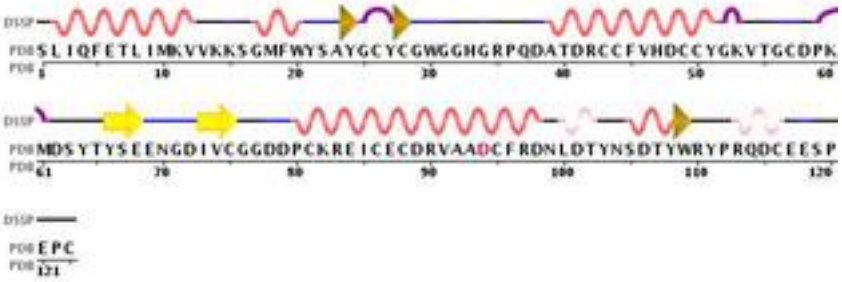
CEH61 1TG1



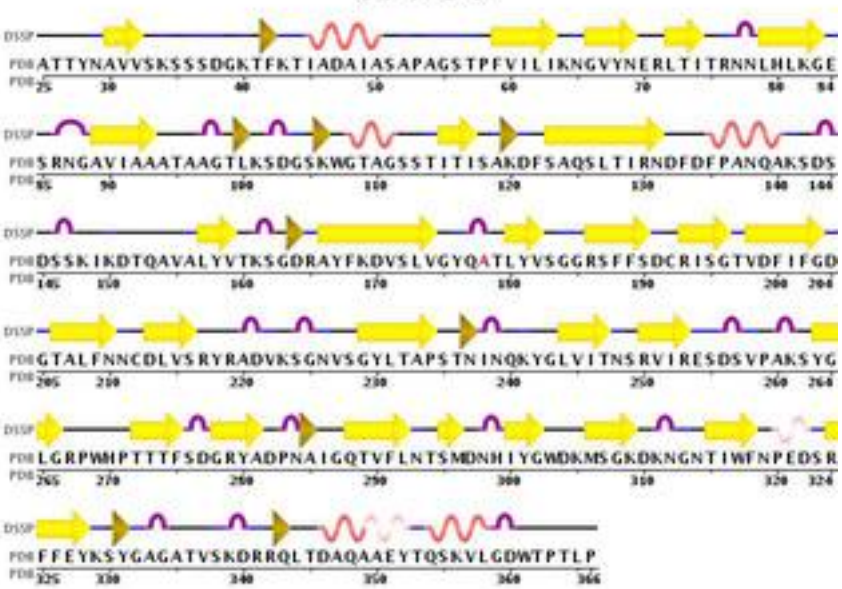
CEH62 1POC



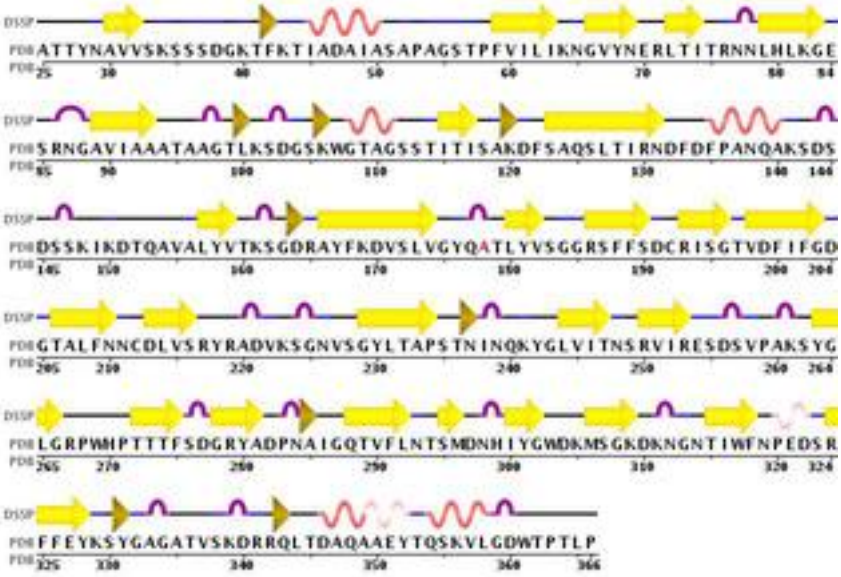
CEH63 1LJL



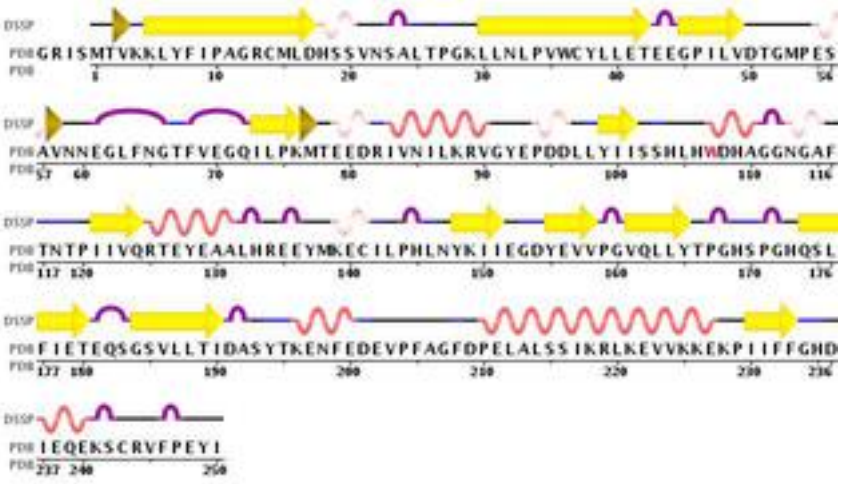
CEH64 2NSP



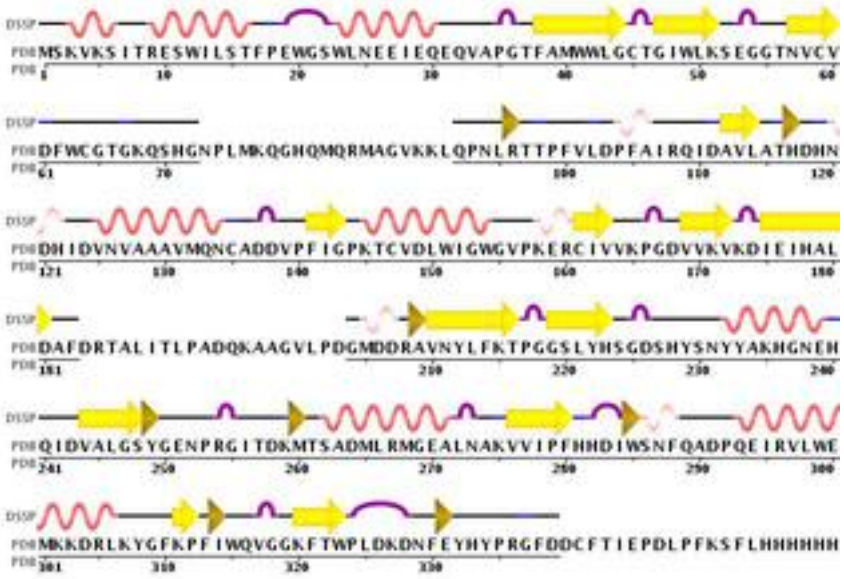
CEH65 2NSP



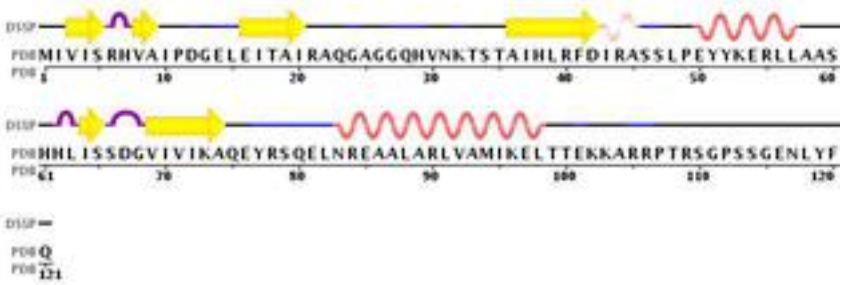
CEH66 4J5H



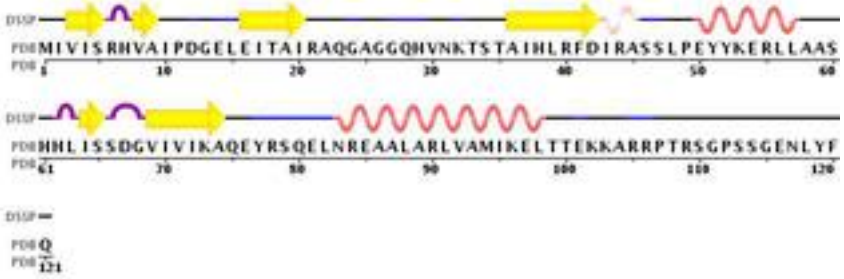
CEH67 2WYM



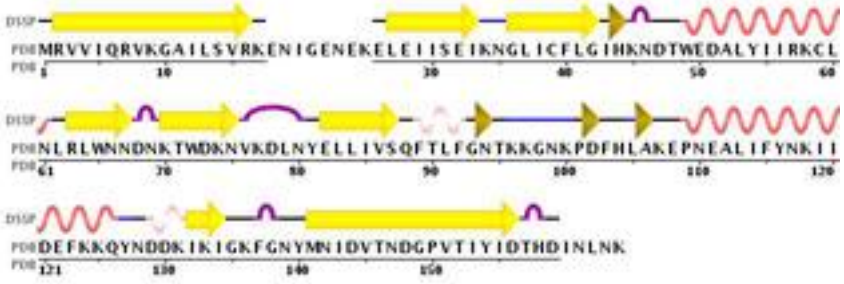
CEH69 2RTX



