POLYMER PROPERTIES THAT PREDICT PROTEIN STRUCTURE CLASS FROM THE PRIMARY SEQUENCE

by

Nathan Khaodeuanepheng, B.S.

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a major in Biochemistry
December 2022

Committee Members:

Steven Whitten, Chair

Karen Lewis

Kevin Lewis

COPYRIGHT

by

Nathan Khaodeuanepheng

2022

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

This work is protected by the Copyright Laws of the United States (Public Law 94-553, section 107). Consistent with fair use as defined in the Copyright Laws, brief quotations from this material are allowed with proper acknowledgement. Use of this material for financial gain without the author's express written permission is not allowed.

Duplication Permission

As the copyright holder of this work I, Nathan Khaodeuanepheng, authorize duplication of this work, in whole or in part, for educational or scholarly purposes only.

DEDICATION

To my parents, Vong Khaodeuanepheng and Diane Khaodeuanepheng, to my sister and brother in-law, Catherine Khaodeuanepheng Gilbert, and Aaron Gilbert. Without the consistent love and support y'all poured into me while I chased my dreams I would not be in this position.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor Dr. Steven T. Whitten for his endless support and guidance throughout this process. Without the patience, dedication, and mentorship Dr. Whitten has provided none of this would be possible – I sincerely appreciate the amount of time you have invested into both this project and me.

I would like to thank my committee members, Dr. Karen Lewis, and Dr. Kevin Lewis for taking the time to invest into my education as well. I would like to thank my extended family, friends, and mentors for their words of encouragement during those rough days. Lastly, I would like to thank Dr. Dana Dean for believing in me at one of the most critical points in my academic career.

TABLE OF CONTENTS

Page

LIST OF TABLES

Tables **Page**

LIST OF FIGURES

Figures **Page**

ABSTRACT

Within cells, membrane-free compartments form spontaneously and reversibly in a process referred to as "phase separation". By forming specific compartments and micro-environments, these membraneless organelles, for example, Cajal bodies, the nucleolus, stress granules, and P-bodies, are used to regulate a myriad of cellular functions via control of the spatial organization of biological matter and the concomitant modulation of biochemical reactivity. Proteins have a prominent role driving phase separation and, among phase-separating (PS) proteins, many have intrinsically disordered regions (IDRs) that are needed for phase separation to occur. Previous work created a computer algorithm called ParSe (Partition Sequence) that successfully identifies PS IDRs from the protein primary sequence starting from predictions of hydrodynamic size, which is indicative of the relative strength of intramolecular as compared to solvent interactions. The key assumption of ParSe is that intramolecular cohesion that compacts monomeric proteins is correlated with intermolecular cohesion that drives phase separation. To assess hydrodynamic size, ParSe uses a sequence-based model of the polymer scaling exponent, ν_{model}, that when paired with a second sequence-based parameter, the intrinsic propensity for a sequence to form β-turns, can distinguish between sequences belonging to one of three classes of protein regions: folded, ID, and PS ID. However, the prior study did not test whether the combination of ν_{model} and β-turn propensity is unique in its predictive power, as would be required if hydrodynamic size and turn structures are indeed mechanistically linked to protein phase separation. Here, it is shown that ν_{model} and β-turn propensity are not unique in their ability

to identify PS IDRs but rather this can be done with similar fidelity using ν_{model} paired with a range of different types of conformational propensity scales or hydrophobicity scales. Thus, structural hypotheses relating to the mechanistic details of protein phase separation cannot be established based on these results. Moreover, when applying ParSe to verified globular proteins, we noticed that these proteins often contain short regions that are incorrectly predicted to be ID. We hypothesize that these predicted short IDRs within known folded regions represent segments within a folded domain that have low structural stability. To test this hypothesis, ParSe calculations were compared to hydrogen-deuterium exchange (HDX) data measured in four folded proteins. Good agreement between the locations of ParSe-predicted IDRs and regions with low stability as inferred by HDX rates were found.

1. INTRODUCTION

1.1 Introduction

Proteins, a class of biomacromolecule, are linear polymers comprised of subunits called amino acids. These macromolecules exist in various subgroups and perform varying functional roles that are essential in maintaining cellular health (1–7). For example, proteins regulate gene expression, catalyze biochemical reactions, act as transporter molecules, and give structural support to the cell (1–3). To understand mechanistically how proteins can perform such tasks, structural analyses are used because they show the arrangement of specific chemical groups in the macromolecule and how these groups are combined to form a functional unit (1, 8).

Specifically, the structural and chemical properties of proteins are defined by the linear sequence of amino acids that make up the polymer chain, which is referred to as the primary sequence (8, 9). There are 20 common, naturally occurring amino acid types. Each amino acid type is composed of the same basic structure: a central carbon atom that is covalently linked to an amine group, a carboxyl group, a side chain group, and a hydrogen (**Figure 1.1**). However, the functional group that resides on the side chain is different among the amino acid types (8, 10). Thus, the side chain at each residue position in the protein will have specific chemical properties such as polarity, size, and hydrophobicity that, by the pattern of these properties in the primary sequence, determines how and if the protein folds, which, in turn, governs function.

The folding of a protein into its native, globular structure is known to involve a hydrophobic-driven collapse of the polypeptide chain (11, 12). The formation of the hydrophobic core is both entropically and enthalpically favorable due to the increased

mobility of water molecules released into the bulk solvent when hydrophobic side chains become buried, and the subsequent formation of internal, favorable contacts between peptide subunits in the folded structure (4, 13, 14). Alternatively, residues containing hydrophilic and charged side chains groups prefer solvent-interactions (i.e., with water) over self-interactions, are often found at the surface of globular protein structures (7, 8, 12). These intramolecular and intermolecular interactions that form due to the side chain properties of the protein determine specific folding motifs, especially for secondary structures that are defined by the pattern of internal hydrogen bonding between backbone hydrogen bond donors and hydrogen bond acceptors. The three predominant types of secondary structures found within proteins are α-helices, β-sheets, and β-turns (16). An α-helix is defined as a right-handed helix that contains 3.6 residues per turn, where the hydrogen of the amine group of an amino acid residue (*i*) interacts with the oxygen of the carboxyl group of a different residue (*i+4*) that is found four residues earlier within the polypeptide chain *(i* to *i+4)* (**Figure 1.2A**). Unlike α-helices, which are discrete units of sequentially contiguous residues, β-sheets form from multiple β-strands that can be distant in the protein sequence (5). An individual β-strand is typically four to ten amino acid residues in length, where side chain groups alternate above and below an extended backbone conformation (17, 18). Hydrogen bonding between β-strands forms β-sheets that are oriented parallel (i.e., when the N- and C-termini of different strands are in the same direction) or antiparallel (i.e., when the N- and C- termini of different strands are in the opposite direction) relative to each other. β-sheets also can be a mixture of both parallel and antiparallel strands (17, 18). β-turns, or reverse turns, are non-repetitive secondary structures that allow the polypeptide chain to reverse direction nearly 180 degrees (**Figure**

1.2C) (19). β-turns often are found in between β-strands in a β-sheets. Structurally, the β-turn is made from four sequential amino acid residues (*i, i+1, i+2,* and *i+3*) with a hydrogen bond connecting the *i* and *i+3* positions (20). β-turns also have functional roles such as sites for post-translational modification, protein-protein interactions, and substrate binding (16, 17, 21, 22).

These secondary structures, supported in part by internal hydrogen bonding, provide stability to the native fold of a protein. Other intramolecular interactions that are known to stabilize the native fold include disulfide bridges between thiol side chains, salt bridges between oppositely charged groups, and van der Waals interactions between hydrophobic groups (23, 24). As such, specific amino acid types are found to have propensities, for or against, in forming the different secondary structure types, due to their physical characteristics. For example, the uncharged amino acid residues alanine and leucine strongly favor α-helical structure, while proline and glycine do not (16, 25). Proline has been defined as a "helix breaker" due to the steric hinderance that occurs from the large side chain group that forms a covalent bond with the preceding backbone amide group. The presence of this side chain group near the main chain backbone causes a kink in the α-helix to avoid a steric clash (26). Further, the covalent link from the side chain to the preceding backbone amide eliminates a hydrogen bond donor group that could otherwise form a α-helix-stabilizing hydrogen bond. When compared to other amino acids, glycine is unique in that its side chain is a single hydrogen atom, making it the smallest side chain of the common amino acid types. Accordingly, the polypeptide chain has higher conformational flexibility at glycine positions, and thus adopting a helical structure is less favorable entropically wherever glycine is present (27, 28).

While many, but not all (29), proteins have been found to adopt stable globular structures, these folded structures are found to exhibit marginal stability (13, 14, 30). The stability of protein structures can be defined thermodynamically by the difference in free energy between the folded and unfolded states under equilibrium conditions (31–33). The overall free energy of the folded state in globular proteins is typically between -5 to -15 kcal/mol (34), and thus protein folding is generally spontaneous and favorable. Free energies at such values, however, also reveal that under physiological-like conditions, the unfolded state is populated at small, but not insignificant, amounts. Indeed, owing to this marginal stability, proteins are observed to unfold and refold repeatably *in vivo* (35).

1.2 Intrinsically disordered proteins

While most biological proteins show some stability for a native conformation, there are some proteins that do not. These proteins do not fold under normal biological conditions and thus have been classified as "intrinsically disordered", or ID. ID proteins (IDPs) are found in all kingdoms of life (36) and, in humans, comprise ~10% of the proteome. An additional ~30% of human proteins have mixed structures, in the sense that these proteins contain both large, folded regions and large ID regions (IDRs) (37). Not surprisingly, the compositions of IDPs and IDRs are depleted in hydrophobic amino acid types (38, 39), and thus their sequences are incapable of forming a globular structure with a hydrophobic core. Instead, IDPs and IDRs are enriched in the charged and polar amino acid types (38, 40). Further, based on the observed frequencies of the amino acid types in ID or folded regions in surveys of proteins (10, 41), some amino acid types have been labeled as "order promoting" (W, C, F, I, Y, V, L and N) and some as "disorder promoting" (A, R, G, Q, S, P, E, and K). As expected, order promoting residues are mostly hydrophobic and

uncharged, while disorder promoting residues are mostly hydrophilic and charged (41). The remaining amino acids (H, M, T and D) are ambiguous and equally common to both ordered and disordered regions (10). IDPs and IDRs also exhibit low sequence complexity when compared to folded proteins, meaning their sequences can be highly repetitive (41, 42). Though the presence of ID seems in conflict with the widely held notion of a direct link between protein structure and protein function, IDPs and IDRs are found to facilitate a range of important biological processes, such as cell cycle regulation, cellular signaling, transcription, and translation (37, 39, 41).

A new class of IDPs has been identified recently due to their underlying importance in maintaining cellular health (43–45), and their significance in various neurogenerative diseases (44, 46, 47). These proteins are differentiated from other IDPs by the ability to drive a process called liquid-liquid phase separation (LLPS), where proteins and other macromolecules spontaneously de-mix from the cellular milieu to form biomolecular condensates, also referred to as membraneless organelles (MLOs) (**Figure 1.3**) (43, 44). Unlike canonical membrane-bound organelles such as the golgi apparatus, nucleus, and the endoplasmic reticulum, MLOs are membrane-free and lack a lipid bilayer. MLOs can be found in the cytoplasm (e.g., P-granules) and the nucleus (e.g., Cajal bodies), and they have been associated with many biological roles, such as driving intracellular compartmentalization, cell differentiation, and ribosome biogenesis (43, 47, 48). Indeed, dysregulation of intracellular LLPS has been linked to certain pathological states, for example neurotoxic amyloid and fibril particles found in Parkinson's and Alzheimer's diseases may begin as biomolecular condensates that harden to plaques (44, 49, 50).

To better understand the mechanism by which MLOs form, the process of LLPS and the components that are believed to drive LLPS have been closely scrutinized (51–53). Characteristics such as highly repetitive low complexity sequences, enrichment in polar residues, side chain properties that promote cation-π interactions (e.g., between arginine and aromatic residues), and π-π interactions (e.g., between aromatic groups) are believed to be essential in driving the formation of some MLOs (52) .

IDRs are hypothesized to serve as points of multivalent interaction for these self-interactions to occur, thus playing a role in the process of LLPS (51, 52). Recently, to further conceptualize LLPS, Banai *et al.* have proposed the model of "scaffolds" and "clients". Scaffolds are the essential regions found within proteins that self-associate through multivalent interactions to drive LLPS, whereas clients are non-essential regions that are recruited by their interactions with scaffold regions (45, 54). Similarly, Martin *et al.* also proposed the model of "stickers" and "spacers" that has been adapted from the field of associative polymers (52) (**Figure 1.4**). Sticker regions act as "sticky" points for multivalent crosslinks to occur with other "stickers" found in the polymer chain (51). "Spacers" are linker regions that separate "stickers" in the chain to provide spacing (14, 40, 54). Though spacers do not directly drive LLPS, spacer regions impact the flexibility of the chain and therefore modulate the ability of sticker regions to interact with one another (14, 40, 54).

1.3 IDP hydrodynamic size as a predictor of LLPS potential

The propensity for a particular protein to phase separate is generally thought to be driven by a preference for protein-protein over protein-solvent interactions (14, 40, 50, 54), whereby protein-protein interactions are needed to drive the formation of a protein-dense phase. The same interactions thought to drive LLPS also have been hypothesized to affect hydrodynamic size in monomeric IDPs (40, 56). Conceptually, IDPs that prefer contact with the solvent will have elongated and swollen structures when compared to the compacted dimensions observed when self-interactions dominate (**Figure 1.5**).

One framework useful for quantifying the balance of self-interactions versus solvent-interactions is the polymer scaling exponent, ν. Polymer scaling exponent was originally derived for use with long, homopolymers where subunit-subunit interactions are equivalent, as are, separately, subunit-solvent interactions (57, 58). This exponent is obtained from the dependence of size (e.g., hydrodynamic radius, R_h, or radius of gyration, R_g) on polymer length, N, in the power law relationship, $R_h \sim N^{\nu}$ (58). Because IDPs are heteropolymers and have varying, spatially organized, local interactions, ν from an IDP represents a phenomenological parameter rather than an exact description of the balance of molecular forces present. Nonetheless, numerous studies have shown that polymer properties such as ν, derived for use in homopolymers, can be successfully applied to biological IDPs to help understand their observed solution behavior (6).

When measured in biological proteins, experimental ν is often ~0.3 for folded proteins, ~0.5 for IDPs in water, and ~0.6 for chemically denatured, unfolded proteins in solutions with high concentrations of guanidine hydrochloride or urea (59). This trend indicates that ν increases as the protein structure is increasingly solvated. Because

biomolecular LLPS includes the exchange of macromolecule-solvent interactions for macromolecule-macromolecule interactions (51, 60, 61), *ν* could be a predictor of LLPS potential among heteropolymeric IDRs. Supporting this idea, numerous studies have found that the hydrodynamic dimensions of some IDRs are correlated to the temperature dependence in LLPS behavior.

To test the hypothesis that *ν* could be a predictor of LLPS potential among IDRs, Whitten and coworkers recently demonstrated that mean *ν*, as predicted by the protein sequence, is indeed smaller in IDRs known to phase separate when compared to IDRs in general (62). Despite the difference in means, there was significant statistical overlap in *ν* between the two IDR sets (i.e., phase-separating IDRs and non-phase-separating IDRs), indicating that sequence-calculated *ν* (referred in the study as ν_{model}) has low predictive power for LLPS by itself. However, the intrinsic propensity for β-turn structures, calculated from sequence, also was different between the two IDR sets, and when combined with ν_{model}, these two sequence-calculated values showed the ability to predict protein class: for folded, phase-separating, and non-phase-separating IDRs . This result was robust to choice of β-turn propensity scale (62). Molecular modeling was used to explore the origins of the ability for turn propensity to predict LLPS potential, and it was found that transient β-turn structures reduce the de-solvation penalty of forming a protein-rich phase and increase exposure of atoms involved in π/sp^2 valence electron interactions (62), which have been identified in other studies as sticker sites (52, 55). By this proposed mechanism, β-turns act as energetically favored nucleation points.

In that study, the ability of other sequence-based properties to predict protein classification (i.e., for folded, phase-separating ID, and non-phase-separating ID) was not

explored, though such information could provide additional mechanistic insight. For example, charge-based protein-protein interactions are thought to contribute to LLPS behavior (51, 56, 60); as are hydrophobicity considerations (11, 12, 60). As such: do sequence-based calculations of the protein charge, or hydrophobicity, yield predictive capabilities? Moreover, LLPS is thought to be driven by multivalent interactions, and the types of these interactions, listed in **Table 1.1**, are chemically diverse (40, 52, 61). Thus, potentially, many sequence-based properties could be used to predict protein class from sequence and, if so, β-turn propensity and ν_{model} would not be unique in this ability. As an added point, in the prior study, the set of folded sequences used to establish ν_{model} and β-turn propensity as predictive parameters was obtained from analysis of the folded regions in LLPS proteins (62). Because LLPS proteins are a small subset of globular proteins in general, it is not confirmed if the mean properties in this limited set of folded sequences is fully representative. Indeed, compositional differences are found when globular proteins from different subsets are compared – for example mesophile *versus* extremophile (63) and membrane protein *versus* cytosolic protein (7, 22, 64)

1.4 Project Goals

The goal of this project is to determine the breadth of amino acid properties that, when combined with ν_{model}, could be used to predict protein class for phase-separating ID, non-phase-separating ID, and folded regions given only the primary sequence. To do this, protein databases containing subsets of proteins that are folded, ID, or ID and known to spontaneously phase separate were analyzed. The database of folded proteins included curated sets of non-homologous human proteins (14), small to large proteins (65), extremophilic proteins (63), membrane proteins, and metamorphic, or "fold-switching",

proteins (66). A set of IDRs from proteins that exhibit LLPS behavior were obtained from Vernon *et al.* (67), the PhaSePro database (68), and the DisProt database (69), chosen because each contains protein lists that have been manually curated for experimentally verified cases of LLPS. A set of IDPs not known to phase separate but that remain monomeric under normal solution conditions was assembled from literature reports (58, 59, 62, 70–76, 76–84) . The Amino Acid Index database (85), which maintains and updates a list of amino acid property scales pulled from the scientific literature, provided a set of 566 amino acid scales that cover a wide range of properties, including many based on charge, hydrophobicity, and conformational propensities, as well as some more exotic scales based on refractivity, polarizability, and even melting points. Added to this list of amino acid scales was a newly developed hydrophobicity scale designed with the specific goal of differentiating phase-separating and non-phase-separating IDRs (86, 87).

Reported herein, we found that β-turn propensity, when paired with ν_{model}, was not alone in its ability to identify protein class from the primary sequence. Instead, sequence-calculated turn propensity could be substituted for almost any intrinsic conformational propensity scale with little loss of prediction fidelity. Our findings suggest that structural differences among the protein classes are strongly encoded in the primary sequence, even when comparing phase-separating and non-phase-separating IDRs that, of course, lack stable structures. We found multiple β-sheet, α-helix, and hydrophobicity scales with similar, and sometimes better, ability to statistically distinguish protein class when compared to β-turn propensity scales. Hydrophobicity-based scales were best at distinguishing folded from ID, with, in general, small statistical evidence for discerning phase-separating and non-phase-separating IDRs.

When applying sequence-based calculations of β-turn propensity and ν_{model} to verified globular proteins, we noticed that these proteins often contain short regions that are incorrectly predicted to be ID. We hypothesize that these predicted short IDRs within known folded regions represent segments within a folded domain that have low structural stability. Many studies have shown that proteins exhibit differences in conformational stability across their folded structure (30, 63) with, generally, higher stability associated with a buried hydrophobic core and lower stability found in solvent-exposed loop structures (11, 12). To test this hypothesis, we compared sequence-based calculations of β-turn propensity and ν_{model} at the residue-level to hydrogen-deuterium exchange (HDX) data measured in four well-characterized and folded proteins: staphylococcal nuclease (88), cytochrome C (89), barnase (90) , and ribonuclease A (91). Rates of HDX are rapid in weakly stable and structurally dynamic regions that undergo local unfolding transitions, while HXD rates are typically slow in buried structures that are conformationally stable (**Figure 1.6**). In these four proteins, we found good agreement between the locations of sequence-predicted IDRs and regions with low stability as inferred by HDX rates. When testing this observation further using a database of HDX results collected from additional proteins (92–107), however, a correlation of measured HDX rates and sequence-predicted IDRs was inconclusive. We find that many proteins lack the protection factor data necessary to further test our hypothesis, therefore our calculations were limited only to a small subset of HDX proteins.

Tables.

Table 1.1 Weak multivalent protein-protein interactions that are hypothesized to drive liquid-liquid phase separation.

Type of Interaction	*Description*	*Amino Acid Type*
π-π	Non-covalent interactions occurring between two electrons rich π system	Phenylalanine (F), Tyrosine (Y), Tryptophan (W)
Cation-π	Non-covalent interaction that occurs between an electron rich π system and an adjacent cation	Phenylalanine (F), Histidine(H)*, Lysine (K), Arginine(R), Tyrosine (Y), Tryptophan (W),
Cation-Anion	Non-covalent interaction between ions of opposite charges (i.e., positive and negative)	Aspartic Acid (D), Glutamic Acid (E), Histidine (H)*, Lysine (K), Arginine(R),
Dipole-Dipole	The interaction between the positively charged end of a polar molecule and the negative end of a different polar molecule	Serine(S), Threonine (T), Tyrosine (Y), Asparagine (N), Glutamine (Q)

Figures.

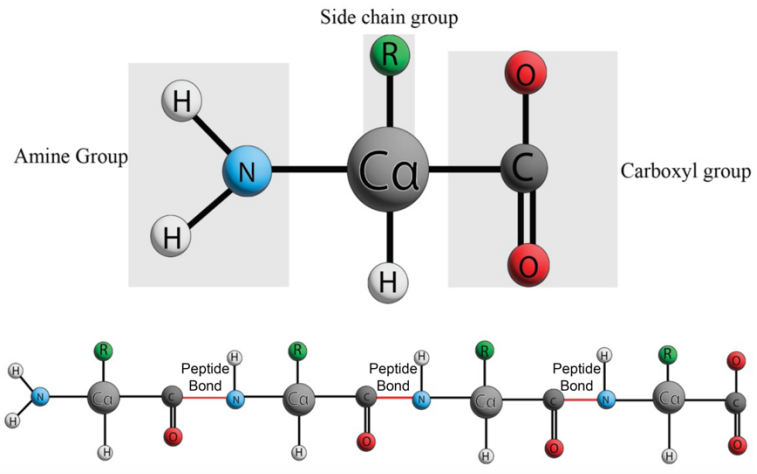

Figure 1.1 Proteins are linear polymers of amino acids linked by peptide bonds. For the 20 biologically common amino acids, each is composed of a central carbon atom, called the alpha carbon (C_α), an amine group, a carboxyl group, a side chain group, and a hydrogen atom, as shown in the top panel. A condensation reaction forms a peptide bond (shown in red) between the amine and carboxyl groups of adjacent amino acids in the polypeptide (i.e., protein) chain.

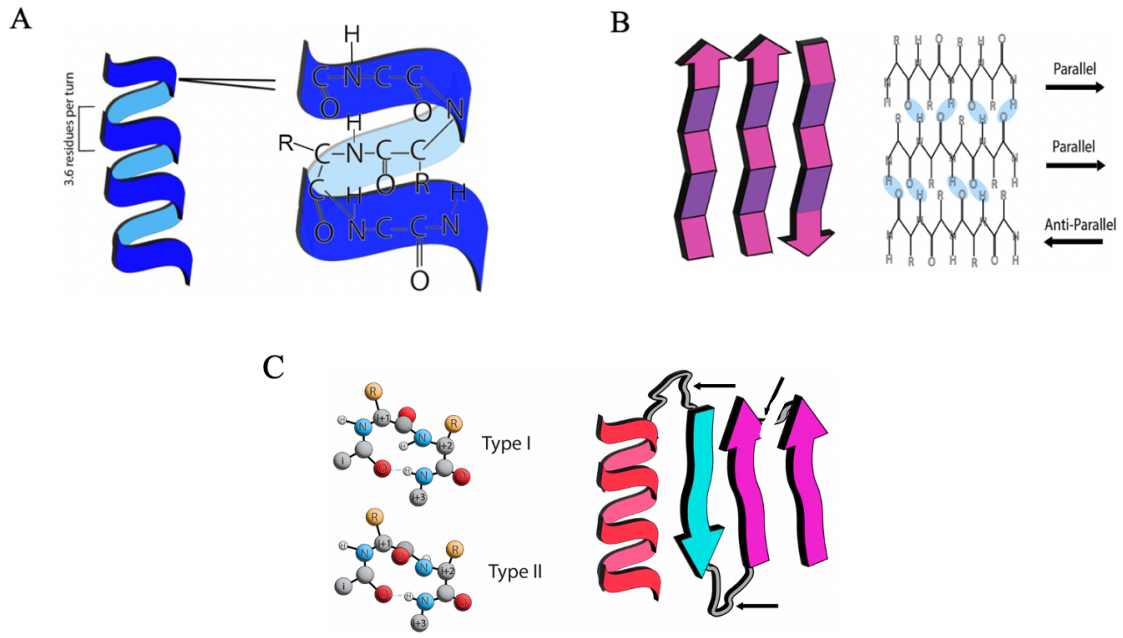

Figure 1.2 Internal hydrogen bonding defines the common secondary structure types: α-helix, β-sheet, and β-turn. Hydrogen bonding between the carboxyl groups and amine groups of amino acids allows for secondary structures to form. *A,* Cartoon representation of an α-helix, defined by approximately 3.6 residues where the hydrogen of the amine group of an amino acid residue interacts with the oxygen of the carbonyl group of a different residue that is found four residues earlier within the polypeptide chain. *B,* Individual β-strands consist of four to ten amino acid residues and form β-sheets due to the pattern of internal hydrogen bonding that occurs between individual strands. β-sheets may be parallel, anti-parallel, or a mixture of both, meaning strands may bond in the same direction (parallel) or opposite direction (anti-parallel). *C,* β-turns are non-regular secondary structures that allow for proteins to change direction nearly 180 degrees to fold back onto themselves. Turns are commonly found between other forms of secondary structure, as indicated by the black arrows. β-turns can further be classified by the torsional angle between *i+1* and *i+3*. The two most commonly occurring turn structures are Type I and Type II, where the orientation of the amide bond between *i+1* and *i+2* differ in the plane of the β-turn.

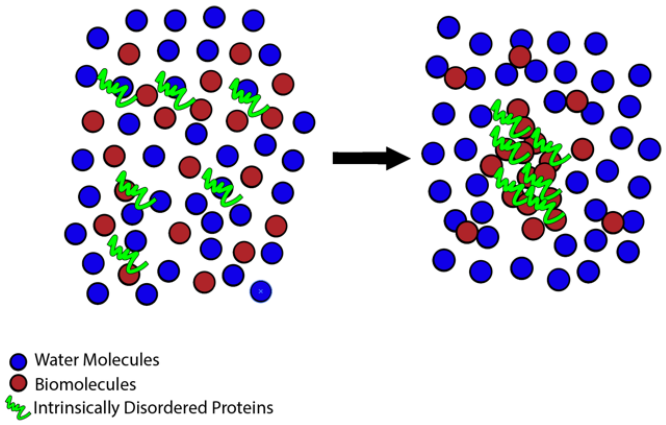

Figure 1.3 Schematic representation of the formation of membraneless organelles via liquid-liquid phase separation. Cellular contents spontaneously de-mix into liquid-like droplets consisting of a highly concentrated "dense" phase and a "dilute" phase to form membraneless organelles. IDPs are often found in the "dense" phase along with other macromolecules and are believed to be a key component in the process of liquid-liquid phase separation by serving as points of multivalent interactions.

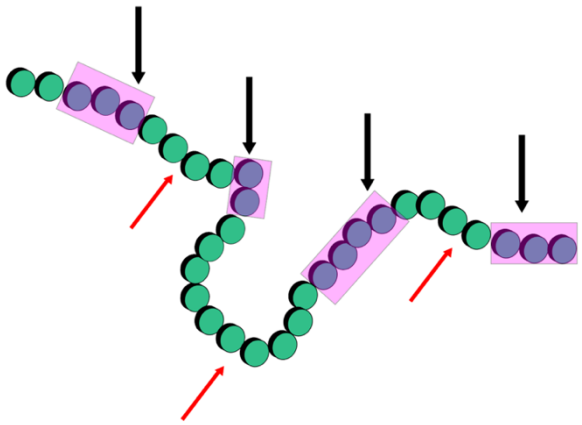

Figure 1.4 Stickers and spacers model of an IDP. "Sticker" regions, indicated by black arrows, are inter-spread within a polymer chain and serve as points of multivalent chain-chain interactions for liquid-liquid phase separation to occur. "Spacers", indicated by red arrows, are found in between "sticker" regions and limit the flexibility of the chain, they therefore modulate the ability for "sticker" regions to interact.

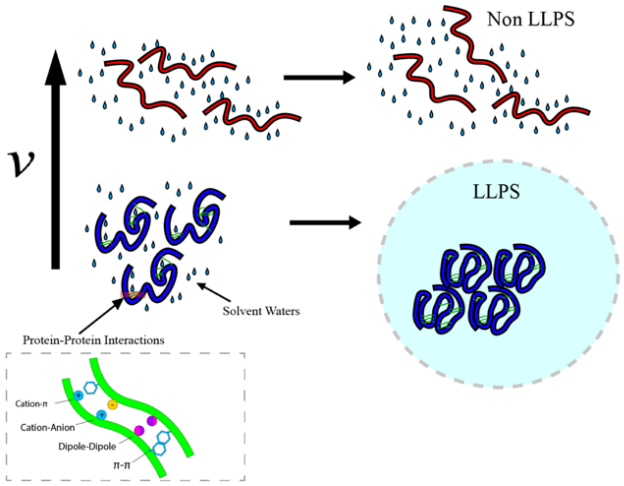

Figure 1.5 Hydrodynamic size predicts the propensity of IDRs to undergo LLPS. IDRs that prefer solvent interactions over protein-protein interactions are predicted to be more swollen and elongated when compared to IDRs that prefer self-interactions, thus displaying a larger hydrodynamic size. Proteins that undergo LLPS prefer protein-protein interactions (i.e., Cation- π, Cation-Anion, Dipole-Dipole, π- π), which would manifest in a smaller hydrodynamic size when compared to conventional IDRs.

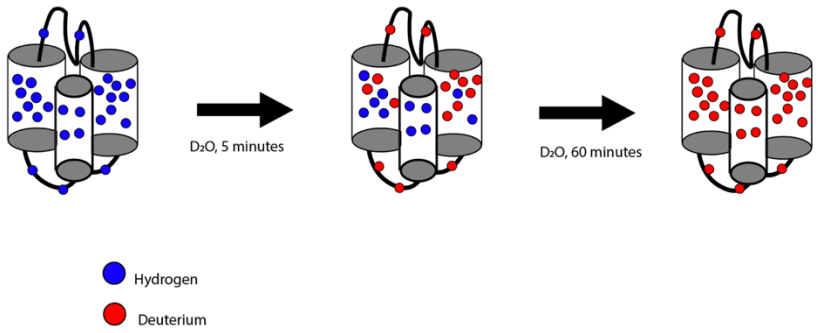

Figure 1.6 Hydrogen Deuterium Exchange provides insight to local protein structure The rate at which backbone amide hydrogen atoms exchange with deuterium atoms from the solvent is measured over time to reveal structural characteristics of proteins. IDRs lack structural stability and are more accessible to the solvent. Therefore, these regions are expected to exchange their atoms faster when compared to regions that are defined (i.e., folded).

2. METHODS

2.1 Protein databases

A set of 224 IDRs from proteins that exhibit LLPS behavior, used for the PS ID set, was obtained from Paiz *et al*. (62). For the ID set, we used the same 23 IDR sequences used previously (2, 62). The folded set started with the 82 folded sequences used previously, and then added a set of human proteins with nonhomologous structures (14), proteins with small to large structures (65), extremophile proteins (63), metamorphic proteins (66), and membrane proteins that were found by searching the PDB (92) for the phrase "membrane protein." Using the PISCES Server (93), the human, extremophile, metamorphic, and membrane proteins had a maximum of 50% sequence identity within each folded subset and only X-ray structures with a resolution better than 2.5 Å.

2.2 Calculation for ν_{model} and β-turn propensity

The propensity to form β-turn structures can be calculated by $\sum scale_i / N$, where $scale_i$ can be defined by value for amino acid type i in the normalized frequencies for β-turn structures from Levitt (15). The summation is over the number of residues, N, present in the protein sequence. ν_{model} was introduced previously (62) as a phenomenological substitute to the polymer scaling exponent (94, 95) and used to normalize protein hydrodynamic size to the chain length,

$$\nu_{model} = \log(R_h/R_o) / \log(N), \quad [1]$$

where R_o is a constant set to 2.16 Å, and the hydrodynamic radius, R_h, is calculated from sequence using an equation found to be accurate for monomeric IDPs (58, 59, 70, 96, 97).

The hydrodynamic dimensions of intrinsically disordered proteins are strongly dependent on primary sequence. The mean R_h of IDPs has been accurately predicted from intrinsic chain bias for polyproline II (PPII) (98) and overall net charge estimates (99). The equation to calculate R_h from sequence is

$$R_h = 2.16\text{Å} \cdot N^{(0.503-0.11\cdot\ln(1-f_{PPII}))} + 0.26\cdot|Q_{net}| - 0.29 \cdot N^{0.5}, \quad [2]$$

where the chain sequence N represents the number of residues, f_{PPII} is the fractional number of residues in the PPII conformation, and Q_{net} represents net charge. f_{PPII} is estimated from $\sum P_{\mathrm{PPII},i}/N$, where $f_{\mathrm{PPII},\,i}$ represents the experimentally determined value for amino acid type i in unfolded peptides (98) and the summation is over the protein sequence. Q_{net} is calculated by the number of lysine and arginine residues minus the number of glutamic acid and aspartic acid residues.

2.3 Calculating $\nu_{\mathbf{model}}$, β-turn propensity, β-sheet propensity, α-helix propensity, and hydrophobicity (ϕ) in 25-residue windows for identifying LLPS regions

Based on previous work (62, 100) , ν_{model} for a 25-residue window was calculated from sequence by first multiplying the number of each amino acid type in the window sequence by 4 and then calculating ν_{model} for the resulting 100-residue length. The β-turn propensity for a 25-residue window was calculated without modification to the method as described above and used the normalized turn frequencies from Levitt (15). The β-sheet propensity for a 25-residue window was calculated without modification to the method as described above and used the normalized turn frequencies from Qian and coworkers (20). The α-helix propensity for a 25-residue window was calculated without modification to the

method as described above and used the normalized helical frequencies by Ptitsyn-Finkelstein (21). ϕ for a 25-residue window was calculated without modification to the method as described above and used the normalized ϕ frequencies by Naderi-Manesh (22).

2.4 ParSe sequence calculations

For an input primary sequence, whereby the amino acids are restricted to the 20 common types, ParSe first reads the sequence to determine its length, *N*. Next, the algorithm uses a sliding window scheme (**Chapter 3, Figure 3.8A**) to calculate vmodel, β-turn propensity, β-sheet propensity, α-helix propensity, and ϕ for every 25-residue segment of the primary sequence. This window scheme can be applied to proteins with *N* >25. R_h is calculated by Equation 2, which in turn is used to determine ν_{model} by Equation 1, by the same method used in the original ParSe described previously (62). β-sheet is calculated as the sequence sum divided by *N* using the scale by Qian and coworkers (20). α-helix propensity is calculated as the sequence sum divided by *N* using the scale by Ptitsyn-Finkelstein (21). ϕ is calculated as the sequence sum divided by *N* using the hydrophobicity scale by Naderi-Manesh (22). The mean values of ν_{model} and each amino acid sequence property is calculated for each window (i.e., 25 residues in a sequence) and determines the localization to a PS, ID, or folded sector in a scale versus ν_{model} plot (**Chapter 3**, **Figure 3.5**). The sector boundaries are shown in (**Chapter 3, Figure 3.5**), and these boundaries are defined by the mean and standard deviation in each sequence property and ν_{model} calculated in the null set (**Chapter 3, Tables 3.5**). For example, if a window, based on its β-turn propensity and ν_{model} values, is localized to the PS sector, the central residue in that window is labeled "P," whereas localization to the ID sector labels the central residue position "D", and localization to the Folded sector labels the residue "F."

N- and C-terminal residues not belonging to a central window position are assigned the label of the central residue in the first and last window, respectively, of the whole sequence. Protein regions predicted by ParSe to be PS, ID, or Folded are determined by finding contiguous residue positions of length ≥20 that are ≥90% of only one label P, D, or F, respectively. When overlap occurs between adjacent predicted regions, owing to the up to 10% label mixing allowed, this overlap is split evenly between the two adjacent regions.

2.5 Calculating the separation, Γ, between set distributions in multiple properties

Separation, Γ, between distributions from different sets can be defined by the distance between the set means ± the standard deviations (σ). This can be represented visually by ellipses, where, for one set, the ellipse origin is defined by the means associated with two properties, while the lengths of the major and minor axes are defined by the two values of σ. The equation for such an ellipse is given

$$\frac{(x-\mu_1)^2}{\sigma_1} + \frac{(y-\mu_2)^2}{\sigma_2} = 1, \qquad [3]$$

where μ_1 and μ_2 are the two means, and σ_1 and σ_2 are the two standard deviations. To determine if the mean ± σ in two properties for any two sequence sets (or additional sequence sets) produce overlapping distributions, we simply need to calculate if the ellipses overlap or not. As shown schematically in Figure 3.3, this was computed by first finding the line that connects the origins of our two ellipses (**Chapter 3, Figure 3.3A**). The points that fall on that line and intersect the ellipses are determined (i.e., m_1, n_1, and m_2, n_2 in the figure). The point (m_1, n_1) on the first ellipse is substituted for *x* and *y* in the equation that defines the second ellipse, and *vice versa*, the point (m_2, n_2) on the second ellipse is

substituted for x and y in the equation that defines the first. If the two ellipses overlap, the resulting calculation from both substitutions will be less than one. Γ is assigned the lower of these two values. If the two ellipses are separated and do not overlap, the resulting calculation from both substitutions will be greater than one. Here, Γ is assigned the value of one. When comparing three sequence sets in two properties (i.e., v_{model} and any of the 567 property scales in the PS-IDR test, IDR null, and folded sets) Γ is given as a product,

$$\Gamma = \Gamma_{PSID-ID} \times \Gamma_{PSID-Folded} \times \Gamma_{ID-Folded} \quad [4]$$

where the subscripts represent each possible combination of sequence sets.

2.6 Start2Fold and protection factors

The Start2Fold Database contains curated proteins from past works that have been used in hydrogen deuterium experiments (23). As of August 2021, Start2Fold listed 57 proteins in the database. Of the 57 proteins, 3 proteins were NMR structures and were removed from the data set for our studies. The remaining proteins were manually curated by referencing literature containing each protein for protection factor data at individual residue positions, resulting in 18 proteins in total for our studies. The Start2Fold Database is available at: *https://www.bio2byte.be/start2fold/residues*

2.7 Amino Acid Index

The amino acid index is a database containing 566 amino acid indices representing various physicochemical and biochemical properties of the amino acids such as size, composition, charge, flexibility, hydrophobicity, aperiodicity, structure, and other characteristics (24). The structure classification was further sub-categorized by the

propensity to form specific secondary structures such as α-helix, sheet, turn, coil, and loop structures. Hydrophobicity scales were also sub-categorized based on whether the scale was derived from the analysis of a set of protein structures, for example if the amino acid type prefers buried or exposed locations, or if the scale is from solution-based measurements, such as measuring solubility limits in water *versus* organic solvents. Scales labeled "other" represent physiochemical properties corresponding to polarity, pKa values, mutability, refractivity, and other indices that do not easily bin into the "non-other" classifications defined above. The Amino Acid Index is available at: https://www.genome.jp/aaindex/

2.8 Protein structural rendering using ChimeraX

Protein models were rendered using ChimeraX for **Figure 3.11**, available for use at https://www.cgl.ucsf.edu/chimera/.

3. RESULTS AND DISCUSSION

3.1 Introduction

Liquid-liquid phase separation (LLPS) describes the reversible process where proteins and other macromolecules spontaneously de-mix from the cellular milieu to form membrane-free compartments that have been termed "membraneless organelles", or "MLOs" (43, 44, 53, 104–108). MLOs are essential to the cell and utilized to facilitate key life processes including transcription, translation, metabolism, and signaling (1, 4, 8). While it is known that MLO formation is driven primarily by proteins (43, 44, 53, 104–108), unfortunately, neither the physical mechanisms underlying protein LLPS nor the range of proteins exhibiting LLPS behavior are fully understood. Current literature suggests that LLPS can be driven by multivalent interactions occurring between protein chains such as aromatic stacking, charge-based interactions, and π-π bonding (Table 1.1) (52, 55, 67). Because these interactions are based upon intrinsic properties that can be identified from the primary sequence, a number of sequence-based methods have been developed to predict the potential for a protein to undergo LLPS from the primary sequence (109–111).

Recently, an analysis of protein databases has revealed that the polymer scaling exponent, ν, and β-turn propensity, both calculated from the sequence, can be used to identify regions within proteins that are folded, ID, or ID with high potential to undergo phase separation (62). For flexible homopolymers, ν scales with the hydrodynamic size and is used to report on the balance of self and solvent interactions (16,17,18). Small values for ν (~0.3) indicate a net preference for self-interactions, while larger values (~0.6) suggest chain-solvent interactions are preferred instead. Because proteins are heteropolymers, the

parameter ν_{model} was introduced as a phenomenological substitute to ν and used to normalize the protein hydrodynamic size to its chain length:

$$\nu_{model} = \log (R_h/R_o)/ \log(N), \qquad [3.1]$$

where N is the number of residues, R_o is a constant set to 2.16 Å, and the hydrodynamic radius, R_h, is calculated from sequence using an equation that has been found to be accurate for monomeric IDPs (58, 59, 70).

The ability of sequence calculated ν_{model} and β-turn propensity to predict regions within proteins that are ID and with the potential to drive LLPS was demonstrated using IDRs obtained from proteins found in manually curated lists for experimentally verified cases of LLPS (67). Negative control sequence sets representing IDRs from proteins not known to phase separate and folded protein regions, however, were comparably smaller and possibly less representative of proteins in general. For example, the set of folded protein regions used previously was obtained from the folded regions found within the set of known phase-separating proteins. Whether or not folded regions within proteins that exhibit phase-separating behavior are biased differently in ν_{model} and β-turn propensity when compared to folded proteins in general is not known. Moreover, the range of intrinsic sequence properties that could be used as classifiers for identifying protein regions that are folded, ID, or ID with high potential for driving LLPS also was not investigated.

Here, to address these issues, we sought to expand the folded sequence set by including folded regions from additional protein types. To the previous, original folded set, we added folded regions from extremophile proteins (63), metamorphic proteins (66),

membrane proteins, structurally non-homologous human proteins (14), and folded structures that varied in sequence length (65). Expansion of the sequence set representing IDRs from proteins not known to phase separate, and the results arising from that change, are reported elsewhere (112).

Next, we investigated the range of intrinsic sequence properties that identify phase-separating (PS) IDRs to better understand the mechanisms and protein features that possibly drive LLPS. To do this, we obtained a large list of amino acid property scales from the Amino Acid Index database which is a curated set of 566 numerical indices representing various physicochemical and biochemical properties of the amino acids, and then determined the ability of each scale when paired with ν_{model} to distinguish folded, ID, and PS-IDR populations within the sequence sets. A newly developed hydrophobicity scale designed to predict sequences that drive LLPS (86) was added to this set of intrinsic sequence properties. The reason for continuing to use ν_{model} in each pair, rather than exploring all possible combinations, was based upon the long-held interest in the field for using ν as a predictor of LLPS potential in IDRs (62, 105, 113). The results from removing this dependence on ν_{model} as a LLPS classifier are reported elsewhere (112).

Lastly, when applying our sequence-based calculations to proteins that are known to fold into globular structures, we noticed that short regions within these folded proteins are often incorrectly predicted by ν_{model} and β-turn propensity to be ID. We sought to determine if these short, incorrectly predicted IDRs could be explained by differences in stability within the folded protein structure. Specifically, we hypothesize that these incorrectly labeled ID regions map to short segments of folded structure (e.g., loops) with low stability. To test this idea, we compared our window-based calculations in β-turn

propensity and ν_{model} to experimental protection factors obtained from hydrogen-deuterium exchange (HDX) experiments, which trend quantitatively with structural stability (88, 114–116). Here, we used HDX data from four well-characterized proteins, staphylococcal nuclease (88), cytochrome C (89), barnase (117), and ribonuclease A (118), as well as curated HDX data found in the Start2Fold database (103).

3.2 Folded regions exhibit common ν_{model} and β-turn propensity characteristics

Sequences from folded protein regions were used as a control in the previous study (i.e., sequences not enriched for ID (62). However, this set was limited to folded sequences found in proteins known to exhibit phase separation behavior (67) and may not be representative of folded proteins in general. To determine this, the folded set was expanded to include: 122 human proteins with nonhomologous folded structures (14), 32 proteins with small (N=36) to large (N=415) folded structures (65), 54 folded extremophile proteins (63), 53 folded metamorphic proteins (66), and 90 folded membrane proteins. Combined, this increased the number of sequences in the folded set from 82 to 433. The combined set of sequences in the expanded folded set are listed in **Table 3.1**.