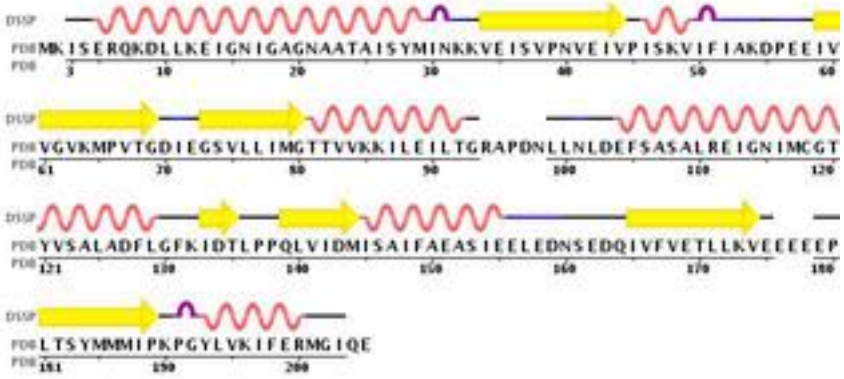
CEH69 2RTX



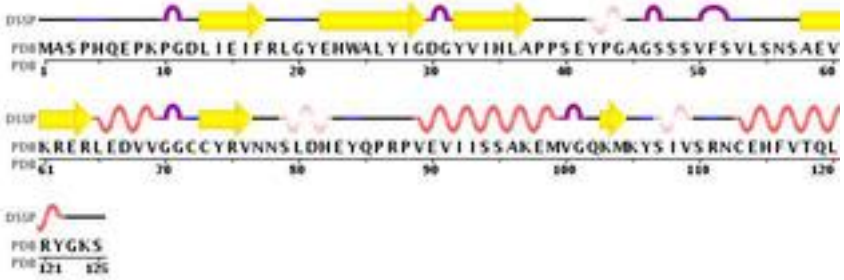
CEH71 3KO9



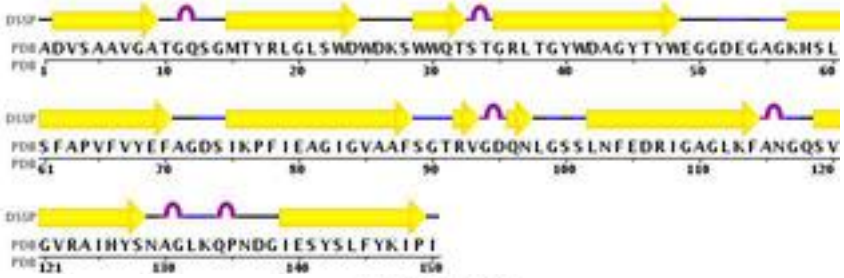
CEH73 2F9Z



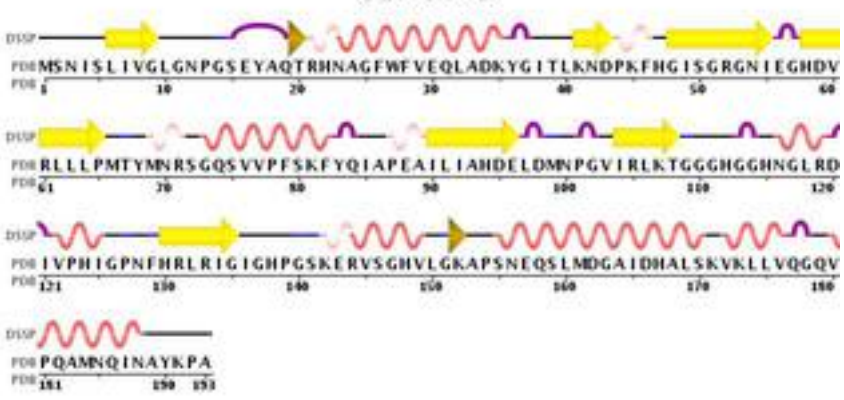
CEH74 2MY9



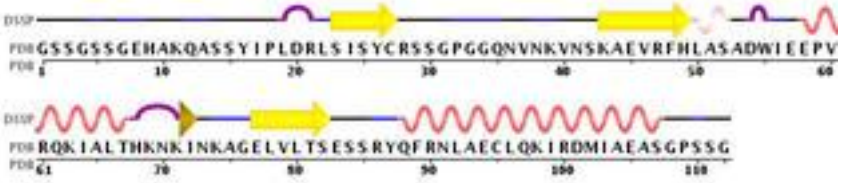
CEH76 2ERV



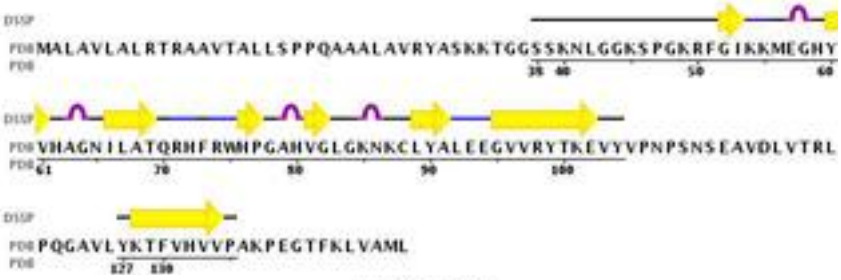
CEH78 4HOY



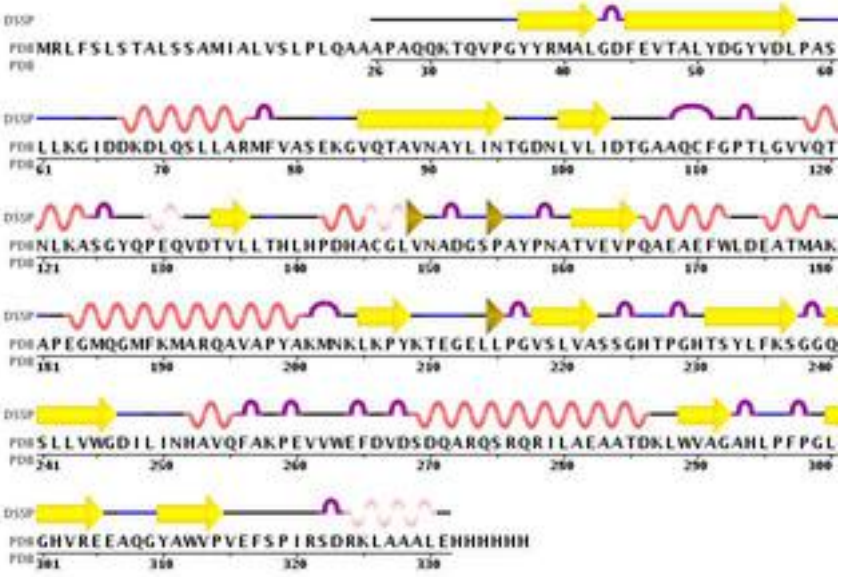
CEH79 1J26

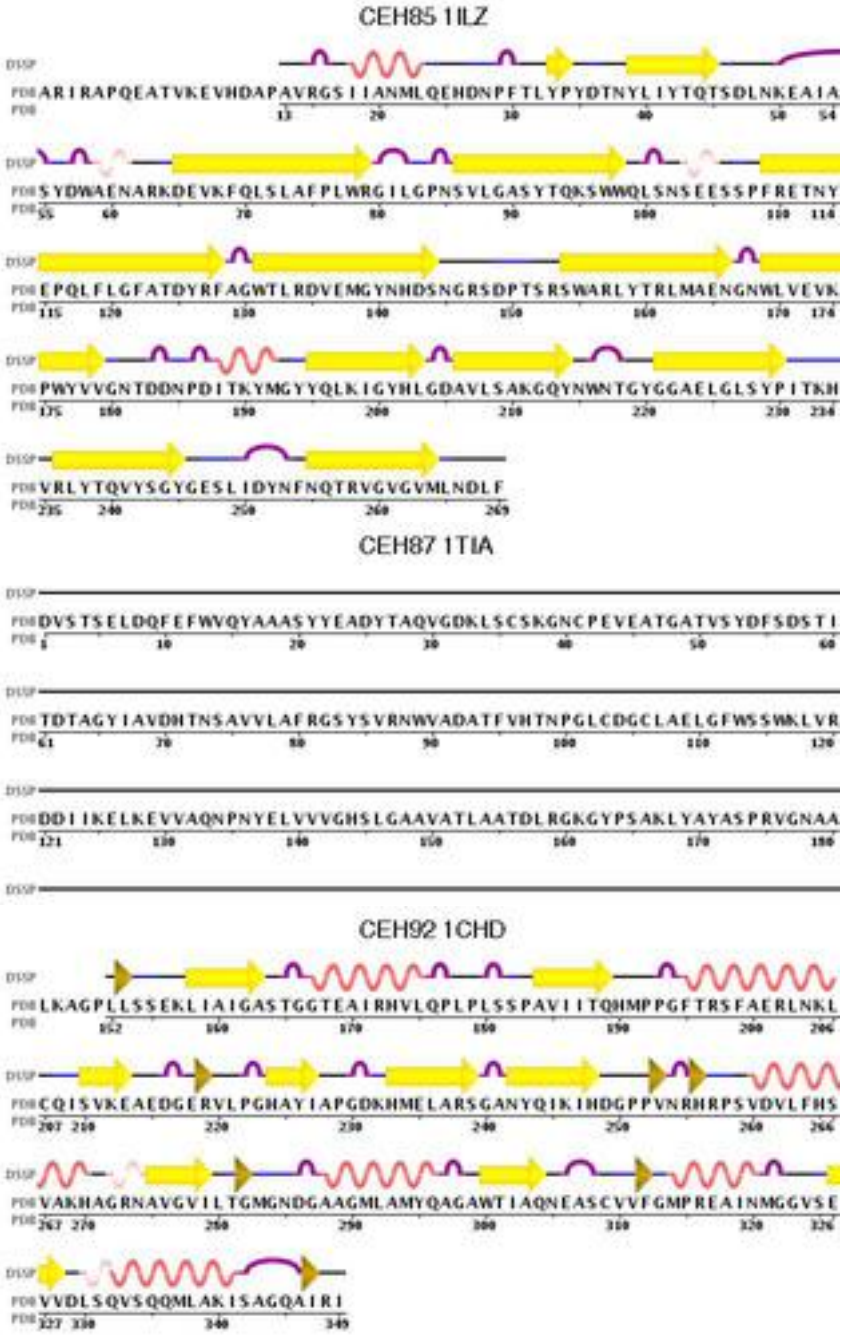