When comparing the different types of folded proteins, means in ν_{model} and β-turn propensity were, overall, similar (**Tables 3.2 and 3.3**). This was determined statistically by calculating one-tail p-values using Welch's unequal variances t-test (119), which assumes a normal distribution, and the nonparametric Mann-Whitney U-test (34, 35), which does not. For ν_{model}, the mean ± σ (standard deviation) was highly similar and with p-values >0.05 (indicating similarity in means and distributions) when the human, small-to-large, membrane, and metamorphic folded protein sets were compared to the previous folded set. Folded extremophile proteins, however, reported a mean ν_{model} (0.542 ± 0.011)

that was essentially identical to the PS-IDR set mean (0.542 ± 0.020). As such, the *p*-value obtained when comparing means in ν_{model} between the extremophile and previous folded set was <0.05. The PS-IDR test set represents a set of 224 IDRs obtained from proteins verified to exhibit phase separation behavior. The identities of the sequences in the PS-IDR test set have been published elsewhere (112).

For β-turn propensity, set means were statistically similar when the small-to-large (0.968 ± 0.027) and metamorphic (0.972 ± 0.040) sets were compared to the previous folded set (0.969 ± 0.039). However, *p*-values <0.05 were found when comparing the human (0.980 ± 0.039), extremophile (0.983 ± 0.030), and membrane proteins (0.956 ± 0.046) with the previous folded set. Despite these statistical differences between types of folded proteins, *p*-values were lowest, meaning the statistical difference was more pronounced, when the combined folded set (433 sequences) was compared to the PS-IDR test set. This can be observed visually by plotting set means in a β-turn propensity *versus* ν_{model} plot, where the different folded sets have a tight grouping compared to the PS-IDR and IDR null set means (**Figure 3.1**). Substantial overlap in mean ± σ in this plot conceptually trends with statistical similarity (i.e., *p*-values >0.05), for example when comparing the folded subsets. In contrast, separations in mean ± σ with minimal overlap (**Figure 3.1**) trends with statistical differences (i.e., *p*-values <0.05), for example when comparing the combined folded set with either of the PS-IDR test or IDR null sets (Figure 3. Though not listed in Table 3.2, the *p*-values from comparing means in ν_{model} between the combined folded and IDR null sets were $2.3E^{-05}$ and $4.0E^{-09}$ for the *t*- and *U*-test, respectively. Likewise, the *p*-values were $1.2E^{-05}$ and $3.2E^{-09}$ (*t*-test and *U*-test,

respectively) when comparing mean β-turn propensities between the combined folded and IDR null sets.

In summary, these results indicate that ν_{model} and β-turn propensity are indeed statistically different in mean value among some types of folded proteins, but the observed statistical differences are small when compared to the differences between the folded, ID, and PS-IDR classes.

3.3 Amino Acid preferences for folded, ID, and PS-ID protein regions

To assess how the different amino acid types contribute to a protein region being classified as folded, ID, or PS-IDR, we calculated ν_{model} and β-turn propensity in long homopolymers (N=100) of the common amino acids. The mean values in β-turn propensity and ν_{model} are plotted in Figure 3.2, with the plot divided into sectors as done previously (**Chapter 1, Figure 1.1**). The homopolymer results in this figure are consistent with the idea that Trp, Cys, Phe, Ile, Tyr, Val, Leu, Ala, His, Met, and Thr act as "order promoting" amino acids (i.e., favoring folding because they are located in the folded sector of the plot), while Arg, Gln, Pro, Glu, Lys, and Asp are "disorder promoting", and Asn, Ser, and Gly are "phase separation promoting". This result is similar, but not identical, to conclusion from protein database analyses that determined Trp, Cys, Phe, Ile, Tyr, Val, Leu, and Asn are enriched in folded proteins, and thus classified as "order promoting", while Ala, Arg, Gln Pro, Glu, Lys, Gly, and Ser are enriched in IDPs, and "disorder promoting", and His, Met, Thr, and Asp are "ambiguous" amino acid types, meaning they can be found in either folded or IDRs (10, 15, 41).

3.4 Intrinsic sequence properties that identify folded, ID, and PS-ID regions.

Previously, we demonstrated that two sequence-calculated properties, ν_{model} and β-turn propensity, can be used to identify folded, ID, and PS-IDRs from the primary sequence (62). To determine if other intrinsic sequence-based properties in addition to β-turn propensity can predict LLPS regions when combined with ν_{model}, we obtained 566 amino acid property scales from the Amino Acid Index database (85), as well as a newly developed hydrophobicity scale designed specifically to identify PS-IDRs (86). For each of the 567 properties, the mean ± σ in the folded, PS-IDR test, and IDR null sequence sets was calculated. We designed a numerical parameter, referred to as "separation", Γ, to find those properties that, when combined with ν_{model}, have minimally overlapping distributions among the three sequence sets, like the separated distributions that are observed when β-turn propensity is paired with ν_{model}.

The use of Γ is based on the idea that the mean ± σ in two properties for a sequence set, for example ν_{model} paired with any of the 567 property scales, can be represented visually by an ellipse. The ellipse origin is defined by the two means, while the lengths of the major and minor axes are defined by the two values of σ. The equation for such an ellipse is given by,

$$\frac{(x-\mu_1)^2}{\sigma_1} + \frac{(y-\mu_2)^2}{\sigma_2} = 1, \qquad [3.2]$$

where μ_1 and μ_2 are the two means, and σ_1 and σ_2 are the two standard deviations. To determine if the mean ± σ in two properties for any two sequence sets (or additional sequence sets) produce overlapping distributions, we simply need to calculate if the ellipses

overlap or not. As shown schematically in **Figure 3.3**, this was computed by first finding the line that connects the origins of our two ellipses. Next, the points that fall on that line and intersect the ellipses are determined (i.e., m_1, n_1, and m_2, n_2 in the figure). The point (m_1, n_1) on the first ellipse is substituted for *x* and *y* in the equation that defines the second ellipse, and *vice versa*, the point (m_2, n_2) on the second ellipse is substituted for x and y in the equation that defines the first. If the two ellipses overlap, the resulting calculation from both substitutions will be less than one. Γ is assigned the lower of these two values. If the two ellipses are separated and do not overlap, the resulting calculation from both substitutions will be greater than one. Here, Γ is assigned the value of one. When comparing three sequence sets in two properties (i.e., ν_{model} and any of the 567 property scales in the PS-IDR test, IDR null, and folded sets) Γ is given as a product,

$$\Gamma = \Gamma_{PSID-ID} \times \Gamma_{PSID-Folded} \times \Gamma_{ID-Folded}, \qquad [3.3]$$

where the subscripts represent each possible combination of sequence sets.

Using equation 3.2, Γ was calculated for ν_{model} when paired individually with each of the 567 property scales (**Figure 3.4**). To organize these results, we binned each property scale into one of eight classifications: size, composition, charge, flexibility, hydrophobicity, aperiodicity, structure, and other. The structure classification was further sub-categorized by the propensity to form specific secondary structures such as α-helix, sheet, turn, coil, and loop structures. Hydrophobicity scales were also sub-categorized based on if the scale was derived from the analysis of a set of protein structures (85), for example if the amino acid type prefers buried or exposed locations, or if the scale is from

solution-based measurements, such as measuring solubility limits in water *versus* organic solvents. Scales labeled "other" represent physiochemical properties corresponding to polarity, pKa values, mutability, refractivity, and other indices that do not easily bin into the "non-other" classifications defined above.

The largest value of Γ was 0.44, showing that all properties gave mean ± σ in the three sequence sets that overlapped to some degree when paired in a second dimension with v_{model}. The highest Γ value was from a β-sheet propensity scale by Qian and coworkers (101). To show which scales correlate with Levitt's β-turn propensity scale, the computed correlation, R^2, was used as the y-axis in **Figure 4.8A**. We found that a few property scales, especially other turn propensity scales as well as aperiodic and coil scales, exhibited strong correlations to Levitt's β-turn propensity scale. Γ computed for Levitt's β-turn propensity scale was 0.110, which was among the top 5% of scales in terms of the highest Γ values. This shows that Levitt's β-turn propensity scale outperforms most amino acid property scales in separating the three sequence sets when paired with v_{model}. Only 27 of the 567 amino acid scales produced Γ > 0.110. On average, the better performing properties also showed higher correlations to Levitt's β-turn propensity scale. Overall, it was mostly hydrophobicity scales, Φ, and secondary structure propensities (e.g., β-sheet, α-helix, turn) that gave the largest Γ values. Of the top 28 scales (including Levitt's β-turn propensity scale), 22 were based upon secondary structure propensities, 4 were based upon Φ determined from surveys of protein structures, and 2 were aperiodic conformational propensities. These 28 scales are listed in **Table 3.4**.

Next, for the β-sheet, α-helix, and Φ scales with the three largest values in Γ, we plotted the mean ± σ for each property in the sequence sets against v_{model} to determine if

separation of the dataset means visually increased relative to when ν_{model} is paired with β-turn propensity (**Figure 3.5**). Compared to **Figure 3.1**, which shows ν_{model} paired with β-turn propensity, this seems to be the case. For example, the β-sheet propensity scale from Qian and coworkers produced the largest Γ (=0.44) and clearly resulted in the smallest amount of overlap between the IDR and PS-IDR sequence sets (**Figure 3.5A**). Moreover, the Φ scale from Naderi-Manesh (102) gave complete separation between the folded and IDR sets, however some overlap is observed between the IDR and PS-IDR sets. The α-helix scale from Ptitsyn-Finkelstein (121) produced the smallest Γ amongst the top three performing scales and visually produced the largest total overlap. In summary, we find that these results are consistent with the idea that calculations of Γ can be used to identify and quantify those property scales that can discern the folded, ID, and PS-IDR sequence sets.

To assess if changing β-turn propensity for β-sheet propensity, α-helix propensity, or hydrophobicity alters how the different amino acid types contribute to a protein region being classified as folded, ID, or PS-IDR, we again calculated their values in long homopolymers (N=100) of the common amino acids. Comparing our results to our previous calculations using β-turn propensity (**Figure 3.2**), the amino acids classified as "phase separation promoting" (Asn, Ser, Gly), "disorder promoting" (Arg, Gln, Pro, Glu, Lys, and Asp), or "order promoting" (Trp, Cys, Phe, Ile, Val, Leu, Ala, and Met) remain localized to their same respective sectors (**Figure 3.6**). However, His, Tyr, and Thr were previously classified as order promoting when evaluated with β-turn propensity, and now are "phase separation promoting" with either β-sheet and α-helix indices used to define the sectors. When using hydrophobicity to define the sectors instead of β-turn propensity, only the aromatic residues Tyr and His are "phase separation promoting".

3.5 Using β-sheet propensity, α-helix propensity, or hydrophobicity paired with ν_{model} to predict phase-separating protein regions

Previously, mean ν_{model} and mean β-turn propensity in the folded, ID-null, and PS-ID test sets were used as the basis for identifying regions within proteins that match the LLPS class, and thus used as a PS IDR predictor (62). We sought to determine if the mean values of ν_{model} paired with β-sheet propensity, α-helix propensity, or Φ properties could be similarly used (i.e., in lieu of using β-turn propensity). For our initial test, β-sheet propensity, α-helix propensity, Φ, and ν_{model} were calculated for each of the known domains of the yeast prion protein, Sup35 (**Figure 3.7**). The N-terminal domain (residues 1-124) is reported to mediate the phase separation of Sup35 (122), and, when paired with ν_{model}, the sequence-calculated values of β-sheet propensity, α-helix propensity, and Φ were similar to the mean values obtained from the PS-IDR sequences, and thus consistent with this region driving phase separation (**Figure 3.7**). Similarly, β-sheet propensity, α-helix propensity, and Φ calculated for the C-terminal domain (residues 125-254) each were similar to the folded set means and predict a folded region when paired ν_{model}. The C-terminal domain of Sup35 is known to fold into a stable, globular structure (122). Likewise, the middle domain (residues 255-685) of Sup35 has calculated values of β-sheet propensity, α-helix propensity, and Φ that are similar to the mean values of the IDR null sequence set. This region of Sup35 is known to be ID and does not drive phase separation (122)

Next, to analyze proteins without predefined boundaries for the different regions, we used an algorithm that applies a 25-residue window across the whole protein sequence, in 1-residue steps (**Figure 3.8A**). For each 25-residue window, values for β-sheet

propensity, α-helix propensity, Φ, and ν_{model} are calculated and then used to label a window "F", "D", or "P" (**Figure 3.8B-E**). The window label is assigned to the central residue of the window, thus converting any primary sequence into a string of F, D, and P letters. The algorithm then identifies folded, ID, and PS-ID regions as those regions with lengths ≥ 20 residues that are at least 90% of the same letter (e.g., F, D, P). To demonstrate this scheme, **Figure 3.8** shows the results when applying our algorithm to the full sequence of Sup35. Each small dot in panel B is representative of ν_{model} and β-turn propensity for a different 25-residue window, and each value is mapped onto a β-turn propensity versus ν_{model} plot where sector boundaries for folded, ID, and PS-ID were defined by the mean and standard deviation of the IDR null sequence set.

Substituting β-sheet propensity, α-helix propensity, or Φ for β-turn propensity in this algorithm yields similar prediction results when compared across Sup35 (**Figure 3.8**) and six additional well-studied proteins that have been verified to exhibit LLPS behavior (**Figure 3.9A-F**). FUS is reported to require the entire sequence to undergo LLPS (123), including a folded the RNA recognition motif (RRM) mapping to residues 282-371 (123). The silk wrapping protein, Spidroin-1, contains repeat folded regions with intervening short PS-IDRs that promote phase separation via hydrophobic interactions (45, 46). The N-terminus (residues 1-236) of DDX4 is reported to mediate phase separation by a network of charge, hydrophobic, cation-π, and aromatic interactions, mostly from F and R residues (87, 126). The G- and R-rich N-terminus of LAF-1 is reported to contribute to RNA-protein interactions and drives phase separation (127) , while the middle domain of LAF-1 (residues 231-628) was identified as folded and represents a RecA-like DEAD box helicase domain (128). The ID C-terminus of LAF-1 (residues 648-708) is not required for *in vitro*

phase separation and may have been incorrectly predicted to be PS-ID by each algorithm. The phase separation of bacterial single-stranded DNA binding protein, SSB, is reported to be driven by the low sequence complexity ID-linker region found between the ID C-terminal domain and the highly conserved N-terminus OB-fold (52, 53). For eIGF4G2, a small segment (residues 13-97) consisting of mostly N and Q residues was reported to drive phase separation (130).

Comparing these results, overall, we find that when using β-sheet propensity and α-helix propensity, ParSe identifies more regions as PS ID. For example, when β-sheet propensity is used, the central region of FUS, which is folded, is predicted by ParSe to be mostly PS-ID and an extension of the C-terminal PS region that was predicted by the original ParSe when using β-turn propensity. Moreover, when using α-helix propensity in the ParSe prediction, Spidroin-1 no longer has any predicted folded regions located between each predicted PS IDR (which again are, on average, longer). When using Φ in the prediction, we find that our results closely resemble predictions using β-turn propensity, however Φ predictions had more mixed regions (colored white in the figure) where fewer residues were labeled one of F, D, or P in contiguous stretches at the 90% or higher threshold (**Figure 3.9C**).

In summary, β-sheet propensity, α-helix propensity, and Φ, when paired with ν_{model}, predict regions promoting phase separation behavior in proteins at a similar level (e.g., number of domains and locations), but with an overall slight increase in PS ID content and corresponding decreases in folded regions. Based on this, we conclude that ν_{model} can be paired with multiple sequence-based properties in addition to β-turn propensity to identify folded, ID, and PS-ID regions from sequence.

3.6 Long protein regions labeled "P" by ParSe are unique to proteins that undergo LLPS

Previously, when β-turn propensity and ν_{model} were used to evaluate the PS-ID test set, most of the sequences in this set contained long stretches ($N > 50$) with at least 90% of residue positions labeled "P" (62). By comparison, there were very few proteins in the human proteome (~5%) with long segments containing 90% of residue positions labeled "P" (**Figure 3.10**). We sought to determine if the rarity of long, contiguous stretches of "P" labeled positions in sequences of the human proteome persisted when β-turn propensity is substituted for β-sheet propensity, α-helix propensity, or Φ. When using Φ, β-sheet, or α-helix propensities, 62%, 86%, and 85%, respectively, of the sequences in the human proteome had regions at least one residue in length with high LLPS potential (as compared to ~70% when β-turn propensity is used). Moreover, Φ, β-sheet, or α-helix propensities identified 2%, 10%, and 12%, respectively, of the human proteome to contain regions at least 50 residues in length with potential to promote phase separation, while β-turn propensity identified ~5% of the set.

Next, as a negative control, we repeated these calculations on the ~14,000 sequences in the SCOPe (Structural Classification of Proteins extended, version 2.07) database, which represent folded proteins across families and superfamilies (131), and the ~1,500 consensus ID sequences in the DisProt database (excluding ID sequences annotated as phase separating) (54, 55). When using β-sheet propensity, α-helix propensity, or Φ, ParSe predicted less than ~10% of the DisProt database to contain long regions ($N \geq 50$) with the potential to promote phase separation. For comparison, when using β-turn propensity in ParSe, the rate is ~5%. Similarly, when using β-sheet propensity or α-helix

propensity in ParSe, ~6% of the SCOPe dataset is predicted to contain long regions ($N \geq$ 50) with the potential to promote phase separation. When using Φ in ParSe, this rate is < 1%, whereas the rate is ~0.41% when using β-turn propensity.

When using α-helix propensity, β-sheet propensity, or Φ for a set of 43 proteins that have been characterized *in vitro* and verified to undergo phase separation, our modified algorithms predicted ~97% ,~95% , and ~ 60% to contain PS regions 50 residues or longer in length, compared to ~88% when using β-turn propensity (**Figure 3.10A**). Of note, when principal component analysis (PCA) was performed on the human proteome using all intrinsic sequence properties from the Amino Acid Index, researchers found that Φ and ν_{model} yielded strongly correlated modes of variation, meaning both of these sequence-calculated properties partitioned sequences similarly. This result provides an explanation for the lower performance of ParSe when Φ is paired with ν_{model} (23).

To assess the predictive performance of our modified versions of ParSe in identifying PS-IDR, we produced a recall plot (**Figure 3.10B)** comparing the human proteome to the curated set of LLPS proteins that have been characterized *in vitro* and verified to undergo phase separation. The data in a recall plot is typically quantified by the area under the curve (AUC) when comparing the test set (LLPS *in vitro* sufficient) versus a comparison data set (Human proteome) (133–135). We find that replacing β-turn propensity with β-sheet propensity, α-helix propensity, or Φ property resulted in AUC values > ~0.99 for β-turn propensity, β-sheet propensity, α-helix propensity and ~0.90 for Φ property. Overall, our modified versions of ParSe did not marginally improve our abilities to identify PS-IDR in protein sequences.

3.7 HDX protection factor values in folded proteins trend with ParSe predicted short IDRs

When using ParSe with β-turn propensity and ν_{model} on sequences from folded proteins, we noticed short segments to be incorrectly predicted as IDRs (**Figure 3.11**). We hypothesized that regions that have been incorrectly predicted as IDRs in these folded sequences would correspond to regions of lower structural stability within the folded domains. To test this idea, we compared our window-based calculations with ParSe to protection factor (PF) data from hydrogen-deuterium exchange experiments (HDX). HDX measures the rate of exchange between backbone amide hydrogen atoms and deuterium atoms of the solvent to provide information on both the structure and the stability of a protein (4,136–138). For example, most backbone amide hydrogen atoms of well-structured globular proteins are often found at the tightly packed hydrophobic core , and thus are less accessible to the solvent to undergo exchange. For regions of proteins that lack structural stability, such as IDRs, the backbone amide hydrogen atoms are more accessible to the solvent and undergo exchange more rapidly (27, 60, 61).

Protection factors use the rate of exchange between hydrogen and deuterium atoms to quantitatively assess the structural parameters of proteins by assigning values at individual residues, where higher protection factors correspond to regions of high stability and regions that have low structural stability correspond to lower protection factors (60, 61). Because our window-based calculations can be used to identify regions of ID (i.e., residues labeled D or P) and regions of structure (i.e., residues labeled F), we compared individual F, D, and P residue positions to their respective protection factor value to determine if regions of ID corresponded to, on average, lower protection factor values. To

test this idea, we applied ParSe to four well-studied proteins in regards to hydrogen deuterium exchange experiments: staphylococcal nuclease (88), cytochrome c (89) , barnase (117), ribonuclease A (118) (**Figure 3.11**). We determined the overall structural classification of each HDX protein (i.e., folded, ID, or PS-ID) by evaluating the percent composition of each letter code in the entire sequence. We classified proteins with > 50% F residues as folded, while proteins with > 50% D or P residues as ID (**Table 3.6**). ParSe identified Staphylococcal nuclease, ribonuclease A, and barnase as folded, while cytochrome C was classified as ID. To determine if regions that were classified as ID (i.e., labeled D or P) by ParSe corresponded to lower protection factors, we evaluated the individual residue positions to determine an overall mean protection factor for structured positions (F) and unstructured positions (D and P) (**Table 3.7**). In general, among the proteins, there were very few P-labeled positions, less so than D. Overall, the sum total for protection factor was higher for positions labeled F by ParSe than positions labeled D or P, consistent with our hypothesis. For example, cytochrome C, although overall classified incorrectly as ID by ParSe, containing 53% of residues as labeled D or P, but nonetheless yielded average protection factor values of 6.3 ± 1.4 for structured regions and 5.1 ± 1.1 for ID regions. For Staphylococcal nuclease, barnase, and ribonuclease the mean protection factor values gave higher values at positions labeled F by ParSe when compared to positions labeled D or P (**Table 3.7**). Of note, when comparing mean ± σ protection factor of these proteins, we find that the mean values for positions labeled F are only slightly higher than mean values for positions labeled D or P.

3.8 Average hydrogen-deuterium exchange protection factor of F, D, and P positions in the Start2Fold database

Because our preliminary evaluation comparing ParSe output to HDX protection factors was limited to four folded proteins, we next sought to expand this analysis by using HDX data found in the Start2Fold database (103). The Start2Fold database contains position-specific HDX data from 14 proteins that were curated from the literature (**Table 3.8**). Additional proteins in the Start2Fold database that did not have position-specific data were not included in our analysis. Of the 14 proteins with position specific HDX data in Start2Fold, 12 of these contained ≥ 68% F residues, meaning most of these proteins were classified overall as folded.

Twelve of the fourteen proteins are consistent with our findings above, where lower mean protection factors corresponded to regions identified as ID by ParSe. Bovine pancreatic trypsin inhibitor (63) and Tendamistat (140) were found to have PF values higher for regions that were classified as ID. However, these proteins are abnormally short (<100 residues) and contain multiple disulfide linkages in the folded structure. Bovine pancreatic trypsin inhibitor consisted of 43% F residues and 57% ID residues and yielded an overall PF value of 1.97 ± 0.74 for folded regions and 1.99 ± 2.2 for ID regions. Tendamistat was classified as structured but yielded slightly higher average protection factor values in regions of ID (**Table 3.9**).

In summary, when proteins that have been used in hydrogen deuterium experiments are evaluated by ParSe, regions labeled “F” contain, on average, have higher protection factors when compared to regions labeled “D” or “P”. This was observed in 16 of the 18 proteins (83%) that we analyzed. However, short, disulfide rich proteins seem to show

opposite behavior, where average protein factors were slightly higher in ParSe-predicted ID positions compared to predicted folded positions.

Tables.

Table 3.1. Sequence set of folded proteins.

NONHOMOLOGOUS HUMAN PROTEIN DATASET.

NAME	N	SEQUENCE	PDB-ID
KUNITZ TYPE DOMAIN C5	58	ETDICKLPKDEGTCRDFILKWYYDPNTKSCARFW YGGCGGNENKFGSQKECEKVCAPV	1KTH
HUMAN HYPERPLASTIC DISCS PROTEIN	61	HRQALGERLYPRVQAMQPAFASKITGMLLELSPA QLLLLLASEDSLRARVDEAMELIIAHG	1I2T
INTERLEUKIN 8	68	LRCQCIKTYSKPFHPKFIKELRVIESGPHCANTEIIV KLSDGRELCLDPKENWVQRVVEKFLKRAENS	3IL8
HUMAN TRANSCRIPTION FACTOR IIF (TFIIF)	73	GPLGSGDVQVTEDAVRRYLTRKPMTTKDLLKKF QTKKTGLSSEQTVNVLAQILKRLNPERKMINDKM HFSLKE	1I27
GRANULYSIN	74	GRDYRTCLTIVQKLKKMVDKPTQRSVSNAATRV CRTGRSRWRDVCRNFMRRYQSRVIQGLVAGETA QQICEDLR	1L9L
HUMAN MONOCYTE CHEMOTACTIC PROTEIN-2	75	PDSVSIPITCCFNVINRKIPIQRLESYTRITNIQCPKE AVIFKTQRGKEVCADPKERWVRDSMKHLDQIFQ NLKP	1ESR
CYTOCHROME B5 DOMAIN	80	STHIYTKEEVSSHTSPETGIWVTLGSEVFDVTEFVD LHPGGPSKLMLAAGGPLEPFWALYAVHNQSHVR ELLAQYKIGEL	1MJ4
APOLIPOPROTEIN(A) KRINGLE IV TYPE 7	82	CYHGDGQSYRGSFSTTVTGRTCQSWSSMTPHWH QRTTEYYPNGGLTRNYCRNPDAEIRPWCYTMDPS VRWEYCNLTQCPVME	1I71
PHOSPHATIDYLINOS ITOL 3-KINASE P85-ALPHA SUBUNIT SH3 DOMAIN	83	AEGYQYRALYDYKKEREEDIDLHLGDILTVNKGS LVALGFSDGQEARPEEIGWLNGYNETTGERGDFP GTYVEYIGRKKISPP	1PHT
FIBRONECTIN TYPE III DOMAIN	90	RLDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTY GIKDVPGDRTTIDLTEDENQYSIGNLKPDTEYEVS LISRRGDMSSNPAKETFTT	1TEN
HUMAN FIBRONECTIN	91	RDLEVVAATPTSLLISWDAPAVTVRYYRITYGETG GNSPVQEFTVPGSKSTATISGLKPGVDYTITVYAV TGRGDSPASSKPISINYRTEI	1FNA
HE APAF-1 CARD	93	MDAKARNCLLQHREALEKDIKTSYIMDHMISDGF LTISEEEKVRNEPTQQQRAAMLIKMILKKDNDSY VSFYNALLHEGYKDLAALLHDGIPV	1CY5
MONOMERIC HUMAN BETA-2-MICROGLOBULIN	96	IQRTPKIQVYSRHPAENGKSNFLNCYVSGFHPSDIE VDLLKNGERIEKVEHSDLSFSKDWSFYLLYYTEFT PTEKDEYACRVNHVTLSQPKIVKWD	1LDS
N-TERMINAL DOMAIN OF THE AMYLOID PRECURSOR PROTEIN	96	LLAEPQIAMFCGRLNMHMNVQNGKWDSDPSGTK TCIDTKEGILQYCQEVYPELQITNVVEANQPVTIQ NWCKRGRKQCKTHPHFVIPYRCLVGEFV	1MWP
HUMAN PSORIASIN (S100A7) CA2+	96	SNTQAERSIIGMIDMFHKYTRRDDKIDKPSLLTMM KENFPNFLSACDKKGTNYLADVFEKKDKNEDKKI DFSEFLSLLGDIATDYHKQSHGAAPCS	2PSR

SH2 DOMAIN OF THE CSK HOMOLOGOUS KINASE CHK	97	LSLMPWFHGKISGQEAVQQLQPPEDGLFLVRESA RHPGDYVLCVSFGRDVIHYRVLHRDGHLTIDEAV FFCNLMDMVEHYSKDKGAICTKLVRPKRK	1JWO
PLECKSTRIN HOMOLOGY DOMAIN	100	PSLGTKEGYLTKQGGLVKTWKTRWFTLHRNELK YFKDQMSPEPIRILDLTECSAVQFDYSQERVNCFC LVFPFRTFYLCAKTGVEADEWIKILRWKLSQI	1FAO
BINDING PH DOMAIN OF TAPP1	104	SAVIKAGYCVKQGAVMKNWKRRYFQLDENTIGY FKSELEKEPLRVIPLKEVHKVQECKQSDIMMRDN LFEIVTTSRTFYVQADSPEEMHSWIKAVSGAIVAQ RG	1EAZ
LCK SH2	104	APEPWFFKNLSRKDAERQLLAPGNTHGSFLIRESE STAGSFCLSVRDFDQNQGEVVKHYKIRNLDNGGF YISPRITFPGLHELVRHYTNASDGLCTRLSRPCQT	1IJR
HUMAN GELSOLIN DOMAIN	104	VVQRLFQVKGRRVVRATEVPVSWESFNNGDCFIL DLGNNIHQWCGSNSNRYERLKATQVSKGIRDNER SGRARVHVSEEGTEPEAMLQVLGPKPALPAGTED TA	1KCQ
TGF-BETA TYPE II RECEPTOR LIGAND BINDING DOMAIN	105	ALCKFCDVRFSTCDNQKSCMSNCSITSICEKPQEV CVAVWRKNDENITLETVCHDPKLPYHDFILEDAA SPTCIMKEKKKPGETFFMCSCSSDECNDNIIFSEEY	1M9Z
FK506 BINDING PROTEIN FKBP MUTANT R42K/H87V	107	GVQVETISPGDGRTFPKRGQTCVVHYTGMLEDGK KFDSSRDKNKPFKFMLGKQEVIRGWEEGVAQMS VGQRAKLTISPDYAYGATGVPGIIPPHATLVFDVE LLKLE	1BKF
HUMAN THIOREDOXIN-LIKE PROTEIN	107	VGVKPVGSDPDFQPELSGAGSRLAVVKFTMRGCG PCLRIAPAFSSMSNKYPQAVFLEVDVHQCQGTAA TNNISATPTFQFFRNKVRIDQYQGADAVGLEEKIK QHLE	1GH2
CALPONIN HOMOLOGY (CH) DOMAIN	108	KSAKDALLLWCQMKTAGYPNVNIHNFTTSWRDG MAFNALIHKHRPDLIDFDKLKKSNAHYNLQNAFN LAEQHLGLTKLLDPEDISVDHPDEKSIITYVVTYY HYFSKM	1BKR
HUMAN PRION PROTEIN	108	GAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRE NMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIK QHTVTTTTKGENFTETDVKMMERVVEQMCITQY ERESQAYY	1I4M
BIKUNIN FROM THE HUMAN INTER-ALPHA-INHIBITOR COMPLEX	110	SCQLGYSAGPCMGMTSRYFYNGTSMACETFQYG GCMGNGNNFVTEKECLQTCRTVAACNLPIVRGPC RAFIQLWAFDAVKGKCVLFPYGGCQGNGNKFYS EKECREYCGV	1BIK
M2BP SCAVENGER RECEPTOR CYSTEINE-RICH DOMAIN	111	VNDGDMRLADGGATNQGRVEIFYRGQWGTVCD NLWDLTDASVVCRALGFENATQALGRAAFGQGS GPIMLDEVQCTGTEASLADCKSLGWLKSNCRHER DAGVVCTNETTL	1BY2
HUMAN CYSTATIN C; DIMERIC FORM	111	VGGPMDASVEEEGVRRALDFAVGEYNKASNDM YHSRALQVVRARKQIVAGVNYFLDVELGRTTCTK TQPNLDNCPFHDQPHLKRKAFCSFQIYAVPWQGT MTLSKSTCQDA	1G96
P14TCL1	111	CPTLGEAVTDHPDRLWAWEKFVYLDEKQHAWLP LTIEIKDRLQLRVLLRREDVVLGRPMTPTQIGPSLL PIMWQLYPDGRYRSSDSSFWRLVYHIKIDGVEDM LLELLPDD	1JSG
PKCI-SUBSTRATE ANALOG	111	DTIFGKIIRKEIPAKIIFEDDRCLAFHDISPQAPTHFL VIPKKHISQISVAEDDDESLLGHLMIVGKKCAADL GLNKGYRMVVNEGSDGGQSVYHVHLHVLGGRQ MHWPPG	1KPF
TRANSFORMING GROWTH FACTOR-BETA2	112	ALDAAYCFRNVQDNCCLRPLYIDFKRDLGWKWI HEPKGYNANFCAGACPYLWSSDTQHSRVLSLYNT INPEASASPCCVSQDLEPLTILYYIGKTPKIEQLSN MIVKSCKCS	2TGI

LAMIN A/C GLOBULAR DOMAIN	113	GSHRTSGRVAVEEVDEEGKFVRLRNKSNEDQSM GNWQIKRQNGDDPLLTYRFPPKFTLKAGQVVTIW AAGAGATHSPPTDLVWKAQNTWGCGNSLRTALI NSTGEEVAMRKLV	1IFR
THROMBOSPONDIN-1 TYPE 1	113	QDGGWSHWSPWSSCSVTCGDGVITRIRLCNSPSP QMNGKPCEGEARETKACKKDACPINGGWGPWSP WDICSVTCGGGVQKRSRLCNNPTPQFGGKDCVG DVTENQICNKQDC	1LSL
LIGANDED STEROL CARRIER PROTEIN TYPE 2 (SCP-2)	115	LQSTFVFEEIGRRLKDIGPEVVKKVNAVFEWHITK GGNIGAKWTIDLKSGSGKVYQGPAKGAADTTIILS DEDFMEVVLGKLDPQKAFFSGRLKARGNIMLSQK LQMILKDYAKL	1IKT
HUMAN FKBP25	116	PKYTKSVLKKGDKTNFPKKGDVVHCWYTGTLQD GTVFDTNIQTSAKKKKNAKPLSFKVGVGKVIRGW DEALLTMSKGEKARLEIEPEWAYGKKGQPDAKIP PNAKLTFEVELVDID	1PBK
HUMAN CD69	117	SSCSEDWVGYQRKCYFISTVKRSWTSAQNACSEH GATLAVIDSEKDMNFLKRYAGREEHWVGLKKEP GHPWKWSNGKEFNNWFNVTGSDKCVFLKNTEVS SMECEKNLYWICNKPYK	1E 87
GABA(A) RECEPTOR ASSOCIATED PROTEIN GABARAP	117	MKFVYKEEHPFEKRRSEGEKIRKKYPDRVPVIVEK APKARIGDLDKKKYLVPSDLTVGQFYFLIRKRIHL RAEDALFFFVNNVIPPTSATMGQLYQEHHEEDFFL YIAYSDESVYGL	1GNU
RIBONUCLEASE 1 DES1-7	120	AFQRQHMDSDSSPSSSSTYCNQMMRRRNMTQGR CKPVNTFVHEPLVDVQNVCFQEKVTCKNGQGNC YKSNSSMHITDCRLTNGSRYPNCAYRTSPKERHII VACEGSPYVPVHFDASVED	1E21
BTB DOMAIN FROM PLZF	121	MGMIQLQNPSHPTGLLCKANQMRLAGTLCDVVI MVDSQEFHAHRTVLACTSKMFEILFHRNSQHYTL DFLSPKTFQQILEYAYTATLQAKAEDLDDLLYAA EILEIEYLEEQCLKMLETIQ	1BUO
HUMAN ANGIOGENIN VARIANT Q117G	122	DNSRYTHFLTQHYDAKPQGRDDRYCESIMRRRGL TSPCKDINTFIHGNKRSIKAICENKNGNPHRENLRI SKSSFQVTTCKLHGGSPWPPCQYRATAGFRNVVV ACENGLPVHLDGSIFRRP	1K59
HUMAN ALPHA-LACTALBUMIN	123	KQFTKCELSQLLKDIDGYGGIALPELICTMFHTSG YDTQAIVENDESTEYGLFQISNKLWCKSSQVPQSR NICDISCDKFLDDDITDDIMCAKKILDIKGIDYWLA HKALCTEKLEQWLCEKL	1B9O
HUMAN SECRETORY PHOSPHOLIPASE A2	124	NLVNFHRMIKLTTGKEAALSYGFYGCHCGVGGR GSPKDATDRCCVTHDCCYKRLEKRGCGTKFLSYK FSNSGSRITCAKQDSCRSQLCECDKAAATCFARN KTTYNKKYQYYSNKHCRGSTPRC	1POD
CALCIUM-PHOSPHOLIPID BINDING DOMAIN	126	SSHKFTVVVLRATKVTKGAFGDMLDTPDPYVELF ISTTPDSRKRTRHFNNDINPVWNETFEFILDPNQEN VLEITLMDANYVMDETLGTATFTVSSMKVGEKK EVPFIFNQVTEMVLEMSLEVASS	1RLW
H1 SUBUNIT	128	CPVNWVEHERSCYWFSRSGKAWADADNYCRLE DAHLVVVTSWEEQKFVQHHIGPVNTWMGLHDQ NGPWKWVDGTDYETGFKNWRPEQPDDWYGHGL GGGEDCAHFTDDGRWNDDVCQRPYRWVCETEL	1DV8
FIBROBLAST GROWTH FACTOR 4 (FGF4)	128	GIKRLRRLYCNVGIGFHLQALPDGRIGGAHADTR DSLLELSPVERGVVSIFGVASRFFVAMSSKGKLYG SPFFTDECTFKEILLPNNYNAYESYKYPGMFIALG KNGKTKKGNRVSPTMKVTHFLPRL	1IJT
INTERLEUKIN-4 MUTANT E9A	129	HKCDITLQAIIKTLNSLTEQKTLCTELTVTDIFAAS KNTTEKETFCRAATVLRQFYSHHEKDTRCLGATA QQFHRHKQLIRFLKRLDRNLWGLAGLNSCPVKEA NQSTLENFLERLKTIMREKYSKCSS	1HZI

HUMAN LYSOZYME	130	KVFERCELARTLKRLGMDGYRGISLANWMCLAK WESGYNTRATNYNAGDRSTDYGIFQINSRYWCN DGKTPGAVNACHLSCSALLQDNIADAVACAKRV VRDPQGIRAWVAWRNRCQNRDVRQYVQGCGV	1JSF
HUMAN MUSCLE FATTY ACID BINDING PROTEIN:	131	VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVA SMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEF DETTADDRKVKSIVTLDGGKLVHLQKWDGQETT LVRELIDGKLILTLTHGTAVCTRTYEKE	1HMT
HUMAN COLLAGEN X NC1 TRIMER	132	MPVSAFTVILSKAYPAIGTPIPFDKILYNRQQHYDP RTGIFTCQIPGIYYFSYHVHVKGTHVWVGLYKNG TPVMYTYDEYTKGYLDQASGSAIIDLTENDQVWL QLPNAESNGLYSSEYVHSSFSGFLVAPM	1GR3
EPIDERMAL FATTY ACID BINDING PROTEIN	133	TVQQLEGRWRLVDSKGFDEYMKELGVGIALRKM GAMAKPDCIITCDGKNLTIKTESTLKTTQFSCTLG EKFEETTADGRKTQTVCNFTDGALVQHQEWDGK ESTITRKLKDGKLVVECVMNNVTCTRIYEKVE	1B56
HUMAN IGF2R DOMAIN 11	133	DCQVTNPSTGHLFDLSSLSGRAGFTAAYSEKGLV YMSICGENENCPPGVGACFGQTRISVGKANKRLR YVDQVLQLVYKDGSPCPSKSGLSYKSVISFVCRPE AGPTNRPMLISLDKQTCTLFFSWHTPLACE	1GP0
HUMAN CRBP IV	133	PADLSGTWTLLSSDNFEGYMLALGIDFATRKIAKL LKPQKVIEQNGDSFTIHTNSSLRNYFVKFKVGEEF DEDNRGLDNRKCKSLVIWDNDRLTCIQKGEKKN RGWTHWIEGDKLHLEMFCEGQVCKQTFQRA	1LPJ
EOSINOPHIL-DERIVED NEUROTOXIN	134	KPPQFTWAQWFETQHINMTSQQCTNAMQVINNY QRRCKNQNTFLLTTFANVVNVCGNPNMTCPSNK TRKNCHHSGSQVPLIHCNLTTPSPQNISNCRYAQT PANMFYIVACDNRDQRRDPPQYPVVPVHLDRII	1GQV
GALECTIN-3 CARBOHYDRATE RECOGNITION DOMAIN (CRD)	137	LIVPYNLPLPGGVVPRMLITILGTVKPNANRIALDF QRGNDVAFHFNPRFNENNRRVIVCNTKLDNNWG REERQSVFPFESGKPFKIQVLVEPDHFKVAVNDAH LLQYNHRVKKLNEISKLGISGDIDLTSASYTMI	1A3K
CELLULAR RETINOIC-ACID-BINDING PROTEINS I	137	PNFSGNWKIIRSENFEELLKVLGVNVMLRKIAVAA ASKPAVEIKQEGDTFYIKTSTTVRTTEINFKVGEEF EEQTVDGRPCKSLVKWESENKMVCEQKLLKGEG PKTSWTRELTNDGELILTMTADDVVCTRVYVRE	1CBS
C-TYPE LECTIN CARBOHYDRATE RECOGNITION DOMAIN	137	ALQTVCLKGTKVHMKCFLAFTQTKTFHEASEDCI SRGGTLSTPQTGSENDALYEYLRQSVGNEAEIWL GLNDMAAEGTWVDMTGARIAYKNWETEITAQPD GGKTENCAVLSGAANGKWFDKRCRDQLPYICQF GIV	1TN3
HUMAN PLATELET PROFILIN I	139	AGWNAYIDNLMADGTCQDAAIVGYKDSPSVWA AVPGKTFVNITPAEVGVLVGKDRSSFYVNGLTLG GQKCSVIRDSLLQDGEFSMDLRTKSTGGAPTFNV TVTKTDKTLVLLMGKEGVHGGLINKKCYEMASH LRRSQY	1FIL
MMS2	139	VKVPRNFRLLEELEEGQKGVGDGTVSWGLEDDE DMTLTRWTGMIIGPPRTNYENRIYSLKVECGPKYP EAPPSVRFVTKINMNGINNSSGMVDARSIPVLAK WQNSYSIKVVLQELRRLMMSKENMKLPQPPEGQ TYNN	1J74
HUMAN GGA1 VHS DOMAIN	139	PETLEARINRATNPLNKELDWASINGFCEQLNEDF EGPPLATRLLAHKIQSPQEWEAIQALTVLETCMKS CGKRFHDEVGKFRFLNELIKVVSPKYLGSRTSEKV KNKILELLYSWTVGLPEEVKIAEAYQMLKKQGIV	1JWF
MANNOSE BINDING PROTEIN	141	AASERKALQTEMARIKKWLTFSLGKQVGNKFFLT NGEIMTFEKVKALCVKFQASVATPRNAAENGAIQ NLIKEEAFLGITDEKTEGQFVDLTGNRLTYTNWNE GEPNNAGSDEDCVLLLKNGQWNDVPCSTSHLAV CEFPI	1HUP