CEH79 4CE4

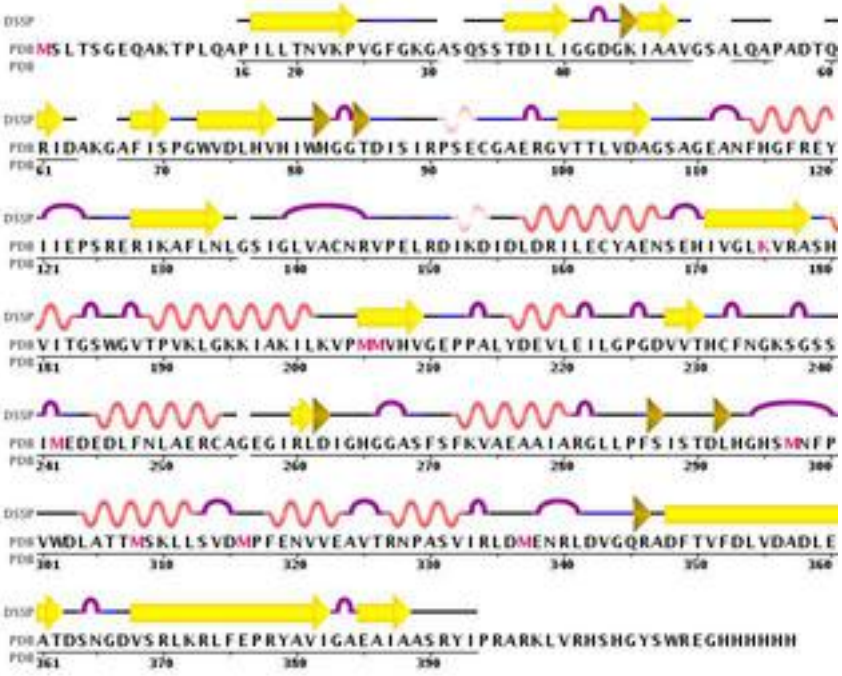


CEH84 4O98

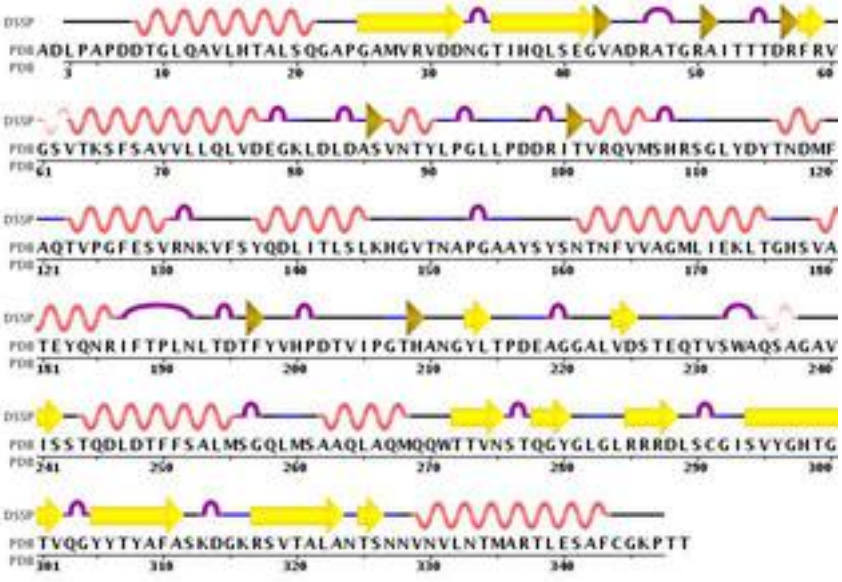


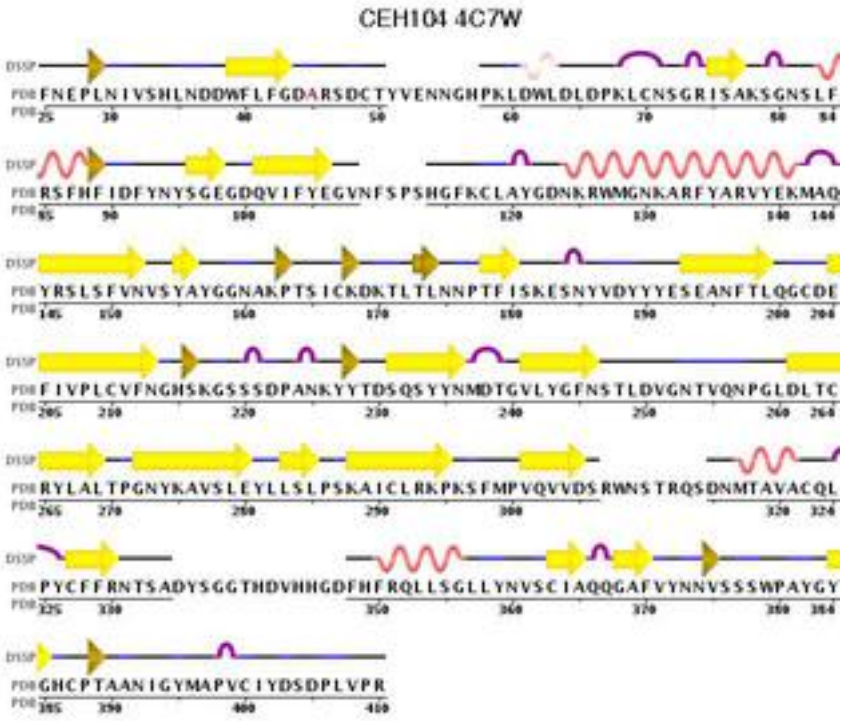

CEH85 1ILZ
CEH87 1TIA
CEH92 1CHD

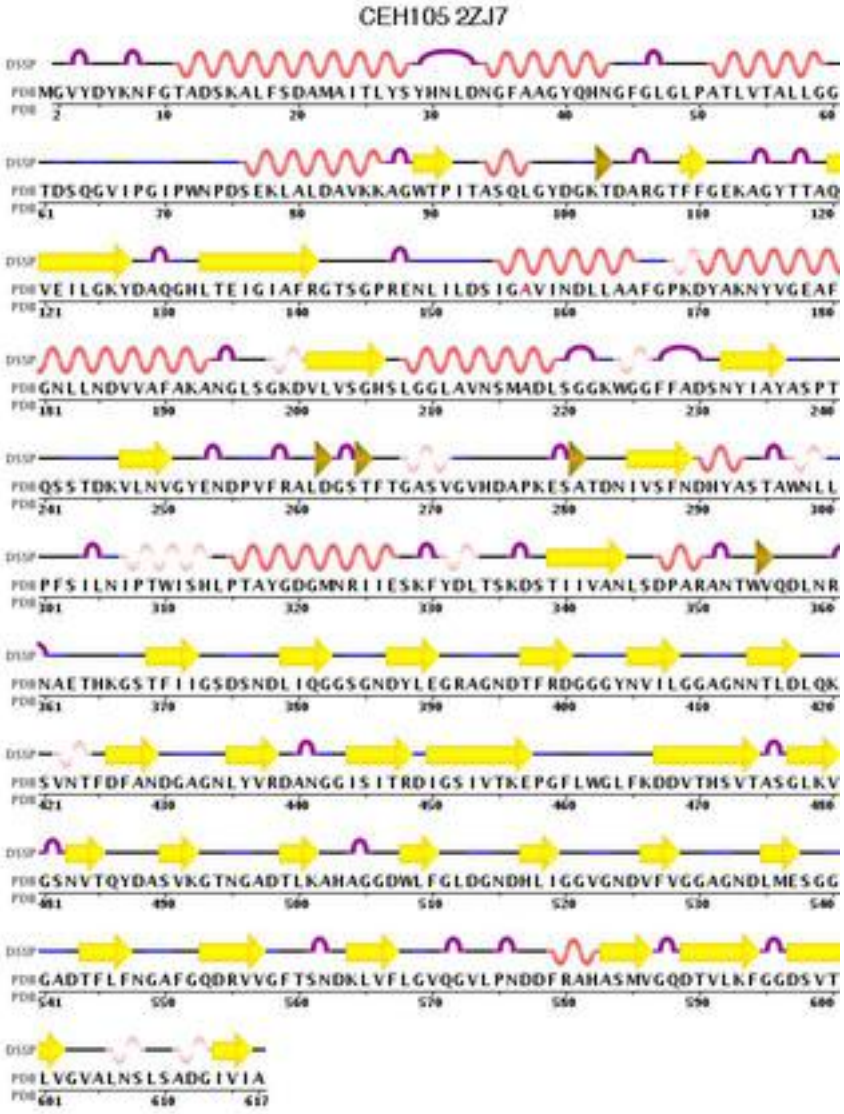
CEH95 2OGJ

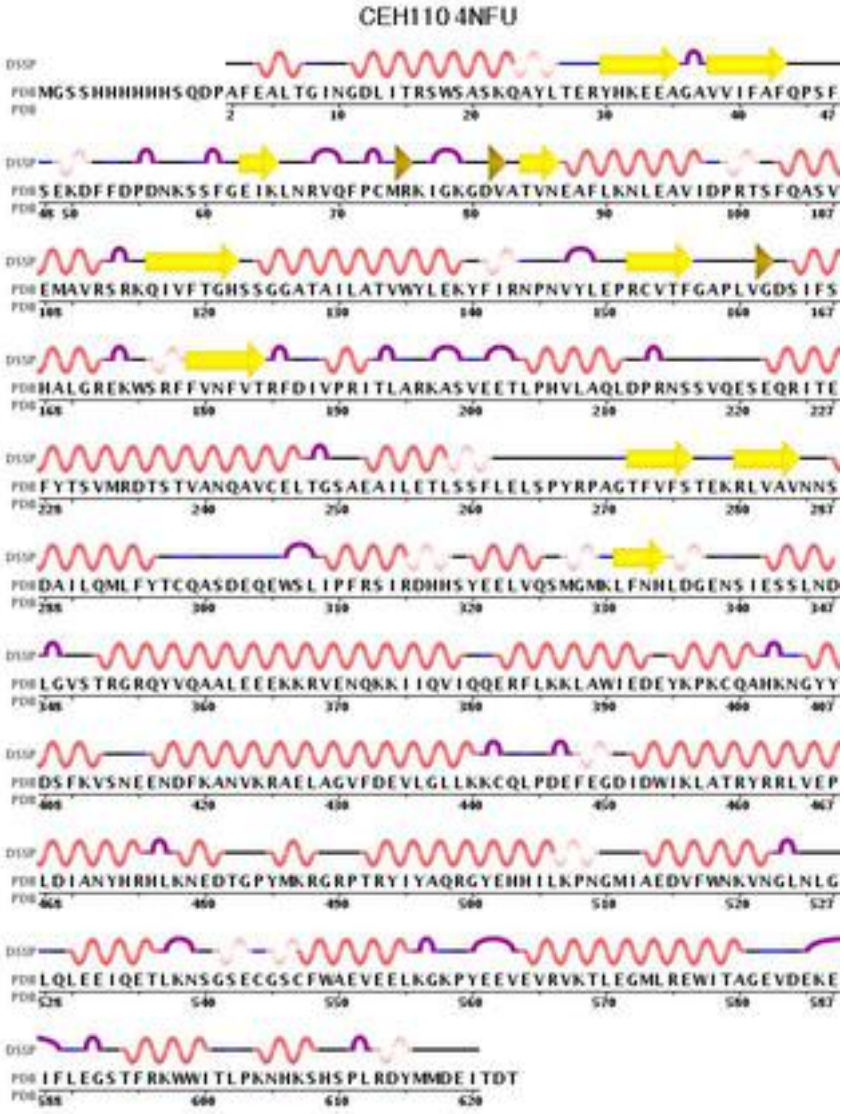


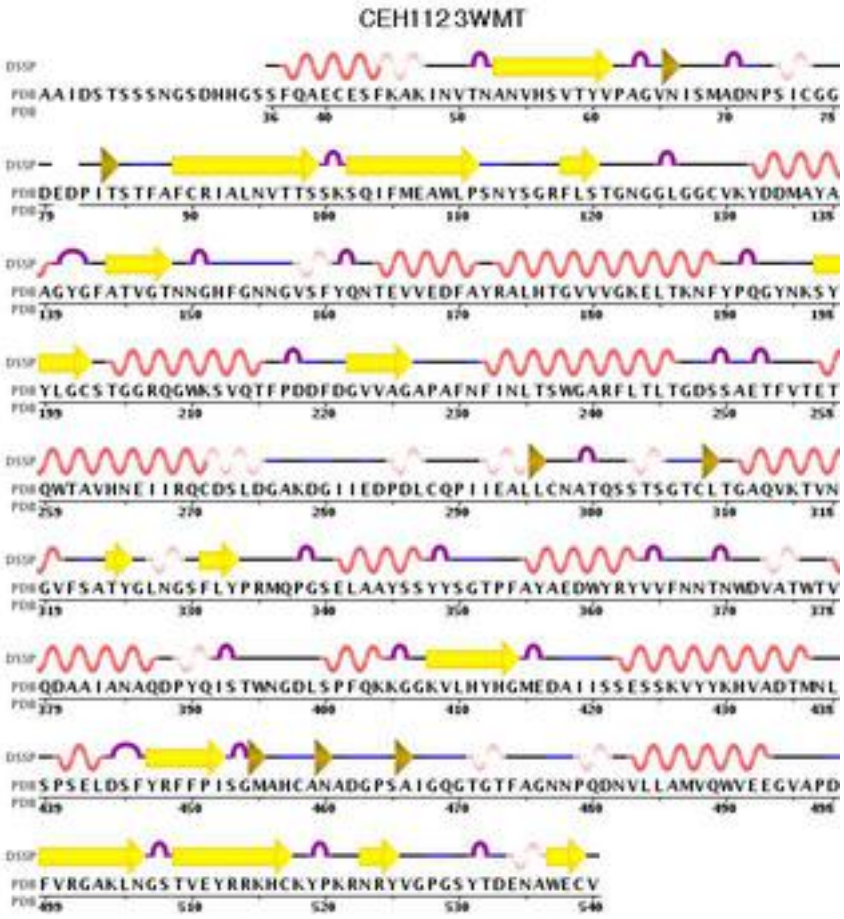
CEH98 1PW1

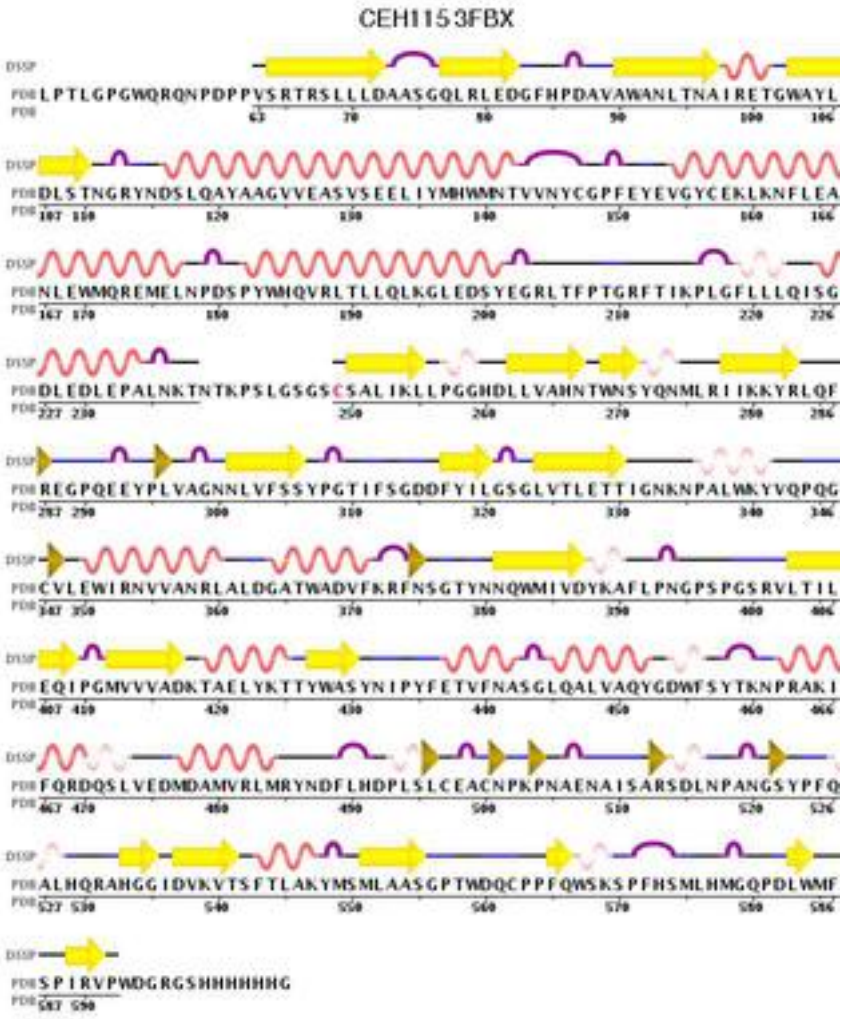



CEH104 4C7W
DSSP
PDB FNEPLNIVSHLNDDWFLFGDARSDCTYVENNGNPKLDWLDLDPKLCNSGRISAKSGNSLF