CHARCOT-LEYDEN	141	SLLPVPYTEAASLSTGSTVTIKGRPLVCFLNEPYLQVDFHTEMKEESDIVFHFQVCFGRRVVMNSREYGAWKQQVESKNMPFQDGQEFELSISVLPDKYQVMVNGQSSYTFDHRIKPEAVKMVQVWRDISLTKFNVSYLKR	1LCL
FOCAL ADHESION KINASE	142	EISPPPTANLDRSNDKVYENVTGLVKAVIEMSSKIQPAPPEEYVPMVKEVGLALRTLLATVDETIPLLPASTHREIEMAQKLLNSDLGELINKMKLAQQYVMTSLQQEYKKQMLTAAHALAVDAKNLLDVIDQARLKMLGQT	1K04
PX DOMAIN FROM P40PHOX	143	AVAQQLRAESDFEQLPDDVAISANIADIEEKRGFTSHFVFVIEVKTKGGSKYLIYRRYRQFHALQSKLEERFGPDSKSSALACTLPTLPAKVYVGVKQEIAEMRIPALNAYMKSLLSLPVWVLMDEDVRIFFYQSPYDSEQVP	1H6H
CALMODULIN STRUCTURE	144	LTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTA	1CLL
CALMODULIN-LIKE PROTEIN (HCLP)	144	LTEEQVTEFKEAFSLFDKDGDGCITTRELGTVMRSLGQNPTEAELRDMMSEIDRDGNGTVDFPEFLGMMARKMKDTDNEEEIREAFRVFDKDGNGFVSAAELRHVMTRLGEKLSDEEVDEMIRAADTDGDGQVNYEEFVRVLVS	1GGZ
HUMAN LITHOSTATHINE	144	QEAQTELPQARISCPEGTNAYRSYCYYFNEDRETWVDADLYCQNMNSGNLVSVLTQAEGAFVASLIKESGTDDFNVWIGLHDPKKNRAWHWSSGSLVSYKSWGIGAPSSVNPGYCVSLTSSTGFQKWKDVPCEDKFSFVCKFKN	1QDD
HUMAN CD40 LIGAND	146	GDQNPQIAAHVISEASSKTTSVLQWAEKGYYTMSNNLVTLENGKQLTVKRQGLYYIYAQVTFCSNREASSQAPFIASLCLKSPGRFERILLRAANTHSSAKPCGQQSIHLGGVFELQPGASVFVNVTDPSQVSHGTGFTSFGLLKL	1ALY
EXTRACELLULAR CA2+-BINDING MODULE	151	PPCLDSELTEFPLRMRDWLKNVLVTLYERDEDNNLLTEKQKLRVKKIHENEKRLEAGDHPVELLARDFEKNYNMYIFPVHWQFGQLDQHPIDGYLSHTELAPLRAPLIPMEHCTTRFFETCDLDNDKYIALDEWAGCFGIKQKDIDKDLVI	1SRA
MONOMERIC HUMAN SOD MUTANT	153	ATKAVAVLKGDGPVQGIINFEQKESNGPVKVWGSIKGLTEGLHGFHVHEEEDNTAGCTSAGPHFNPLSRKHGGPKDEERHVGDLGNVTADKDGVADVSIEDSVISLSGDHSIIGRTLVVHEKADDLGKGGNEQSTKTGNAGSRLACGVIGIAQ	1MFM
B1 DOMAIN OF HUMAN NEUROPILIN-1	155	FKCMEALGMESGEIHSDQITASSQYSTNWSAERSRLNYPENGWTPGEDSYREWIQVDLGLLRFVTAVGTQGAISKETKKKYYVKTYKIDVSSNGEDWITIKEGNKPVLFQGNTNPTDVVVAVFPKPLITRFVRIKPATWETGISMRFEVYGCKIT	1KEX
HUMAN INTERLEUKIN-10	155	TQSENSCTHFPGNLPNMLRDLRDAFSRVKTFFQMKDQLDNLLLKESLLEDFKGYLGCQALSEMIQFYLEEVMPQAENQDPDIKAHVNSLGENLKTLRLRLRRCHRFLPCENKSKAVEQVKNAFNKLQEKGIYKAMSEFDIFINYIEAYMTMKIRN	2ILK
CDK INHIBITOR P19INK4D	156	RAGDRLSGAAARGDVQEVRRLLHRELVHPDALNRFGKTALQVMMFGSTAIALELLKQGASPNVQDTSGTSPVHDAARTGFLDTLKVLVEHGADVNVPDGTGALPIHLAVQEGHTAVVSFLAAESDLHRRDARGLTPLELALQRGAQDLVDILQGHM	1BD8

E-SELECTIN LECTIN/EGF DOMAINS	157	WSYNTSTEAMTYDEASAYCQQRYTHLVAIQNKE EIEYLNSILSYSPSYYWIGIRKVNNVWVWVGTQKP LTEEAKNWAPGEPNNRQKDEDCVEIYIKREKDVG MWNDERCSKKKLALCYTAACTNTSCSGHGECVE TINNYTCKCDPGFSGLKCEQIV	1G1T
FIBROBLAST GROWTH FACTOR 9 (FGF9)	157	TDLDHLKGILRRRQLYCRTGFHLEIFPNGTIQGTR KDHSRFGILEFISIAVGLVSIRGVDSGLYLGMNEK GELYGSEKLTQECVFREQFEENWYNTYSSNLYKH VDTGRRYYVALNKDGTPREGTRTKRHQKFTHFLP RPVDPDKVPELYKDILSQS	1IHK
PHOSPHOTYROSYL PHOSPHATASE	157	AEQATKSVLFVCLGNICRSPIAEAVFRKLVTDQNI SENWRVDSAATSGYEIGNPPDYRGQSCMKRHGIP MSHVARQITKEDFATFDYILCMDESNLRDLNRKS NQVKTCKAKIELLGSYDPQKQLIIEDPYYGNDSDF ETVYQQCVRCCRAFLEKAH	5PNT
HUMAN MMP-12	158	GPVWRKHYITYRINNYTPDMNREDVDYAIRKAFQ VWSNVTPLKFSKINTGMADILVVFARGAHGDFHA FDGKGGILAHAFGPGSGIGGDAHFDEDEFWTTHS GGTNLFLTAVHAIGHSLGLGHSSDPKAVMFPTYK YVDINTFRLSADDIRGIQSLYG	1JK3
PHOSPHATASE 5	159	PPADGALKRAEELKTQANDYFKAKDYENAIKFYS QAIELNPSNAIYYGNRSLAYLRTECYGYALGDAT RAIELDKKYIKGYYRRAASNMALGKFRAALRDYE TVVKVKPHDKDAKMKYQECNKIVKQKAFERAIA GDEHKRSVVDSLDIESMTIEDEYS	1A17
HUMAN COAGULATION FACTOR V	159	CSTPLGMENGKIENKQITASSFKKSWWGDYWEPF RARLNAQGRVNAWQAKANNNKQWLEIDLLKIKK ITAIITQGCKSLSSEMYVKSYTIHYSEQGVEWKPY RLKSSMVDKIFEGNTNTKGHVKNFFNPPIISRFIRV IPKTWNQSITLRLELFGCDIY	1CZT
C2 DOMAIN OF HUMAN FACTOR VIII	159	LNSCSMPLGMESKAISDAQITASSYFTNMFATWSP SKARLHLQGRSNAWRPQVNNPKEWLQVDFQKT MKVTGVTTQGVKSLLTSMYVKEFLISSSQDGHQ WTLFFQNGKVKVFQGNQDSFTPVVNCLDPPLLTR YLRIHPQSWVHQIALRMEVLGCEAQ	1D7P
APC10/DOC1 SUBUNIT OF THE HUMAN ANAPHASE-PROMOTING COMPLEX	161	ATPNKTPPGADPKQLERTGTVREIGSQAVWSLSSC KPGFGVDQLRDDNLETYWQSDGSQPHLVNIQFRR KTTVKTLCIYADYKSDESYTPSKISVRVGNNFHNL QEIRQLELVEPSGWIHVPLTDNHKKPTRTFMIQIA VLANHQNGRDTHMRQIKIYTPV	1JHJ
2-(BIPHENYL-4-SULFONYL)-1,2,3,4-TETRAHYDRO-ISOQUINOLINE-3-CARBOXYLIC ACID (D-TIC DERIVATIVE)	163	MLTPGNPKWERTNLTYRIRNYTPQLSEAEVERAIK DAFELWSVASPLIFTRISQGEADINIAFYQRDHGD NSPFDGPNGILAHAFQPGQGIGGDAHFDAEETWT NTSANYNLFLVAAHEFGHSLGLAHSSDPGALMYP NYAFRETSNYSLPQDDIDGIQAIYG	1I76
SH3-SH2 DOMAIN FRAGMENT	163	MGPSENDPNLFVALYDFVASGDNTLSITKGEKLR VLGYNHNGEWCEAQTKNGQGWVPSNYITPVNSL EKHSWYHGPVSRNAAEYLLSSGINGSFLVRESESS PGQRSISLRYEGRVYHYRINTASDGKLYVSSESRF NTLAELVHHHSTVADGLITTLHYPAP	2ABL
EMAP2/RNA-BINDING DOMAIN	164	IDVSRLDLRIGCIITARKHPDADSLYVEEVDVGEIA PRTVVSGLVNHVPLEQMQNRMVILLCNLKPAKM RGVLSQAMVMCASSPEKIEILAPPNGSVPGDRITF DAFPGEPDKELNPKKKIWEQIQPDLHTNDECVAT YKGVPFEVKGKGVCRAQTMSNSGIKL	1FL0
HUMAN CYCLOPHILIN A	164	VNPTVFFDIAVDGEPLGRVSFELFADKVPKTAENF RALSTGEKGFGYKGSCFHRIIPGFMCQGGDFTRHN GTGGKSIYGEKFEDENFILKHTGPGILSMANAGPN TNGSQFFICTAKTEWLDGKHVVFGKVKEGMNIVE AMERFGSRNGKTSKKITIADCGQLE	2CPL

P21RAS IN COMPLEX WITH GPPNHP	166	MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDP TIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRD QYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVK DSDDVPMVLVGNKCDLAARTVESRQAQDLARSY GIPYIETSAKTRQGVEDAFYTLVREIRQH	1CTQ
SMALL G PROTEIN RAP2A WITH GDP	167	MREYKVVVLGSGGVGKSALTVQFVTGTFIEKYDP TIEDFYRKEIEVDSSPSVLEILDTAGTEQFASMRDL YIKNGQGFILVYSLVNQQSFQDIKPMRDQIIRVKR YEKVPVILVGNKVDLESEREVSSSEGRALAEEWG CPFMETSAKSKTMVDELFAEIVRQMNYA	1KAO
HUMAN RAB5A	167	GNKICQFKLVLLGESAVGKSSLVLRFVKGQFHEF QESTIGAAFLTQTVCLDDTTVKFEIWDTAGQERY HSLAPMYYRGAQAAIVVYDITNEESFARAKNWV KELQRQASPNIVIALSGNKADLANKRAVDFQEAQ SYADDNSLLFMETSAKTSMNVNEIFMAIAKKL	1N6H
ENDOTHELIAL PROTEIN C RECEPTOR	170	LQRLHMLQISYFRDPYHVWYQGNASLGGHLTHV LEGPDTNTTIIQLQPLQEPESWARTQSGLQSYLLQF HGLVRLVHQERTLAFPLTIRCFLGCELPPEGSRAH VFFEVAVNGSSFVSFRPERALWQADTQVTSGVVT FTLQQLNAYNRTRYELREFLEDTCVQYVQKHI	1L8J
HUMAN FERRITIN	172	TSQVRQNYHQDSEAAINRQINLELYASYVYLSMS YYFDRDDVALKNFAKYFLHQSHEEREHAEKLMK LQNQRGGRIFLQDIQKPDCDDWESGLNAMECALH LEKNVNQSLLELHKLATDKNDPHLCDFIETHYLN EQVKAIKELGDHVTNLRKMGAPESGLAEYLFDKH TLG	2FHA
HUMAN FCGRIII	173	EDLPKAVVFLEPQWYSVLEKDSVTLKCQGAYSPE DNSTQWFHNESLISSQASSYFIDAATVNDSGEYRC QTNLSTLSDPVQLEVHIGWLLLQAPRWVFKEEDPI HLRCHSWKNTALHKVTYLQNGKDRKYFHHNSDF HIPKATLKDSGSYFCRGLVGSKNVSSETVNITITQA	1FNL
FC GAMMA RECEPTOR IIB ECTODOMAIN (CD32)	173	APPKAVLKLEPQWINVLQEDSVTLTCRGTHSPESD SIQWFHNGNLIPTHTQPSYRFKANNNDSGEYTCQT GQTSLSDPVHLTVLSEWLVLQTPHLEFQEGETIVL RCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQ ANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPA	2FCB
SERUM RETINOL BINDING PROTEIN	175	ERDCRVSSFRVKENFDKARFSGTWYAMAKKDPE GLFLQDNIVAEFSVDETGQMSATAKGRVRLLNN WDVCADMVGTFTDTEDPAKFKMKYWGVASFLQ KGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCAD SYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYR LIVHNGYCD	1RBP
GLYCOPROTEIN CD4 MUTANT	178	KKVVLGKKGDTVELTCTASQKKSIQFHWKNSNQI KILGNQGSFLTKSPSKLNDRADSRRSLWDQGNFPL IIKNLKIEDSDTYICEVEDQKEEVQLLVFGLTANSD THLLQGQSLTLTLESPPGSSPSVQCRSPRGKNIQGG KTLSVSQLELQDSGTWTCTVLQNQKKVEFKIDIV VLA	1CDY
CDC25B CATALYTIC DOMAIN	178	DHRELIGDYSKAFLLQTVDGKHQDLKYISPETMV ALLTGKFSNIVDKFVIVDCRYPYEYEGGHIKTAVN LPLERDAESFLLKSPIAPCSLDKRVILIFHCEFSSER GPRMCRFIRERDRAVNDYPSLYYPEMYILKGGYK EFFPQHPNFCEPQDYRPMNHEAFKDELKTFRLKT RSWA	1QB0
RND3/RHOE	179	VKCKIVVVGDSQCGKTALLHVFAKDCFPENYVPT VFENYTASFEIDTQRIELSLWDTSGSPYYDNVRPL SYPDSDAVLICFDISRPETLDSVLKKWKGEIQEFCP NTKMLLVGCKSDLRTDVSTLVELSNHRQTPVSYD QGANMAKQIGAATYIECSALQSENSVRDIFHVAT LACVNK	1M7B

CD11A I-DOMAIN	181	GNVDLVFLFDGSMSLQPDEFQKILDFMKDVMKK LSNTSYQFAAVQFSTSYKTEFDFSDYVKRKDPDA LLKHVKHMLLLTNTFGAINYVATEVFREELGARP DATKVLIIITDGEATDSGNIDAAKDIIRYIIGIGKHF QTKESQETLHKFASKPASEFVKILDTFEKLKDLFT ELQKKIYV	1ZON
HUMAN TISSUE INHIBITOR OF METALLOPROTEINA SE-2	182	CSCSPVHPQQAFCNADVVIRAKAVSEKEVDSGND IYGNPIKRIQYEIKQIKMFKGPEKDIEFIYTAPSSAV CGVSLDVGGKKEYLIAGKAEGDGKMHITLCDFIV PWDTLSTTQKKSLNHRYQMGCECKITRCPMIPCYI SSPDECLWMDWVTEKNINGHQAKFFACIKRSDGS CAWYRGAA	1BR9
SMALL G-PROTEIN	183	GSPQAIKCVVVGDGAVGKTCLLISYTTNAFPGEYI PTVFDNYSANVMVDGKPVNLGLWDTAGQEDYD RLRPLSYPQTDVSLICFSLVSPASFENVRAKWYPE VRHHCPNTPIILVGTKLDLRDDKDTIEKLKEKKLT PITYPQGLAMAKEIGAVKYLECSALTQRGLKTVF DEAIRAVLCPPP	1MH1
HUMAN RETINOBLASTOMA TUMOR SUPPRESSOR	185	VMNTIQQLMMILNSASDQPSENLISYFNNCTVNPK ESILKRVKDIGYIFKEKFAKAVGQGCVEIGSQRYK LGVRLYYRVMESMLKSEEERLSIQNFSKLLNDNIF HMSLLACALEVVMATYSRSTSQNLDSGTDLSFPW ILNVLNLKAFDFYKVIESFIKAEGNLTREMIKHLER CEHRIMESLA	1AD6
INTERCELLULAR ADHESION MOLECULE-1, ICAM-1	185	QTSVSPSKVILPRGGSVLVTCSTSCDQPKLLGIETP LPKKELLLPGNNRKVYELSNVQEDSQPMCYSNCP DGQSTAKTFLTVYWTPERVELAPLPSWQPVGKQL TLRCQVEGGAPRAQLTVVLLRGEKELKREPAVGE PAEVTTTVLVRRDHHGAQFSCRTELDLRPQGLEL FENTSAPYQLQTF	1IAM
DIHYDROFOLATE REDUCTASE)	186	VGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYF QRMTTTSSVEGKQNLVIMGKKTWFSIPEKNRPLK GRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQP ELANKVDMVWIVGGSSVYKEAMNHPGHLKLFVT RIMQDFESDTFFPEIDLEKYKLLPEYPGVLSDVQEE KGIKYKFEVYEKND	1KMV
P115RHOGEF RGRGS DOMAIN	190	SQFQSLEQVKRRPAHLMALLQHVALQFEPGPLLC CLHADMLGSLGPKEAKKAFLDFYHSFLEKTAVLR VPVPPNVAFELDRTRADLISEDVQRRFVQEVVQS QQVAVGRQLEDFRSKRLMGMTPWEQELAQLEA WVGRDRASYEARERHVAERLLMHLEEMQHTIST DEEKSAAVVNAIGLYMRHLGVRT	1IAP
ICAM-2	192	KVFEVHVRPKKLAVEPKGSLEVNCSTTCNQPEVG GLETSLNKILLDEQAQWKHYLVSNISHDTVLQCH FTCSGKQESMNSNVSVYQPPRQVILTLQPTLVAV GKSFTIECRVPTVEPLDSLTLFLFRGNETLHYETFG KAAPAPQEATATFNSTADREDGHRNFSCLAVLDL MSRGGNIFHKHSAPKMLEIY	1ZXQ
DEOXYRIBONUCLEOTIDASE	194	RALRVLVDMDGVLADFEGGFLRKFRARFPDQPFI ALEDRRGFWVSEQYGRLRPGLSEKAISIWESKNFF FELEPLPGAVEAVKEMASLQNTDVFICTSPIKMFK YCPYEKYAWVEKYFGPDFLEQIVLTRDKTVVSAD LLIDDRPDITGAEPTPSWEHVLFTACHNQHLQLQP PRRRLHSWADDWKAILDSKRP	1MH9
INHIBITORY RECEPTOR (P58-CL42)	195	RKPSLLAHPGPLVKSEETVILQCWSDVMFEHFLLH REGMFNDTLRLIGEHHDGVSKANFSISRMTQDLA GTYRCYGSVTHSPYQVSAPSDPLDIVIIGLYEKPSL SAQPGPTVLAGENVTLSCSSRSSYDMYHLSREGE AHERRLPAGPKVNGTFQADFPLGPATHGGTYRCF GSFHDSPYEWSKSSDPLLVSVT	1NKR
EXCHANGE FACTOR ARNO	195	ANEGSKTLQRNRKMAMGRKKFNMDPKKGIQFLV ENELLQNTPEEIARFLYKGEGLNKTAIGDYLGERE	1PBV

		ELNLAVLHAFVDLHEFTDLNLVQALRQFLWSFRL PGEAQKIDRMMEAFAQRYCLCNPGVFQSTDTCY VLSFAVIMLNTSLHNPNVRDKPGLERFVAMNRGI NEGGDLPEELLRNLYDSIRNEPFKIP	
C GELATINASE A	200	LGPVTPEICKQDIVFDGIAQIRGEIFFFKDRFIWRTV TPRDKPMGPLLVATFWPELPEKIDAVYEAPQEEK AVFFAGNEYWIYSASTLERGYPKPLTSLGLPPDVQ RVDAAFNWSKNKKTYIFAGDKFWRYNEVKKKM DPGFPKLIADAWNAIPDNLDAVVDLQGGGHSYFF KGAYYLKLENQSLKSVKFGSIKSDWLGC	1GEN
PHOSPHATIDYLCHO LINE TRANSFER	203	FSEEQFWEACAELQQPALAGADWQLLVETSGISIY RLLDKKTGLYEYKVFGVLEDCSPTLLADIYMDSD YRKQWDQYVKELYEQECNGETVVYWEVKYPFP MSNRDYVYLRQRRDLDMEGRKIHVILARSTSMPQ LGERSGVIRVKQYKQSLAIESDGKKGSKVFMYYF DNPGGQIPSWLINWAAKNGVPNFLKDMARACQN Y	1LN1
BILIVERDIN IX BETA REDUCTASE	205	MAVKKIAIFGATGQTGLTTLAQAVQAGYEVTVLV RDSSRLPSEGPRPAHVVVGDVLQAADVDKTVAG QDAVIVLLGTRNDLSPTTVMSEGARNIVAAMKAH GVDKVVACTSAFLLWDPTKVPPRLQAVTDDHIR MHKVLRESGLKYVAVMPPHIGDQPLTGAYTVTL DGRGPSRVISKHDLGHFMLRCLTTDEYDGHSTYP SHQY	1HDO
MADCAM-1	206	VKPLQVEPPEPVVAVALGASRQLTCRLACADRGA SVQWRGLDTSLGAVQSDTGRSVLTVRNASLSAA GTRVCVGSCGGRTFQHTVQLLVYAFPNQLTVSPA ALVPGDPEVACTAHKVTPVDPNALSFSLLVGGQE LEGAQALGPEVQEEEEEPQGDEDVLFRVTERWRL PPLGTPVPPALYCQATMRLPGLELSHRQAIPVLIEG R	1GSM
IGE-FC CEPSILON3-CEPSILON4	208	VSAYLSRPSPFDLFIRKSPTITCLVVDLAPSKGTVN LTWSRASGKPVNHSTRKEEKQRNGTLTVTSTLPV GTRDWIEGETYQCRVTHPHLPRALMRSTTKTSGP RAAPEVYAFATPEWPGSRDKRTLACLIQNFMPEDI SVQWLHNEVQLPDARHSTTQPRKTKGSGFFVFSR LEVTRAEWEQKDEFICRAVHEAASPSQTVQRAVS V	1FP5
GLUTATHIONE TRANSFERASE	208	KPILYSYFRSSCSWRVRIALALKGIDYKTVPINLIK DGGQQFSKDFQALNPMKQVPTLKIDGITIHQSLAII EYLEETRPTPRLLPQDPKKRASVRMISDLIAGGIQP LQNLSVLKQVGEEMQLTWAQNAITCGFNALEQIL QSTAGIYCVGDEVTMADLCLVPQVANAERFKVD LTPYPTISSINKRLLVLEAFQVSHPCRQPDTPT	1FW1
CCG1/TAFII250-INTERACTING FACTOR B	208	AASVEQREGTIQVQGQALFFREALPGSGQARFSVL LLHGIRFSSETWQNLGTLHRLAQAGYRAVAIDLP GLGHSKEAAAPAPIGELAPGSFLAAVVDALELGPP VVISPSLSGMYSLPFLTAPGSQLPGFVPVAPICTDK INAANYASVKTPALIVYGDQDPMGQTSFEHLKQL PNHRVLIMKGAGHPCYLDKPEEWHTGLLDFLQGL	1IMJ
HSP90 N-TERMINAL DOMAIN BOUND TO ADP-MG	213	PMEEEEVETFAFQAEIAQLMSLIINTFYSNKEIFLR ELISNSSDALDKIRYETLTDPSKLDSGKELHINLIPN KQDRTLTIVDTGIGMTKADLINNLGTIAKSGTKAF MEALQAGADISMIGQFGVGFYSAYLVAEKVTVIT KHNDDEQYAWESSAGGSFTVRTDTGEPMGRGTK VILHLKEDQTEYLEERRIKEIVKKHSQFIGYPITLFV E	1BYQ
CYS25SER MUTANT	217	LPDSVDWREKGCVTEVKYQGSCGASWAFSAVGA LEAQLKLKTGKLVSLSAQNLVDCSTEKYGNKGC NGGFMTTAFQYIIDNKGIDSDASYPYKAMDLKCQ YDSKYRAATCSKYTELPYGREDVLKEAVANKGP VSVGVDARHPSFFLYRSGVYYEPSCTQNVNHGVL	1GLO

		VVGYGDLNGKEYWLVKNSWGHNFGEEGYIRMA RNKGNHCGIASFPSYPEI	
MU GLUTATHIONE TRANSFERASE GSTM2-2	217	PMTLGYWNIRGLAHSIRLLLEYTDSSYEEKKYTM GDAPDYDRSQWLNEKFKLGLDFPNLPYLIDGTHK ITQSNAILRYIARKHNLCGESEKEQIREDILENQFM DSRMQLAKLCYDPDFEKLKPEYLQALPEMLKLYS QFLGKQPWFLGDKITFVDFIAYDVLERNQVFEPSC LDAFPNLKDFISRFEGLEKISAYMKSSRFLPRPVFT KMAVFGNK	1HNA
PROTEIN L-ISOASPARTATE O-METHYLTRANSFERASE	224	WKSGGASHSELIHNLRKNGIIKTDKVFEVMLATD RSHYAKCNPYMDSPQSIGFQATISAPHMHAYALE LLFDQLHEGAKALDVGSGSGILTACFARMVGCTG KVIGIDHIKELVDDSVNNVRKDDPTLLSSGRVQLV VGDGRMGYAEEAPYDAIHVGAAAPVVPQALIDQ LKPGGRLILPVGPAGGNQMLEQYDKLQDGSIKMK PLMGVIYVPLTDKEKQWSRW	1I1N
ANKYRIN REPEAT DOMAIN OF BCL-3	228	EDGDTPLHIAVVQGNLPAVHRLVNLFQQGGRELD IYNNLRQTPLHLAVITTLPSVVRLLVTAGASPMAL DRHGQTAAHLACEHRSPTCLRALLDSAAPGTLDL EARNYDGLTALHVAVNTECQETVQLLLERGADID AVDIKSGRSPLIHAVENNSLSMVQLLLQHGANVN AQMYSGSSALHSASGRGLLPLVRTLVRSGADSSL KNCHNDTPLMVARSRRVIDILRG	1K1B
DIHYDROPTERIDINE REDUCTASE	236	EARRVLVYGGRGALGSRCVQAFRARNWWVASV DVVENEEASASIIVKMTDSFTEQADQVTAEVGKL LGEEKVDAILCVAGGWAGGNAKSKSLFKNCDLM WKQSIWTSTISSHLATKHLKEGGLLTLAGAKAAL DGTPGMIGYGMAKGAVHQLCQSLAGKNSGMPPG AAAIAVLPVTLDTPMNRKSMPEADFSSWTPLEFL VETFHDWITGKNRPSSGSLIQVVTTEGRTELTPAY F	1HDR
MUTANT ESTROGEN NUCLEAR RECEPTOR	248	NSLALSLTADQMVSALLDAEPPILYSEYDPTRPFS EASMMGLLTNLADRELVHMINWAKRVPGFVDLT LHDQVHLLESAWLEILMIGLVWRSMEHPGKLLFA PNLLLDRNQGKSVEGMVEIFDMLLATSSRFRMMN LQGEEFVCLKSIILLNSGVYTFLSSTLKSLEEKDHI HRVLDKITDTLIHLMAKAGLTLQQQHQRLAQLLL ILSHIRHMSNKGMEHLYSMKSKNVVPLYDLLLEM LDAHRLHA	1QKT
ALPHA-ACTININ	248	GSSNEIRRLERLEHLAEKFRQKASTHETWAYGKE QILLQKDYESASLTEVRALLRKHEAFESDLAAHQ DRVEQIAAIAQELNELDYHDAVNVNDRCQKICDQ WDRLGTLTQKRREALERMEKLLETIDQLHLEFAK RAAPFNNWMEGAMEDLQDMFIVHSIEEIQSLITAH EQFKATLPEADGERQSIMAIQNEVEKVIQSYNIRIS SSNPYSTVTMDELRTKWDKVKQLVPIRDQSLQEE LARQHAN	1QUU
RECOMBINANT HUMAN GAMMA-FIBRINOGEN CARBOXYL	249	QIHDITGKDCQDIANKGAKQSGLYFIKPLKANQQF LVYCEIDGSGNGWTVFQKRLDGSVDFKKNWIQY KEGFGHLSPTGTTEFWLGNEKIHLISTQSAIPYALR VELEDWNGRTSTADYAMFKVGPEADKYRLTYAY FAGGDAGDAFDGFDFGDDPSDKFFTSHNGMQFST WDNDNDKFEGNCAEQDGSGWWMNKCHAGHLN GVYYQGGTYSKASTPNGYDNGIIWATWKTRWYS MKKTTMKIIPFNRL	3FIB

SMALL TO LARGE PROTEIN DATA SET.

NAME	N	SEQUENCE	PDB-ID
ANGIOTENSIN II	8	DRVYIHPF	1N9V
CHICKEN VILLIN HEADPICE	36	MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEK GLF	1VII
PKC CYS2 DOMAIN	50	HRFKVYNYMSPTFCDHCGSLLWGLVKQGLKCED CGMNVHHKCREKVANLC	1PTQ
PROTEIN G	56	MTYKLILNGKTLKGETTTEAVDAATAEKVFKQY ANDNGVDGEWTYDDATKTFTVTE	2GB1
FYN SH3	59	VTLFVALYDYEARTEDDLSFHKGEKFQILNSSEGD WWEARSLTTGETGYIPSNYVAPVD	1SHF
CSPB	67	MLEGKVKWFNSEKGFGFIEVEGQDDVFVHFSAIQ GEGFKTLEEGQAVSFEIVEGNRGPQAANVTKEA	1CSP
UBIQUITIN	76	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGI PPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVL RLRGG	1UBQ
REPRESSOR	87	PLTQEQLEDARRLKAIYEKKKNELGLSQESVADK MGMGQSGVGALFNGINALNAYNAALLAKILKVS VEEFSPSIAREIYEMYEAVS	1LMB
BARSTAR	89	KKAVINGEQIRSISDLHQTLKKELALPEYYGENLD ALWDCLTGWVEYPLVLEWRQFEQSKQLTENGAE SVLQVFREAKAEGADITIILS	1A19
CTACP	98	AEGDTLISVDYEIFGKVQGVFFRKYTQAEGKKLG LVGWVQNTDQGTVQGQLQGPASKVRHMQEWLE TKGSPKSHIDRASFHNEKVIVKLDYTDFQIVK	2ACY
PLASTOCYANIN	99	IDVLLGADDGSLAFVPSEFSISPGEKIVFKNNAGFP HNIVFDEDSIPSGVDASKISMSEEDLLNAKGETFEV ALSNKGEYSFYCSPHQGAGMVGKVTVN	2PCY
HORSE CYTOCHROME C	104	GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLH GLFGRKTGQAPGFTYTDANKNKGITWKEETLME YLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKA TNE	1HRC
PI3K SH2 (RAT)	111	GMNNNMSLQDAEWYWGDISREEVNEKLRDTAD GTFLVRDASTKMHGDYTLTLRKGGNNKSIKIFHR DGKYGFSDPLTFNSVVELINHYRNESLAQYNPKL DVKLLYPVSKY	1FU6
MYOHEMERYTHRIN	113	GFPIPDPYCWDISFRTFYTIIDDEHKTLFNGILLLSQ ADNADHLNELRRCTGKHFLNEQQLMQASQYAGY AEHKKAHDDFIHKLDTWDGDVTYAKNWLVNHIK TIDFKYRGKI	2HMQ
BOVINE-LACTALBUMIN	122	EQLTKCEVFRELKDLKGYGGVSLPEWVCTTFHTS GYDTQAIVQNNDSTEYGLFQINNKIWCKDDQNPH SSNICNISCDKFLDDDLTDDIMCVKKILDKVGINY WLAHKALCSEKLDQWLCEK	1F6S
BOVINE RIBONUCLEASE A	124	KETAAAKFERQHMDSSTSAASSSNYCNQMMKSR NLTKDRCKPVNTFVHESLADVQAVCSQKNVACK NGQTNCYQSYSTMSITDCRETGSSKYPNCAYKTT QANKHIIVACEGNPYVPVHFDASV	1XPT
CHEY	128	ADKELKFLVVDKFSTMRRIVRNLLKELGFNNVEE AEDGVDALNKLQAGGYGFVISDWNMPNMDGLE LLKTIRADGAMSALPVLMVTAEAKKENIIAAAQA GASGYVVKPFTAATLEEKLNKIFEKLGM	1EHC
LYSOZYME	129	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAA KFESNFNTQATNRNTDGSTDYGILQINSRWWCND GRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDG NGMNAWVAWRNRCKGTDVQAWIRGCRL	1HEL

INTESTINAL FA BINDING PROTEIN	131	AFDGTWKVDRNENYEKFMEKMGINVVKRKLGA HDNLKLTITQEGNKFTVKESSNFRNIDVVFELGVD FAYSLADGTELTGTWTMEGNKLVGKFKRVDNGK ELIAVREISGNELIQTYTYEGVEAKRIFKKE	1IFB
STAPHYLOCOCCAL NUCLEASE	141	ATSTKKLHKEPATLIKAIDGDTVKLMYKGQPMTF RLLLVDTPETKHPKKGVEKYGPEASAFTKKMVEN AKKIEVEFNKGQRTDKYGRGLAYIYADGKMVNE ALVRQGLAKVAYVYKPNNTHEQHLRKSEAQAKK EKLNIWS	2SNS
CALMODULIN	143	LTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRS LGQNPTEAELQDMINEVDADGNGTIDFPEFLTMM ARKMKDTDSEEEIREAFRVFDKDGNGYISAAELR HVMTNLGEKLTDEEVDEMIREADIDGDGQVNYE EFVQMMT	1CM1
MYOGLOBIN	153	VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLF KSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVT VLTALGAILKKKGHHEAELKPLAQSHATKHKIPIK YLEFISEAIIHVLHSRHPGDFGADAQGAMNKALEL FRKDIAAKYKELGYQG	1MBO
RIBONUCLEASE H	155	MLKQVEIFTDGSCLGNPGPGGYGAILRYRGREKT FSAGYTRTTNNRMELMAAIVALEALKEHCEVILS TDSQYVRQGITQWIHNWKKRGWKTADKKPVKN VDLWQRLDAALGQHQIKWEWVKGHAGHPENER CDELARAAAMNPTLEDTGYQVEV	2RN2
ASV INTEGRASE CORE	162	PLREPRGLGPLQIWQTDFTLEPRMAPRSWLAVTV DTASSAIVVTQHGRVTSVAAQHHWATAIAVLGRP KAIKTDNGSCFTSKSTREWLARWGIAHTTGIPGNS QGQAMVERANRLLKDKIRVLAEGDGFMKRIPTSK QGELLAKAMYALNHFERGENTKTNL	1ASU
T4 PHAGE LYSOZYME	164	MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKS PSLNAAKSELDKAIGRNCNGVITKDEAEKLFNQD VDAAVRGILRNAKLKPVYDSLDAVRRCALINMVF QMGETGVAGFTNSLRMLQQKRWDEAAVNLAKS RWYNQTPNRAKRVITTFRTGTWDAYKNL	2LZM
DHFR	192	MLKPNVAIIVAALKPALGIGYKGKMPWRLRKEIR YFKDVTTRTTKPNTRNAVIMGRKTWESIPQKFRPL PDRLNIILSRSYENEIIDDNIIHASSIESSLNLVSDVE RVFIIGGAEIYNELINNSLVSHLLITEIEHPSPESIEM DTFLKFPLESWTKQPKSELQKFVGDTVLEDDIKEG DFTYNYTLWTRK	1AI9
MUTY CATALYIC DOMAIN	225	MQASQFSAQVLDWYDKYGRKTLPWQIDKTPYKV WLSEVMLQQTQVATVIPYFERFMARFPTVTDLAN APLDEVLHLWTGLGYYARARNLHKAAQQVATLH GGKFPETFEEVAALPGVGRSTAGAILSLSLGKHFPI LNGNVKRVLARCYAVSGWPGKKEVENKLWSLSE QVTPAVGVERFNQAMMDLGAMICTRSKPKCSLC PLQNGCIAAANNSWALYPGKKPK	1MUN
TRIOSEPHOSPHATE, ISOMERASE	249	SKPQPIAAANWKCNGSQQSLSELIDLFNSTSINHD VQCVVASTFVHLAMTKERLSHPKFVIAAQNAIAK SGAFTGEVSLPILKDFGVNWIVLGHSERRAYYGET NEIVADKVAAAVASGFMVIACIGETLQERESGRT AVVVLTQIAAIAKKLKKADWAKVVIAYEPVWAI GTGKVATPQQAQEAHALIRSWVSSKIGADVAGEL RILYGGSVNGKNARTLYQQRDVNGFLVGGASLKP EFVDIIKATQ	5TIM
HUMAN GLYOXASE II	260	MKVEVLPALTDNYMYLVIDDETKEAAIVDPVQPQ KVVDAARKHGVKLTTVLTTHHHWDHAGGNEKL VKLESGLKVYGGDDRIGALTHKITHLSTLQVGSL NVKCLATPCHTSGHICYFVSKPGGSEPPAVFTGDT LFVAGCGKFYEGTADEMCKALLEVLGRLPPDTRV YCGHEYTINNLKFARHVEPGNAAIREKLAWAKEK	1QH3

		YSIGEPTVPSTLAEEFTYNPFMRVREKTVQQHAGE TDPVTTMRAVRREKDQFKMPRD	
ECORI ENDONUCLEASE	261	SQGVIGIFGDYAKAHDLAVGEVSKLVKKALSNEY PQLSFRYRDSIKKTEINEALKKIDPDLGGTLFVSNS SIKPDGGIVEVKDDYGEWRVVLVAEAKHQGKDII NIRNGLLVGKRGDQDLMAAGNAIERSHKNISEIA NFMLSESHFPYVLFLEGSNFLTENISITRPDGRVVN LEYNSGILNRLDRLTAANYGMPINSNLCINKFVNH KDKSIMLQAASIYTQGDGREWDSKIMFEIMFDIST TSLRVLGRDLFEQLTSK	1ERI
UDP-GALACTOSE 4-EPIMERASE	338	MRVLVTGGSGYIGSHTCVQLLQNGHDVIILDNLC NSKRSVLPVIERLGGKHPTFVEGDIRNEALMTEIL HDHAIDTVIHFAGLKAVGESVQKPLEYYDNNVNG TLRLISAMRAANVKNFIFSSSATVYGDNPKIPYVE SFPTGTPQSPYGKSKLMVEQILTDLQKAQPDWSIA LLRYFNPVGAHPSGDMGEDPQGIPNNLMPYIAQV AVGRRDSLAIFGNDYPTEDGTGVRDYIHVMDLAD GHVVAMEKLANKPGVHIYNLGAGVGNSVLDVV NAFSKACGKPVNYHFAPRREGDLPAYWADASKA DRELNWRVTRTLDEMAQDTWHWQSRHPQGYPD	1NAH
CREATINE KINASE	379	AASERRRLYPPSAEYPDLRKHNNCMASHLTPAVY ARLCDKTTPTGWTLDQCIQTGVDNPGHPFIKTVG MVAGDEETYEVFADLFDPVIQERHNGYDPRTMK HTTDLDASKIRSGYFDERYVLSSRVRTGRSIRGLS LPPACTRAERREVERVVVDALSGLKGDLAGRYYR LSEMTEAEQQQLIDDHFLFDKPVSPLLTAAGMAR DWPDARGIWHNNEKSFLIWVNEEDHTRVISMEKG GNMKRVFERFCRGLKEVERLIQERGWEFMWNER LGYILTCPSNLGTGLRAGVHIKLPLLSKDSRFPKIL ENLRLQKRGTGGVDTAATGGVFDISNLDRLGKSE VELVQLVIDGVNYLIDCERRLERGQDIRIPTPVIHT KH	1QK1
YEAST PGK	415	SLSSKLSVQDLDLKDKRVFIRVDFNVPLDGKKITS NQRIVAALPTIKYVLEHHPRYVVLASHLGRPNGE RNEKYSLAPVAKELQSLLGKDVTFLNDCVGPEVE AAVKASAPGSVILLENLRYHIEEEGSRKVDGQKV KASKEDVQKFRHELSSLADVYINDAFGTAHRAHS SMVGFDLPQRAAGFLLEKELKYFGKALENPTRPF LAILGGAKVADKIQLIDNLLDKVDSIIIGGGMAFTF KKVLENTEIGDSIFDKAVGPEIAKLMEKAKAKGV EVVLPVDFIIADAFSASANTKTVTDKEGIPAGWQG LDNGPESRKLFAATVAKATVILWNGPPGVFEFEK FAAGTKALLDEVVKSSAAGNTVIIGGGDTATVAK KYGVTDKISHVSTGGGASLELLEGKELPGVAFLSE KK	3PGK

EXTREMOPHILE PROTIEN DATA SET.

Name	N	Sequence	PDB ID
DODECIN	67	VFKKVLLTGTSEESFTAAADDAIDRAEDTLDNVVWAE VVDQGVEIGAVEERTYQTEVQVAFELDGSQ	1MOG
OXIDIZED HIGH-POTENTIAL IRON-SULFUR PROTEIN	71	MERLSEDDPAAQALEYRHDASSVQHPAYEEGQTCLN CLLYTDASAQDWGPCSVFPGKLVSANGWCTAWVAR	1HPI
HIGH-POTENTIAL IRON-SULFUR PROTEIN	85	SAPANAVAADNATAIALKYNQDATKSERVAAARPGL PPEEQQCANCQFMQADAAGATDEWKGCQLFPGKLIN VNGWCASWTLKAG	1B0Y
ACYLPHOSPHATASE	90	AIVRAHLKIYGRVQGVGFRWSMQREARKLGVNGWV RNLPDGSVEAVLEGDEERVEALIGWAHQGPPLARVTR VEVKWEQPKGEKGFRIVG	1V3Z
LYSOZYME 1	122	KTFTRCSLAREMYALGVPKSELPQWTCIAEHESSYRTN VVGPTNSNGSNDYGIFQINNYYWCQPSNGRFSYNECH LSCDALLTDNISNSVTCARKIKSQQGWTAWSTWKYCS GSLPSINDCF	2FBD
FIBROBLAST GROWTH FACTOR 2	125	DPKRLYCKNGGFFLRIHPDGRVDGVREKSDPHIKLQL QAEERGVVSIKGVSANRYLAMKEDGRLLASKSVTDEC FFFERLESNNYNTYRSRKYTSWYVALKRTGQYKLGSK TGPGQKAILFLPMS	1BAS
FERREDOXIN-1	128	PTVEYLNYEVVDDNGWDMYDDDVFGEASDMDLDDE DYGSLEVNEGEYILEAAEAQGYDWPFSCRAGACANC AAIVLEGDIDMDMQQILSDEEVEDKNVRLTCIGSPDAD EVKIVYNAKHLDYLQNRVI	1DOI
NUCLEOSIDE DIPHOSPHATE KINASE	155	HDERTFVMVKPDGVQRGLIGDIVTRLETKGLKMVGG KFMRIDEELAHEHYAEHEDKPFFDGLVSFITSGPVFAM VWEGADATRQVRQLMGATDAQDAAPGTIRGDYGND LGHNLIHGSDHEDEGANEREIALFFDDDELVDWDRDA SAWVYEDLA	2AZ1
PEROXIREDOXIN 5	156	GSPIKVGDIIPDVLVYEDVPSKSFPIHDVFRGRKGILFSV VGAFVPGSNNHIPEYLSLYDKFKEEGYHTIACIAVNDP FVMAAWGKTVDPEHKIRMLADMHGEFTRALGTELDS SKMLGNNRSRRYAMLIDDNKIRSVSTEPDITGLACLLSI QRQ	2XHF
SECRETED CHORISMATE MUTASE	165	GTSQLAELVDAAAERLEVADPVAAFKWRAQLPIEDSG RVEQQLAKLGEDARSQHIDPDYVTRVFDDQIRATEAIE YSRFSDWKLNPASAPPEPPDLSASRSAIDSLNNRMLSQI WSHWSLLSAPSCAAQLDRAKRDIVRSRHLDSLYQRAL TTATQSYCQALPPA	2AO2
ENDOGLUCANASE	181	VVHDPKGEAVLPSVFEDGTRQGWDWAGESGVKTALT IEEANGSNALSWEFGYPEVKPSDNWATAPRLDFWKSD LVRGENDYVTFDFYLDPVRATEGANINLVFQPPTNGY WVQAPKTYTINFDELEEANQVNGLYHYEVKINVRDIT NIQDDTLLRNIIFADVESDFAGRVFVDNVRFEGA	1UW
IRON SUPEROXIDE DISMUTASE	192	AFELPSLPYAIDALEPHISKETLEFHHGKHHNTYVVKL NGLIPGTKFENKSLEEIVCSSDGGVFNNAAQIWNHTFY WNSLSPNGGGAPTGAVADAINAKWGSFDAFKEALND KAVNNFGSSWTWLVKLADGSLDIVNTSNAATPLTDD GVTPILTVDLWEHAYYIDYRNVRPDYLKGFWSLVNW EFANANFA	3LIO
CHEMOTAXIS PROTEIN CHEC	200	PLLIDIRKLTLITRLIQDGAEQVADSLATLAGVDAAVEI KSLSFVQPEDIATEGGGTIYSARVRLTEPPYGVFLTFET ETAAEIAELTGSSVEDGFTQLHESALQECNILTSGFIDGI ANTLNATINGTPTVVQDDATEIADKALSHVRRDSLTIV LDSLVDIKESDVAFSLRIFLIPDPGSFVHLIDQLDYDTD RETHI	3QTA