PDB 25 30 40 50 60 70 80 84
DSSP
PDB RSFHFIDFYNYSGEGDQVIFYEGVNFSPSHGFKCLAYGDNKRWMGNKARFYARVYEKMAQ
PDB 85 90 100 120 130 140 146
DSSP
PDB YRSLSFVNVSYAYGGNAKPTSICKDKTLTLNNPTFISKESNYVDYYYESEANFTLQGCDE
PDB 145 150 160 170 180 190 200 204
DSSP
PDB FIVPLCVFNGHSKGSSSDPANKYYTDSQSYYNMDTGVLYGFNSTLDVGNTVQNPGLDLTC
PDB 205 210 220 230 240 250 260 264
DSSP
PDB RYLALTPGNYKAVSLEYLLSLPSKAICLRKPKSFMPVQVVDSRWNSTRQSDNMTAVACQL
PDB 265 270 280 290 300 320 324
DSSP
PDB PYCFFRNTSADYSGGTHDVHHGDFHFRQLLSGLLYNVSCIAQQGAFVYNNVSSSWPAYGY
PDB 325 330 350 360 370 380 384
DSSP
PDB GHCPTAANIGYMAPVCIYDSDPLVPR
PDB 385 390 400 410

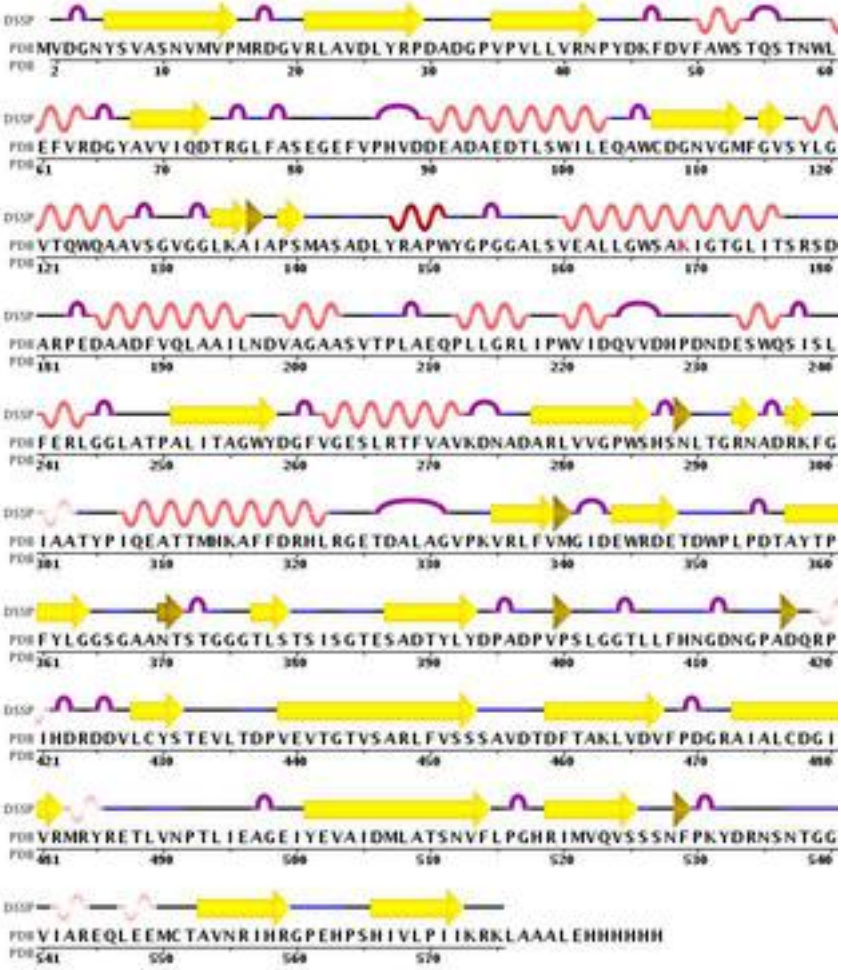

CEH105 2ZJ7
DSSP
PDB MGVYDYKNFGTADSKALFSDAMAITLYSYHNLDNGFAAGYQHNGFGLGLPATLVTALLGG
PDB 2 10 20 30 40 50 60
DSSP
PDB TDSQGVIPGIPWNPDSEKLALDAVKKAGWTPITASQLGYDGKTDARGTFFGEKAGYTTAQ
PDB 61 70 80 90 100 110 120
DSSP
PDB VEILGKYDAQGHLTEIGIAFRGTSGPRENLILDSIGAVINDLLAAFGPKDYAKNYVGEAF
PDB 121 130 140 150 160 170 180
DSSP
PDB GNLLNDVVAFAKANGLSGKDVLVSGHSLGGLAVNSMADLSGGKWGGFFADSNYIAYASPT
PDB 181 190 200 210 220 230 240
DSSP
PDB QSSTDKVLNVGYENDPVFRALDGSTFTGASVGVHDAPKESATDNIVSFNDHYASTAWNLL
PDB 241 250 260 270 280 290 300
DSSP
PDB PFSILNIPTWISHLPTAYGDGMNRIIESKFYDLTSKDSTIIVANLSDPARANTWVQDLNR
PDB 301 310 320 330 340 350 360
DSSP
PDB NAETHKGSTFIIGSDSNDLIQGGSGNDYLEGRAGNDTFRDGGGYNVILGGAGNNTLDLQK
PDB 361 370 380 390 400 410 420
DSSP
PDB SVNTFDFANDGAGNLYVRDANGGISITRDIGSIVTKEPGFLWGLFKDDVTHSVTASGLKV
PDB 421 430 440 450 460 470 480
DSSP
PDB GSNVTQYDASVKGTNGADTLKAHAGGDWLFGLDGNDHLIGGVGNDVFVGGAGNDLMESGG
PDB 481 490 500 510 520 530 540
DSSP
PDB GADTFLFNGAFGQDRVVGFTSNDKLVFLGVQGVLPNDDFRAHASMVGQDTVLKFGGDSVT
PDB 541 550 560 570 580 590 600
DSSP
PDB LVGVALNSLSADGIVIA
PDB 601 610 617