ACETYLXYLAN ESTERASE 2	207	SCPAIHVFGARETTASPGYGSSSTVVNGVLSAYPGSTA EAINYPACGGQSSCGGASYSSSVAQGIAAVASAVNSFN SQCPSTKIVLVGYSQGGEIMDVALCGGGDPNQGYTNT AVQLSSSAVNMVKAAIFMGDPMFRAGLSYEVGTCAA GGFDQRPAGFSCPSAAKIKSYCDASDPYCCNGSNAAT HQGYGSEYGSQALAFVKSKLG	1BS9
FE-SOD	211	VHKLEPKDHLKPQNLEGISNEQIEPHFEAHYKGYVAK YNEIQEKLADQNFADRSKANQNYSEYRELKVEETFNY MGVVLHELYFGMLTPGGKGEPSEALKKKIEEDIGGLD ACTNELKAAAMAFRGWAILGLDIFSGRLVVNGLDAH NVYNLTGLIPLIVIDTYEHAYYVDYKNKRPPYIDAFFK NINWDVVNERFEKAMKAYEALKDFIK	1COJ
SUPEROXIDE DISMUTASE	212	MVSFKRYELPPLPYNYNALEPYIIEEIMKLHHQKHHNT YVKGANAALEKIEKHLKGEIQIDVRAVMRDFSFNYAG HIMHTIFWPNMAPPGKGGGTPGGRVADLIEKQFGGFE KFKALFSAAAKTVEGVGWGVLAFDPLTEELRILQVEK HNVLMTAGLVPILVIDVWEHAYYLQYKNDRGSYVEN WWNVVNWDDVEKRLEQALNNAKPLYLL	3AK1
ADENYLATE KINASE	214	MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAA VKSGSELGKQAKDIMDAGKLVTDELVIALVKERIAQE DCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVP DELIVDRIVGRRVHAPSGRVYHVKFNPPKVEGKDDVT GEELTTRKDDQEETVRKRLVEYHQMTAPLIGYYSKEA EAGNTKYAKVDGTKPVAEVRADLEKILG	1AKE
ADENYLATE KINASE WITH BOUND AP5A	217	MNLVLMGLPGAGKGTQAEKIVAAYGIPHISTGDMFRA AMKEGTPLGLQAKQYMDRGDLVPDEVTIGIVRERLSK DDCQNGFLLDGFPRTVAQAEALETMLADIGRKLDYVI HIDVRQDVLMERLTGRRICRNCGATYHLIFHPPAKPGV CDKCGGELYQRADDNEATVANRLEVNMKQMKPLVD FYEQKGYLRNINGEQDMEKVFADIRELLGGLAR	1ZIN
5'-DEOXY-5'-METHYLTHIOADENOSINE PHOSPHORYLASE	226	PVHILAKKGEVAERVLVVGDPGRARLLSTLLQNPKLT NENRGFLVYTGKYNGETVSIATHGIGGPSIAIVLEELA MLGANVFIRYGTTGALVPYINLGEYIIVTGASYNQGGL FYQYLRDNACVASTPDFELTNKLVTSFSKRNLKYYVG NVFSSDAFYAEDEEFVKKWSSRGNIAVEMECATLFTL SKVKGWKSATVLVVSDNLAKELEKSVMDGAKAVLD TLTS	1JDS
5'-METHYLTHIOADENOSINE	226	MKIGIIGAMEEEVTLLRDKIENRQTISLGGCEIYTGQLN GTEVALLKSGIGKVAAALGATLLLEHCKPDVIINTGSA GGLAPTLKVGDIVVSDEARYHDADVTAFGYEYGQLP GCPAGFKADDKLIAAAEACIAELNLNAVRGLIVSGDAF INGSVGLAKIRHNFPQAIAVEMEATAIAHVCHNFNVPF VVVRAISDVADQSFDEFLAVAAKQSSLMVESLVQKLA	1JYS
RIBOSE-5-PHOSPHATE ISOMERASE A	229	MNVEEMKKIAAKEALKFIEDDMVIGLGTGSTTAYFIKL LGEKLKRGEISDIVGVPTSYQAKLLAIEHDIPIASLDQV DAIDVAVDGADEVDPNLNLIKGRGAALTMEKIIEYRA GTFIVLVDERKLVDYLCQKMPVPIEVIPQAWKAIIEELS IFNAKAELRMGVNKDGPVITDNGNFIIDAKFPRIDDPL DMEIELNTIPGVIENGIFADIADIVIVGTREGVKKLER	1LK5
BACTERIORHODOPSIN-I	236	APGSEGIWLWLGTAGMFLGMLYFIARGWGETDGRRQ KFYIATILITAIAFVNYLAMALGFGLTFIEFGGEQHPIY WARYTDWLFTTPLLLYNLGLLAGADRNTIYSLVSLDV LMIGTGVVATLSAGSGVLSAGAERLVWWGISTAFLLV LLYFLFSSLSGRVANLPSDTRSTFKTLRNLVTVVWLVY PVWWLVGSEGLGLVGIGIETAGFMVIDLVAKVGFGIIL LRSHGVLDGAA	4PXK
THIOREDOXIN PEROXIDASE FROM AEROPYRUM PERNIX K1	240	PGSIPLIGERFPEEVTTDHGVIKLPDHYVSQGKWFVLFS HPADFTPVCTTEFVSFARRYEDFQRLGVDLIGLSVDSV FSHIKWKEWIERHIGVRIPFPIIADPQGTVARRLGLLHA ESATHTVRGVFIVDARGVIRTLYYPELGRLVDEILRIVK ALKLGDSLKRAVPADWPNNEIIGEGLIVPPPTTEDQAR	1X0R

		ARESGQYRSLDWWFCWDTPASRDDVEEARRYLRRAA EKPAKLLYEEA	
3'-PHOSPHOADENOSINE 5'-PHOSPHATE PHOSPHATASE	266	SHHHHHHSTDDLTDAELAADLAADAGKLLLQVRAEI GFDQPWTLGEAGDRQANSLLLRRLQAERPGDAVLSEE AHDDLARLKSDRVWIIDPLDGTREFSTPGRDDWAVHI ALWRRSSNGQPEITDAAVALPARGNVVYRTDTVTSGA APAGVPGTLRIAVSATRPPAVLHRIRQTLAIQPVSIGSA GAKAMAVIDGYVDAYLHAGGQWEWDSAAPAGVML AAGMHASRLDGSPLRYNQLDPYLPDLLMCRAEVAPIL LGAIADAWR	5DJF
LEGUMAIN	267	GGKHWVVIVAGSNGWYNYRHQADACHAYQIIHRNGI PDEQIVVMMYDDIAYSEDNPTPGIVINRPNGTDVYQG VPKDYTGEDVTPQNFLAVLRGDAEAVKGIGSGKVLKS GPQDHVFIYFTDHGSTGILVFPNEDLHVKDLNETIHYM YKHKMYRKMVFYIEACESGSMMNHLPDNINVYATTA ANPRESSYACYYDEKRSTYLGDWYSVNWMEDSDVED LTKETLHKQYHLVKSHTQTSHVMQYGNKTISTMKVM QFQGMKRYVAD	4AW9
APO CUTICLE-DEGRADING PROTEASE (VER112)	279	ITQQQGATWGLTRISHRARGSTAYAYDTSAGAGACVY VIDTGVEDTHPDFEGRAKQIKSYASTARDGHGHGTHC AGTIGSKTWGVAKKVSIFGVKVLDDSGSGSLSNIIAGM DFVASDRQSRNCPRRTVASMSLGGGYSAALNQAAAR LQSSGVFVAVAAGNDNRDAANTSPASEPTVCTVGATD SNDVRSTFSNYGRVVDIFAPGTSITSTWIGGRTNTISGT SMATPHIAGLAAYLFGLEGGSAGAMCGRIQTLSTKNV LTSIPSGTVNYLAFNGAT	3F7M
CATHEPSIN L-LIKE PROTEINASE	306	NDDLWHQWKRMYNKEYNGADDQHRRNIWEKNVKH IQEHNLRHDLGLVTYTLGLNQFTDMTFEEFKAKYLTE MSRASDILSHGVPYEAVPDKIDWRESGYVTEVKDQGN CGSGWAFSTTGTMEGQYMKNERTSISFSEQQLVDCSR PWGNNGCGGGLMENAYQYLKQFGLETESSYPYTAVE GQCRYNKQLGVAKVTGFYTVHSGSEVELKNLVGAEG PAAVAVDVESDFMMYRSGIYQSQTCSPLRVNHAVLA VGYGTQGGTDYWIVKNSWGLSWGERGYIRMVRNRG NMCGIASLASLPMVARFP	2O6X
POLYAMINE AMINOPROPYLTRANSFERASE	309	MDYGMYFFEHVTPYETLVRRMERVIASGKTPFQDYFL FESKGFGKVLILDKDVQSTERDEYIYHETLVHPAMLTH PEPKRVLIVGGGEGATLREVLKHPTVEKAVMVDIDGE LVEVAKRHMPEWHQGAFDDPRAVLVIDDARAYLERT EERYDVVIIDLTDPVGEDNPARLLYTVEFYRLVKAHLN PGGVMGMQTGMILLRVHPVVHRTVREAFRYVRSYKN HIPGFFLNFGFLLASDAFDPAAFSEGVIEARIRERNLAL RHLTAPYLEAMFVLPKDLLEALEKETMVSTDQNPFYV TPEGEARQAPY	1UIR
MALATE DEHYDROGENASE	319	ARSKIALIGAGQIGGTLAHLAGLKELGDVVLFDIVDGV PQGKALDIAESAPVDGFDAKYSGASDYSAIAGADVVI VTAGVPRKPGMSRDDLIGINLKVMEAVGAGIKEHAPD AFVICITNPLDAMVWALQKFSGLPTNKVVGMAGVLDS ARFRHFLAEEFGVSVEDVTAFVLGGHGDDMVPLTRYS TVAGVPLTDLVKLGWTTQEKLDAMVERTRKGGGEIV NLLKTGSAFYAPAASAIAMAESYLRDKKRVLPCAAYL DGQYGIDGLYVGVPVVIGENGVERVLEVTFNDDEKA MFEKSVNSVKGLIEACKSVNDKLA	4ROR
ENDO-1,4-BETA-XYLANASE A	327	AASGLEAAMKAAGKQYFGTALTVRNDQGEIDIINNKN EIGSITPENAMKWEAIQPNRGQFNWGPADQHAAAATS RGYELRCHTLVWHSQLPSWVANGNWNNQTLQAVMR DHINAVMGRYRGKCTHWDVVNEALNEDGTYRDSVFL RVIGEAYIPIAFRMALAADPTTKLYYNDYNLEYGNAK TEGAKRIARLVKSYGLRIDGIGLQAHMTSESTPTQNTP TPSRAKLASVLQGLADLGVDVAYTELDIRMNTPATQQ KLQTNADAYARIVGSCMDVKRCVGITVWGISDKYSW VPGTFPGEGSALLWNDNFQKKPSYTSTLNTINRR	3U7B

PH0655	327	EKVAIKTKPGYGAELVEVDVPKPGPGEVLIKVLATSIC GTDLHIYEWNEWAQSRIKPPQIGHEVAGEVVEIGPGVE GIEVGDYVSVETHIVCGKCYTKIFGVDTDGVFAEYAV VPAQNIWKNPKSIPPEYATLQEPLGNAVDTVLAGPISG KSVLITGAGPLGLLGIAVAKASGAYPVIVSEPSDFRREL AKKVGADYVINPFEEDVVKEVDITDGNGVDVFLEFSG APKALEQGLQAVTPAGRVSLLGLYPGKVTIDFNNLIIF KALTIYGITGRHLWETWYTVSRLLQSGKLNLDPIITHK YKGFDKYEEAFELRAGKTGKVVFL	2D8A
MALATE DEHYDROGENASE	337	FEKGYVDENYIRVPKDRLFSFIVRVLTKLGVPEEDAKI VADNLVADLRGVESHGVQRLKRYVDGIISGGVNLHPK IRVIREGPSYALIDGDEGLGQVVGYRSKLAIKKAKDTG IGIVIARNSNHYGIAGYYALAAEEGIGISTNSRPLVAPT GGIERILGTNPIALAAPTKDKPFLLDATSVVPIGKLEWA INREGNITTKVEEVFNGGALLPLGGFGELLGGHKGYGL SLVDILSGILSGGTWSKYVKNTSEKGSNVCHFFVIDIEH FIPLEEFKEKISQIEEIKSSRKHPEFERIWIHGEKGFLTET RLKLGIPIYRKVLEELNEIAKRVGVEGL	1V9N
CHORISMATE MUTASE	344	EAAVTQSPRNKVAVTGEKVTLSCQQTNNHNNMYWY RQDTGHGLRLIHYSYGVGNTEKGDIPDGYEASRPSQE QFSLILESATPSQTSVYFCASGGGGTLYFGAGTRLSVLS QPDPMPDDLHKSSEFTGTMGNMKYLYDDHYVSATKV KSVDKFLAHDLIYNISDKKLKNYDKVKTELLNEDLAK KYKDEVVDVYGSNYYVNCYFSSKDNVWWPGKTCMY GGITKHEGNHFDNGNLQNVLVRVYENKRNTISFEVQT DKKSVTAQELDIKARNFLINKKNLYEFNSSPYETGYIK FIENNGNTFWYDMMPAPGDKFDQSKYLMMYNDNKT VDSKSVKIEVHLTTK	2AO2
GLUCOSE 1-DEHYDROGENASE	355	MKAIAVKRGRPVVIEKPRPEPESGEALVRTLRVGVDG TDHEVIAGGHGGFPEGEDHLVLGHEAVGVVVDPNDT ELEEGDIVVPTVRRPPASGTNEYFERDQPDMAPDGMY FERGIVGAHGYMSEFFTSPEKYLVRIPRSQAELGFLIEPI SITEKALEHAYASRSAFDWDPSSAFVLGNGSLGLLTLA MLKVDDKGYENLYCLGRRDRPDPTIDIIEELDATYVDS RQTPVEDVPDVYEQMDFIYEATGFPKHAIQSVQALAP NGVGALLGVPSDWAFEVDAGAFHREMVLHNKALVG SVNSHVEHFEAATVTFTKLPKWFLEDLVTGVHPLSEFE AAFDDDDTTIKTAIEFSTV	2B5V
PHOSPHOSERINE AMINOTRANSFERASE	360	VKQVFNFNAGPSALPKPALERAQKELLNFNDTQMSV MELSHRSQSYEEVHEQAQNLLRELLQIPNDYQILFLQG GASLQFTMLPMNLLTKGTIGNYVLTGSWSEKALKEAK LLGETHIAASTKANSYQSIPDFSEFQLNENDAYLHITSN NTIYGTQYQNFPEINHAPLIADMSSDILSRPLKVNQFG MIYAGAQKNLGPSGVTVVIVKKDLLNTKVEQVPTML QYATHIKSDSLYNTPPTFSIYMLRNVLDWIKDLGGAEA IAKQNEEKAKIIYDTIDESNGFYVGHAEKGSRSLMNVT FNLRNEELNQQFLAKAKEQGFVGLNGHRSVGGCRASI YNAVPIDACIALRELMIQFKENA	1W23
PH-SENSITIVE ADENYLATE CYCLASE RV1264	360	DDLLGDLGGTARAERAKLVEWLLEQGITPDEIRATNPP LLLATRHLVGDDGTYVSAREISENYGVDLELLQRVQR AVGLARVDDPDAVVHMRADGEAAARAQRFVELGLN PDQVVLVVRVLAEGLSHAAEAMRYTALEAIMRPGAT ELDIAKGSQALVSQIVPLLGPMIQDMLFMQLRHMMET EAVNAGERAAGKPLPGARQVTVAFADLVGFTQLGEV VSAEELGHLAGRLAGLARDLTAPPVWFIKTIGDAVML VCPDPAPLLDTVLKLVEVVDTDNNFPRLRAGVASGM AVSRAGDWFGSPVNVASRVTGVARPGAVLVADSVRE ALGDADGFQWSFAGPRRLRGIRGDVRLFRVRR	1Y10
CYTOCHROME P450 119	367	MYDWFSEMRKKDPVYYDGNIWQVFSYRYTKEVLNN FSKFSSDLTGYHERLEDLRNGKIRFDIPTRYTMLTSDPP LHDELRSMSADIFSPQKLQTLETFIRETTRSLLDSIDPRE DDIVKKLAVPLPIIVISKILGLPIEDKEKFKEWSDLVAFR	1F4T

		LGKPGEIFELGKKYLELIGYVKDHLNSGTEVVSRVVNS NLSDIEKLGYIILLLIAGNETTTNLISNSVIDFTRFNLWQ RIREENLYLKAIEEALRYSPPVMRTVRKTKERVKLGDQ TIEEGEYVRVWIASANRDEEVFHDGEKFIPDRNPNPHL SFGSGIHLCLGAPLARLEARIAIEEFSKRFRHIEILDTEK VPNEVLNGYKRLVVRLKSN	
SUCCINYL-DIAMINOPIMELATE DESUCCINYLASE	370	MKEKVVSLAQDLIRRPSISPNDEGCQQIIAERLEKLGFQ IEWMPFNDTLNLWAKHGTSEPVIAFAGHTDVVPTGDE NQWSSPPFSAEIIDGMLYGRGAADMKGSLAAMIVAAE EYVKANPNHKGTIALLITSDEEATAKDGTIHVVETLMA RDEKITYCMVGEPSSAKNLGDVVKNGRRGSITGNLYI QGIQYPHLAENPIHKAALFLQELTTYQWDKGNEFFPPT SLQIANIHAGTGSVIPAELYIQFNLRYCTEVTDEIIKQKV AEMLEKHNLKYRIEWNLSGKPFLTKPGKLLDSITSAIE ETIGITPKAETGGGTSDGRFIALMGAEVVEFGPLNSTIH KVNECVSVEDLGKCGEIYHKMLVNLLD	3IC1
HMG-COA SYNTHASE FROM ENTEROCOCCUS FAECALIS	383	MTIGIDKISFFVPPYYIDMTALAEARNVDPGKFHIGIGQ DQMAVNPISQDIVTFAANAAEAILTKEDKEAIDMVIVG TESSIDESKAAAVVLHRLMGIQPFARSFEIKEACYGAT AGLQLAKNHVALHPDKKVLVVAADIAKYGLNSGGEP TQGAGAVAMLVASEPRILALKEDNVMLTQDIYDFWR PTGHPYPMVDGPLSNETYIQSFAQVWDEHKKRTGLDF ADYDALAFHIPYTKMGKKALLAKISDQTEAEQERILA RYEESIIYSRRVGNLYTGSLYLGLISLLENATTLTAGNQ IGLFSYGSGAVAEFFTGELVAGYQNHLQKETHLALLD NRTELSIAEYEAMFAETLDTDIDQTLEDELKYSISAINN TVRSYRN	1X9E
ASPARTATE AMINOTRANSFERASE	388	MKLAARVESVSPSMTLIIDAKAKAMKAEGIDVCSFSA GEPDFNTPKHIVEAAKAALEQGKTRYGPAAGEPRLRE AIAQKLQRDNGLCYGADNILVTNGGKQSIFNLMLAMI EPGDEVIIPAPFWVSYPEMVKLAEGTPVILPTTVETQFK VSPEQIRQAITPKTKLLVFNTPSNPTGMVYTPDEVRAIA QVAVEAGLWVLSDEIYEKILYDDAQHLSIGAASPEAY ERSVVCSGFAKTYAMTGWRVGFLAGPVPLVKAATKI QGHSTSNVCTFAQYGAIAAYENSQDCVQEMLAAFAE RRRYMLDALNAMPGLECPKPDGAFYMFPSIAKTGRSS LDFCSELLDQHQVATVPGAAFGADDCIRLSYATDLDTI KRGMERLEKFLHGIL	1J32
PHOSPHONOACETATE HYDROLASE	404	TNLISVNSRSYRLSSAPTIVICVDGCEQEYINQAIQAGQ APFLAELTGFGTVLTGDCVVPSFTNPNNLSIVTGAPPSV HGICGNFFFDQETQEEVLMNDAKYLRAPTILAEMAKA GQLVAVVTAKDKLRNLLGHQLKGICFSAEKADQVNL EEHGVENILARVGMPVPSVYSADLSEFVFAAGLSLLTN ERPDFMYLSTTDYVQHKHAPGTPEANAFYAMMDSYF KRYHEQGAIVAITADHGMNAKTDAIGRPNILFLQDLL DAQYGAQRTRVLLPITDPYVVHHGALGSYATVYLRD AVPQRDAIDFLAGIAGVEAVLTRSQACQRFELPEDRIG DLVVLGERLTVLGSAADKHDLSGLTVPLRSHGGVSEQ KVPLIFNRKLVGLDGRLRNFDIIDLALNHLA	1EI6
XAA-PRO DIPEPTIDASE HALO	425	MNKLAVLYAEHIATLQKRTREIIERENLDGVVFHSGQ AKRQFLDDMYYPFKVNPQFKAWLPVIDNPHCWIVAN GTDKPKLIFYRPVDFWHKVPDEPNEYWADYFDIELLV KPDQVEKLLPYDKARFAYIGEYLEVAQALGFELMNPE PVMNFYHYHRAYKTQYELACMREANKIAVQGHKAA RDAFFQGKSEFEIQQAYLLATQHSENDNPYGNIVALNE NCAILHYTHFDRVAPATHRSFLIDAGANFNGYAADITR TYDFTGEGEFAELVATMKQHQIALMNQLAPGKLYGE LHLDCHQRVAQTLSDFNIVDLSADEIVAKGITSTFFPH GLGHHIGLQVHDVGGFMALRCTRKIEANQVFTIEPGL YFIDSLLGDLAATDNNQHINWDKVAELKPFGGIRIEDN IIVHEDSLENMTRELRLR	3L24