CEH110 4NFU
DSSP
PDB MGSSHHHHHHSQDPAFEALTGINGDLITRSWSASKQAYLTERYHKEEAGAVVIFAFQPSF
PDB 2 10 20 30 40 47
DSSP
PDB SEKDFFDPDNKSSFGEIKLNRVQFPCMRKIGKGDVATVNEAFLKNLEAVIDPRTSFQASV
PDB 48 50 60 70 80 90 100 107
DSSP
PDB EMAVRSRKQIVFTGHSSGGATAILATVWYLEKYFIRNPNVYLEPRCVTFGAPLVGDSIFS
PDB 108 120 130 140 150 160 167
DSSP
PDB HALGREKWSRFFVNFVTREDIVPRITLARKASVEETLPHVLAQLDPRNSSVQESEQRITE
PDB 168 180 190 200 210 220 227
DSSP
PDB FYTSVMRDTSTVANQAVCELTGSAEAILETLSSFLELSPYRPAGTFVFSTEKRLVAVNNS
PDB 228 240 250 260 270 280 287
DSSP
PDB DAILQMLFYTCQASDEQEWSLIPFRSIRDHHSYEELVQSMGMKLFNHLDGENSIESSLND
PDB 288 300 310 320 330 340 347
DSSP
PDB LGVSTRGRQYVQAALEEEKKRVENQKKIIQVIQQERFLKKLAWIEDEYKPKCQAHKNGYY
PDB 348 360 370 380 390 400 407
DSSP
PDB DSFKVSNEENDFKANVKRAELAGVFDEVLGLLKKCQLPDEFEGDIDWIKLATRYRRLVEP
PDB 408 420 430 440 450 460 467
DSSP
PDB LDIANYHRHLKNEDTGPYMKRGRPTRYIYAQRGYEHHILKPNGMIAEDVFWNKVNGLNLG
PDB 468 480 490 500 510 520 527
DSSP
PDB LQLEEIQETLKNSGSECGSCFWAEVEELKGKPYEEVEVRVKTLEGMLREWITAGEVDEKE
PDB 528 540 550 560 570 580 587
DSSP
PDB IFLEGSTFRKWWITLPKNHKSHSPLRDYMMDEITDT
PDB 588 600 610 620

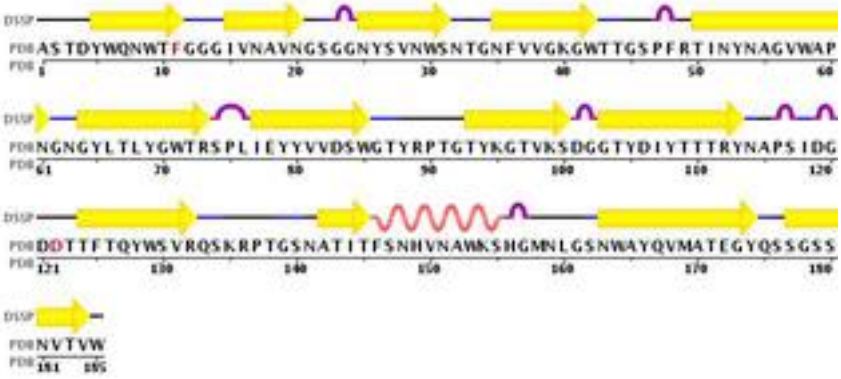

CEH1123WMT
DSSP
PDB AAIDSTSSSNGSDHHGSSFQAECESFKAKINVTNANVHSVTYVPAGVNISMADNPSICGG
PDB 30 40 50 60 70 75
DSSP
PDB DEDPITSTFAFCRIALNVTTSSKSQIFMEAWLPSNYSGRFLSTGNGGLGGCVKYDDMAYA
PDB 75 90 100 110 120 130 135
DSSP
PDB AGYGFATVGTNNGHFGNNGVSFYQNTEVVEDFAYRALHTGVVVGKELTKNFYPQGYNKSY
PDB 135 150 160 170 180 190 195
DSSP
PDB YLGCSTGGRQGWKSVQTFPDDFDGVVAGAPAFNFINLTSWGARFLTLTGDSSAETFVTET
PDB 195 210 220 230 240 250 255
DSSP
PDB QWTAVHNEIIRQCDSLDGAKDGIIEDPDLCQPIIEALLCNATQSSTSGTCLTGAQVKTVN
PDB 255 270 280 290 300 310 315
DSSP
PDB GVFSATYGLNGSFLYPRMQPGSELAAYSSYYSGTPFAYAEDWYRYVVFNNTNWDVATWTV
PDB 315 330 340 350 360 370 375
DSSP
PDB QDAAIANAQDPYQISTWNGDLSPFQKKGGKVLHYHGMEDAIISSESSKVYYKHVADTMNL
PDB 375 390 400 410 420 430 435
DSSP
PDB SPSELDSFYRFFPISGMAHCANADGPSAIGQGTGTFAGNNPQDNVLLAMVQWVEEGVAPD
PDB 435 450 460 470 480 490 495
DSSP
PDB FVRGAKLNGSTVEYRRKHCKYPKRNRYVGPGSYTDENAWECV
PDB 495 510 520 530 540

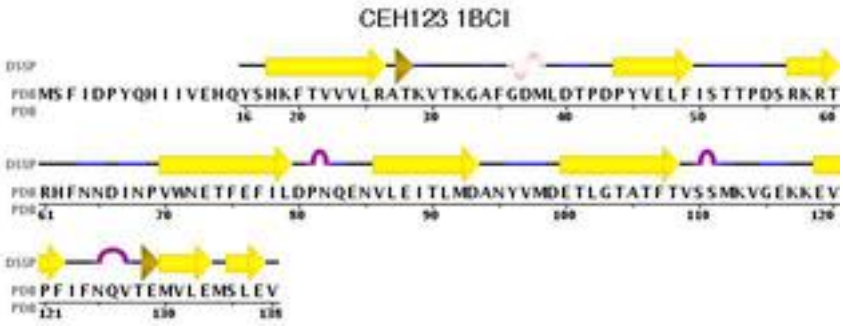

CEH1153FBX

CEH1163I2H



CEH1193EXU




CEH123 1BCI
DSSP
PDB MSFIDPYQHIIVEHQYSHKFTVVVLRATKVTKGAFGDMLDTPDPYVELFISTTPDSRKRT
PDB 10 20 30 40 50 60
DSSP
PDB RHFNNDINPVWNETFEFILDPNQENVLEITLMDANYVMDETLGTATFTVSSMKVGEKKEV
PDB 61 70 80 90 100 110 120
DSSP
PDB PFIFNQVTEMVLEMSLEV
PDB 121 130 135