ALPHA-AMYLASE FROM BACILLUS SUBTILIS	425	LTAPSIKSGTILHAWNWSFNTLKHNMKDIHDAGYTAI QTSPINQVKEGNQGDKSMSNWYWLYQPTSYQIGNRY LGTEQEFKEMCAAAEEYGIKVIVDAVINHTTFDYAAIS NEVKSIPNWTHGNTQIKNWSDRWDVTQNSLLGLYDW NTQNTQVQSYLKRFLERALNDGADGFRFDAAKHIELP DDGSYGSQFWPNITNTSAEFQYGQILQDSASRDAAYA NYMDVTASNYGHSIRSALKNRNLGVSNISHYASDVSA DKLVTWVESHDTYANDDEESTWMSDDDIRLGWAVIA SRSGSTPLFFSRPEGGGNGVRFPGKSQIGDRGSALFED QAITAVNRFHNVMAGQPEELSNPNGNNQIFMNQRGSH GVVLANAGSSSVSINTATKLPDGRYDNKAGAGSFQVN DGKLTGTINARSVAVLYPD	1BAG
ALPHA-AMYLASE FROM ALTEROMONAS HALOPLANCTIS	448	TPTTFVHLFEWNWQDVAQECEQYLGPKGYAAVQVSP PNEHITGSQWWTRYQPVSYELQSRGGNRAQFIDMVNR CSAAGVDIYVDTLINHMAAGSGTGTAGNSFGNKSFPIY SPQDFHESCTINNSDYGNDRYRVQNCELVGLADLDTA SNYVQNTIAAYINDLQAIGVKGFRFDASKHVAASDIQS LMAKVNGSPVVFQEVIDQGGEAVGASEYLSTGLVTEF KYSTELGNTFRNGSLAWLSNFGEGWGFMPSSSAVVFV DNHDNQRGHGGAGNVITFEDGRLYDLANVFMLAYPY GYPKVMSSYDFHGDTDAGGPNVPVHNNGNLECFASN WKCEHRWSYIAGGVDFRNNTADNWAVTNWWDNTN NQISFGRGSSGHMAINKEDSTLTATVQTDMASGQYCN VLKGELSADAKSCSGEVITVNSDGTINLNIGAWDAMAI HKNAKLN	1AQM
YGJG	453	SASALACSAHALNLIEKRTLDHEEMKALNREVIEYFKE HVNPGFLEYRKSVTAGGDYGAVEWQAGSLNTLVDTQ GQEFIDCLGGFGIFNVGHRNPVVVSAVQNQLAKQPLH SQELLDPLRAMLAKTLAALTPGKLKYSFFCNSGTESVE AALKLAKAYQSPRGKFTFIATSGAFHGKSLGALSATA KSTFRKPFMPLLPGFRHVPFGNIEAMRTALNECKKTGD DVAAVILEPIQGEGGVILPPPGYLTAVRKLCDEFGALM ILDEVQTGMGRTGKMFACEHENVQPDILCLAKALGGG VMPIGATIATEEVFSVLFDNPFLHTTTFGGNPLACAAA LATINVLLEQNLPAQAEQKGDMLLDGFRQLAREYPDL VQEARGKGMLMAIEFVDNEIGYNFASEMFRQRVLVA GTLNNAKTIRIEPPLTLTIEQCELVIKAARKALAAMRVS VEEA	4UOX
BACILLUS LICHENIFORMIS ALPHA-AMYLASE	481	LNGTLMQYFEWYMPNDGQHWKRLQNDSAYLAEHGI TAVWIPPAYKGTSQADVGYGAYDLYDLGEFHQKGTV RTKYGTKGELQSAIKSLHSRDINVYGDVVINHKGGAD ATEDVTAVEVDPADRNRVISGEHLIKAWTHFHFPGRG STYSDFKWHWYHFDGTDWDESRKLNRIYKFQGKAW DWEVSNEFGNYDYLMYADIDYDHPDVAAEIKRWGT WYANELQLDGFRLDAVKHIKFSFLRDWVNHVREKTG KEMFTVAEYWSYDLGALENYLNKTNFNHSVFDVPLH YQFHAASTQGGGYDMRKLLNGTVVSKHPLKSVTFVD NHDTQPGQSLESTVQTWFKPLAYAFILTRESGYPQVFY GDMYGTKGDSQREIPALKHKIEPILKARKQYAYGAQH DYFDHHDIVGWTREGDSSVANSGLAALITDGPGGAKR MYVGRQNAGETWHDITGNRSEPVVINSAGWGEFHVN GGSVSIYVQR	1BLI
ALPHA-AMYLASE FROM H.ORENII	488	FEKHGTYYEIFVRSFYDSDGDGIGDLKGIIEKLDYLND GDPETIADLGVNGIWLMPIFKSPSYHGYDVTDYYKINP DYGTLEDFHKLVEAAHQRGIKVIIDLPINHTSERHPWF LKASRDKNSEYRDYYVWAGPDTDTKETKLDGGRVW HYSPTGMYYGYFWSGMPDLNYNNPEVQEKVIGIAKY WLKQGVDGFRLDGAMHIFPPAQYDKNFTWWEKFRQE IEEVKPVYLVGEVWDISETVAPYFKYGFDSTFNFKLAE AVIATAKAGFPFGFNKKAKHIYGVYDREVGFGNYIDA PFLTNHDQNRILDQLGQDRNKARVAASIYLTLPGNPFI YYGEEIGMRGQGPHEVIREPFQWYNGSGEGETYWEPA	1WZA

		MYNDGFTSVEQEEKNLDSLLNHYRRLIHFRNENPVFY TGKIEIINGGLNVVAFRRYNDKRDLYVYHNLVNRPVKI KVASGNWTLLFNSGDKEITPVEDNNKLMYTIPAYTTIV LEKE	
LISTERIOLYSIN O	488	MAPPASPPASPKTPIEKKHADEIDKYIQGLDYNKNNVL VYHGDAVTNVPPRKGYKDGNEYIVVEKKKKSINQNN ADIQVVNAISSLTYPGALVKANSELVENQPDVLPVKR DSLTLSIDLPGMTNQDNKIVVKNATKSNVNNAVNTLV ERWNEKYAQAYPNVSAKIDYDDEMAYSESQLIAKFGT AFKAVNNSLNVNFGAISEGKMQEEVISFKQIYYNVNV NEPTRPSRFFGKAVTKEQLQALGVNAENPPAYISSVAY GRQVYLKLSTNSHSTKVKAAFDAAVSGKSVSGDVELT NIIKNSSFKAVIYGGSAKDEVQIIDGNLGDLRDILKKGA TFNRETPGVPIAYTTNFLKDNELAVIKNNSEYIETTSKA YTDGKINIDHSGGYVAQFNISWDEVNYDPEGNEIVQH KNWSENNKSKLAHFTSSIYLPGNARNINVYAKECTGL AWEWWRTVIDDRNLPLVKNRNISIWGTTLYPKYSNK VDN	4CDB
ALKALINE PHOSPHATASE	497	EVKNVILMIGDGMGPQQVGLLETYANQAPDSIYDGEP TAFHQLAKEGVVGFSLTHPEDAVVVDSACSATQLASG IYSGSEVIGIDAEGNPVETVLELAQARGKATGLVSDTR LTHATPAAFAAHQPHRSLENEIAVDMLEVGPDVMLSG GLRHWVPQSASEDAEVTSLMDGAYEPASKRQDDRNL LAEAVEKGYGLAFSREQLEADQSDKLLGLFANSGMA DGIEYRNTRDDADRREPTLHEMTQAALNRLEQDEDGF FLMVEGGQIDWAGHSNDAGTMLNEMVKFEEAVQGV YDWAKGREDTVILVTADHETGAFGLSYSSADLPEPQS KSGPAFAERDYAPNFNFGDFALLDSLYHQKASFSTLLS EFGALEEEQRTPARLMEMVNANSDFQIDEEQAEAVLA DKPNPYHVEGHSYLEAEEVPAIQDFDAFYPYNDRGNV LGRVLGTAQNVVWGTGTHTHTPVNVFAWGPAETILP VSSIQHHSEVGQYLKSLVE	3WBH
MOUSE AUTOTAXIN	500	WTNTSGSCKGRCFELQEVGPPDCRCDNLCKSYSSCCH DFDELCLKTARGWECTKDRCGEVRNEENACHCSEDC LSRGDCCTNYQVVCKGESHWVDDDCEEIRVPECPAGF VRPPLIIFSVDGFRASYMKKGSKVMPNIEKLRSCGTHA PYMRPVYPTKTFPNLYTLATGLYPESHGIVGNSMYDP VFDATFHLRGREKFNHRWWGGQPLWITATKQGVRAG TFFWSVSIPHERRILTILQWLSLPDNERPSVYAFYSEQP DFSGHKYGPFGPEMTNPLREIDKTVGQLMDGLKQLKL HRCVNVIFVGDHGMEDVTCDRTEFLSNYLTNVDDITL VPGTLGRIRPKIPNNLKYDPKAIIANLTCKKPDQHFKPY MKQHLPKRLHYANNRRIEDLHLLVERRWHVARKPLD VYKKPSGKCFFQGDHGFDNKVNSMQTVFVGYGPTFK YRTKVPPFENIELYNVMCDLLGLKPAPNNGTHGSLNH LLRTNTFRPTLPEEVSRP	3NKM
OXALYL-COA DECARBOXYLASE	500	LQMTDGMHIIVEALKQNNIDTIYGVVGIPVTDMARHA QAEGIRYIGFRHEQSAGYAAAASGFLTQKPGICLTVSA PGFLNGLTALANATVNGFPMIMISGSSDRAIVDLQQGD YEELDQMNAAKPYAKAAFRVNQPQDLGIALARAIRVS VSGRPGGVYLDLPANVLAATMEKDEALTTIVKVENPS PALLPCPKSVTSAISLLAKAERPLIILGKGAAYSQADEQ LREFIESAQIPFLPMSMAKGILEDTHPLSAAAARSFALA NADVVMLVGARLNWLLAHGKKGWAADTQFIQLDIEP QEIDSNRPIAVPVVGDIASSMQGMLAELKQNTFTTPLV WRDILNIHKQQNAQKMHEKLSTDTQPLNYFNALSAVR DVLRENQDIYLVNEGANTLDNARNIIDMYKPRRRLDC GTWGVMGIGMGYAIGASVTSGSPVVAIEGDSAFGFSG MEIETICRYNLPVTIVIFNNGGIYRGDGVDLSGAGAPSP TDLLHHARYDK	2Q27

ARYLSULFATASE FROM PSEUDOMONAS AERUGINOSA	526	KRPNFLVIVADDLGFSDIGAFGGEIATPNLDALAIAGLR LTDFHTASTSPTRSMLLTGTDHHIAGIGTMAEALTPEL EGKPGYEGHLNERVVALPELLREAGYQTLMAGKWHL GLKPEQTPHARGFERSFSLLPGAANHYGFEPPYDESTP RILKGTPALYVEDERYLDTLPEGFYSSDAFGDKLLQYL KERDQSRPFFAYLPFSAPHWPLQAPREIVEKYRGRYDA GPEALRQERLARLKELGLVEADVEAHPVLALTREWEA LEDEERAKSARAMEVYAAMVERMDWNIGRVVDYLR RQGELDNTFVLFMSDNGAEGALLEAFPKFGPDLLGFL DRHYDNSLENIGRANSYVWYGPRWAQAATAPSRLYK AFTTQGGIRVPALVRYPRLSRQGAISHAFATVMDVTPT LLDLAGVRHPGKRWRGREIAEPRGRSWLGWLSGETE AAHDENTVTGWELFGMRAIRQGDWKAVYLPAPVGPA TWQLYDLARDPGEIHDLA	1HDH
SPAP	528	AARSIAATPPKLIVAISVDQFSADLFSEYRQYYTGGLK RLTSEGAVFPRGYQSHAATETCPGHSTILTGSRPSRTGI IANNWFDLDAKREDKNLYCAEDESQPGSSSDKYEASP LHLKVPTLGGRMKAANPATRVVSVAGKDRAAIMMG GATADQVWWLGGPQGYVSYKGVAPTPLVTQVNQAF AQRLAQPNPGFELPAQCVSKDFPVQAGNRTVGTGRFA RDAGDYKGFRISPEQDAMTLAFAAAAIENMQLGKQA QTDIISIGLSATDYVGHTFGTEGTESCIQVDRLDTELGA FFDKLDKDGIDYVVVLTADHGGHDLPERHRMNAMPM EQRVDMALTPKALNATIAEKAGLPGKKVIWSDGPSGD IYYDKGLTAAQRARVETEALKYLRAHPQVQTVFTKAE IAATPSPSGPPESWSLIQEARASFYPSRSGDLLLLLKPR VMSIPEQAVMGSVATHGSPWDTDRRVPILFWRKGMQ HFEQPLGVETVDILPSLAA	3Q3Q
CATALASE-PEROXIDASE	709	KRPKSNQDWWPSKLNLEILDQNARDVGPVEDDFDYA EEFQKLDLEAVKSDLEELMTSSQDWWPADYGHYGPL FIRMAWHSAGTYRTADGRGGAAGGRQRFAPINSWPD NANLDKARRLLLPIKQKYGQKISWADLMILAGNVAIE SMGFKTFGYAGGREDAFEEDKAVNWGPEDEFETQER FDEPGEIQEGLGASVMGLIYVNPEGPDGNPDPEASAKN IRQTFDRMAMNDKETAALIAGGHTFGKVHGADDPEE NLGPEPEAAPIEQQGLGWQNKNMITSGIEGPWTQSPTE WDMGYINNLLDYEWEPEKGPGGAWQWAPKSEELKN SVPDAHDPDEKQTPMMLTTDIALKRDPDYREVMETFQ ENPMEFGMNFAKAWYKLTHRDMGPPERFLGPEVPDE EMIWQDPLPDADYDLIGDEEIAELKEEILDSDLSVSQL VKTAWASASTYRDSDKRGGANGARLRLEPQKNWEV NEPEQLETVLGTLENIQTEFNDSRSD	1ITK

METAMORPHIC PROTIEN DATA SET.

Name	N	SEQUENCE	PDB ID
RESPIRATORY SYNCYTIAL VIRUS FUSION PROTEIN CORE	50	LEGEVNKIKSALLSTNKAVVSLSNGVSVLTSKVLDLK NYIDKQLLPIVNK	1G2C
PARAMYXOVIRUS SV5	62	TAAVALVKANENAAAILNLKNAIQKTNAAVADVVQA TQSLGTAVQAVQDHINSVVSPAITAA	1SVF
PRGI MUTANT	62	GVDNLQTQVTEALDKLAAKPSDPALLAAYQSKLSEYN LYRNAQSNTAKAFKDIDAAIIQNFR	2X9C
HENDRA VIRUS FUSION CORE	63	NINKLKSSIESTNEAVVKLQETAEKTVYVLTALQDSSQ ISSMNQSLQQSKDYIKEAQKILDTV	1WP8
MURINE SAK	75	SVFVKNVGWATQLTSGAVWVQFNDGSQLVMQAGVS SISYTSPDGQTTRYGENEKLPEYIKQKLQLLSSILLMFS N	1MBY
YEAST MATALPHA2	77	GLVFNVVTQDMINKSTKPYRGHRFTKENVRILESWFA KNIENPYLDTKGLENLMKNTSLSRIQIKNWVSNRRRKE KT	1MNM
FIBRONECTIN 8-9FNI DOMAIN PAIR	91	DQCIVDDITYNVQDTFHKKHEEGHMLNCTCFGQGRG RWKCDPVDQCQDSETGTFYQIGDSWEKYVHGVRYQC YCYGRGIGEWHCQPLQTYP	3EJH
DOMAIN-SWAPPED DIMER	92	SGLSVHTDASVTKAAAPESGLEVRDRWLKITIPNAFLG SDVVDWLYHHVEGFPERREARKYASGLLKAGLIRHTV NKITFSEQCYYVFGDLS	5SUZ
MUTANT RABBIT PRP 121-230 (S170N)	97	GGYMLGSAMSRPLIHFGNDYEDRYYRENMYRYPNQV YYRPVDQYNNQNSFVHDCVNITVKQHTVTTTTKGENF TETDIKIMERVVEQMCITQYQQES	4HLS
BETA 2 MICROGLOBULIN DOMAIN-SWAPPED DIMER	99	IQRTPKIQVYSRHPAENGKSNFLNCYVSGFHPSDIEVDL LKNGERIEKVEHSDLSFSKDWSFYLLYYTEFTPTEKDE YACRVNHVTLSQPKIVKWDRDM	3LOW
ISCA	102	MVELTPAAIQELERLQILRIQVQPSECGDWRYDLALVA EPKPTDLLTQSQGWTIAIAAEAAELLRGLRVDYIEDLM GGAFRFHNPNASQTCGCGMAFRVSRS	1X0G
LYMPHOCYTIC CHORIOMENINGITIS VIRUS MEMBRANE FUSION GLYCOPROTEIN	103	EEFSDMLRLIDYNKAALSKFKQDVESALHVFKTTVNS LISDQLLMRNHLRDLMGVPYCNYSKFWYLEHAPKCW LVTNGSYLNETHFSDQIEQEADNMITEMLR	3MKO
CYSTATIN C	107	GPMDASVEEEGVRRALDFAVGEYNKASNDMYHSRAC QVVRARKQIVAGVNYFLDVELCRTTCTKTQLDNCPFH DQPHLKRKAFCSFQIYAVPWQGTMTLSKSTCQDA	3GAX
METHANOCALDOCOC CUS JANNASCHII MONOMERIC SELECASE	109	MKDRKILNEILSNTINELNLNDKKANIKIKIKPLKRKIA SISLTNKTIYINKNILPYLSDEEIRFILAHELLHLKYGKY HINEFEEELLFLFPNKEAILINLINKLHQK	4QHF
NF-KB RELB	110	LVPRGSHMNTSELRICRINKESGPCTGGEELYLLCDKV QKEDISVVFSTASWEGRADFSQADVHRQIAIVFKTPPY EDLEISEPVTVNVFLQRLTDGVCSEPLPFTYLPR	1ZK9
CLOSTRIDIUM HISTOLYTICUM	111	EKLKEKENNDSSDKATVIPNFNTTMQGSLLGDDSRDY YSFEVKEEGEVNIELDKKDEFGVTWTLHPESDRITYGQ VDGNKVSNKVKLRPGKYYLLVYKYSGSGNYELRVNK	1NQD
CERBERUS (PRDC)	111	KEVLASSQEALVVTERKYLKSDWCKTQPLRQTVSEEG CRSRTILNRFCYGQCNSFYIPRHVKKEEDSFQSCAFCKP QRVTSVIVELECPGLDPPFRIKKIQKVKHCRCMSV	4JPH
INFLUENZA HAEMAGGLUTININ	119	TLCLGSTQAAIDQINGKLNRVIEKTNEKFHQIEKEFSEV EGRIQDLEKYVEDTKIDLWSYNAELLVALENQHTIDLT	1HTM

		DSEMNKLFEKTRRQLRENAEEMGNGCFKIYHKCDNA CIESIR	
SARS CORONAVIRUS SPIKE GLYCOPROTEIN	124	GVTQNVLYENQKQIANQFNKAISQIQESLTTTSTALGK LQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLD KVEAEVLGDISGINASVVNIQKEIDRLNEVAKNLNESLI DLQELGKYE	1WYY
KERATIN 4 BINDING DOMAIN	141	QTETYITQINPEGKEMYFASGLGNLYTIIGSDGSPVNLL NAEVKILKTNSKNLTNYDSPEFEDVTSQYSYTNDGSKI TIDWKTNSISSTTSYVVLVKIPKQSGVLYSTVSDINQTY GSKYSYGHTNISGDSDANAEIKLLS	4RMB
A4V MUTANT	152	ATKVVCVLKGDGPVQGIINFEQKESNGPVKVWGSIKG LTEGLHGFHVHEFGDNTAGCTSAGPHFNPLSRKHGGP KDEERHVGDLGNVTADKDGVADVSIEDSVISLSGDHC IIGRTLVVHEKADDLGKGGNEESTKTGNAGSRLACGVI GIA	1UXM
KIWELLIN	158	KCNDDPEVGTHICRGTCKPSGTLTCQGKSHPTYDCSPP VTSSTPAKLTNNDFSEGGDGGGPSECDESYHSNNERIV ALSTGWYNGGSRCGKMIRITASNGKSVSAKVVDECDS RHGCDKEHAGQPPCRNNIVDGSNAVWSALGLNKNVG VVDITWSMA	4PMK
AEROPYRUM PERNIX PEROXIREDOXIN Q ENZYME	160	LVELGEKAPDFTLPNQDFEPVNLYEVLKRGRPAVLIFF PAAFSPVCTKELCTFRDKMAQLEKANAEVLAISVDSP WCLKKFKDENRLAFNLLSDYNREVIKLYNVYHEDLK GLKMVAKRAVFIVKPDGTVAYKWVTDNPLNEPDYDE VVREANKIAGELV	4GQC
ENDOLYSIN R21	165	MPPSLRKAVAAAIGGGAIAIASVLITGPSGNDGLEGVS YIPYKDIVGVWTVCHGHTGKDIMLGKTYTKAECKALL NKDLATVARQINPYIKVDIPETMRGALYSFVYNVGAG NFRTSTLLRKINQGDIKGACDQLRRWTYAGGKQWKG LMTRREIEREICLWGQQ	3HDE
POLYMERASE ALPHA-PRIMASE P58	167	SLDQIDLLSTKSFPPCMRQLHKALRENHHLRHGGRMQ YGLFLKGIGLTLEQALQFWKQEFSYNIRHSFRTDYTPF SCLKIILSNPPSQGDYHGCPFRHSDPELLKQKLQSYKIS PGGISQILDLVKGTHYQVACQKYFEMIHNVDDCGFSL NHPNQFFCESQRILNG	3L9Q
P1 PHAGE ENDOLYSIN LYZ	170	GGAICAIAVITIVGNGNVRTNQAGLELIGNAEGCRRDP YCPAGVWTDGIGNTHGVTPGVRKTDQQIAADWEKNI LIAERCINQHFRGKDPDNAFSATSAAFNGCNSLRTYYS KARGRVETSIHKWAQKGEWVNCNHLPDFVNSNGVPL RGLKIRREKERQLCLTGLVNEH	1XJT
SIGNAL RECOGNITION PARTICLE RECEPTOR BETA	191	SYQPSIIIAGPQNSGKTSLLTLLTTDSVRPTVVSQEPLSA ADYDGSGVTLVDFPGHVKLRYKLSDYLKTRAKFVKG LIFMVDSTVDPKKLTTTAEFLVDILSITESSCENGIDILIA CNKSELFTARPPSKIKDALESEIQKVIERRKKSLNELDV LGFKFANLEASVVAFEGSINKRKISQWREWIDEKL	1NRJ
SUN2-KASH1	196	GVTEEQVHHIVKQALQRYSEDRIGLADYALESGGASVI STRCSETYETKTALLSLFGIPLWYHSQSPRVILQPDVHP GNCWAFQGPQGFAVVRLSARIRPTAVTLEHVPKALSP NSTISSAPKDFAIFGFDEDLQQEGTLLGKFTYDQDGEPI QTFHFQAPTMATYQVVELRILTNWGHPEYTCIYRFRV HGEPAH	4DXR
MAD2 DIMER	202	MALQLSREQGITARGSAEIVAEFFSFGINSILYQRGIYPS ETFTRVQKYGLTLLVTTDLELIKYLNNVVEQLKDWLY KSSVQKLVVVISNIESGEVLERWQFDIESDKTAKAPRE KSQKAIQDEIRSVIRQITATVTFLPLLEVSCSFDLLIYTD KDLVVPEKWEESGPQFITNSEEVRLRSFTTTIHKVNSM VAYKIPVND	2VFX
HTRAP1 N-TERMINAL DOMAIN-APO	205	QGSTSKHEFQAETKKLLDIVARSLYSEKEVFIRELISNA SDALEKLRHKLVSDGQALPEMEIHLQTNAEKGTITIQD TGIGMTQEELVSNLGTIARSGSKAFLDALQIIGQFGVGF YSAFMVADRVEVYSRSAAPGSLGYQWLSDGSGVFEIA	5F3K

		EASGVRTGTKIIIHLKSDCKEFSSEARVRDVVTKYSNF VSFPLYLNGRRMNT	
HUMAN APO-COMT	207	TKEQRILNHVLQHAEPGNAQSVLEAIDTYCEQKGDKK GKIVDAVIQEHQPSVLLELGAYCGYSAVRMARLLSPG ARLITIEINPDCAAITQRMVDFAGVKDKVTLVVGASQD IIPQLKKKYDVDTLDMVFLDHWKDRYLPDTLLLEECG LLRKGTVLLADNVICPGAPDFLAHVRGSSCFECTHYQS FLEYREVVDGLEKAIYKGPG	4PYI
SPLICING FACTOR CWC2	225	SWRDKSAKVQVKESELPSSIPAQTGLTFNIWYNKWSQ GFAGNTRFVSPFALQPQLHSGKTRGDNDGQLFFCLFFA KGMCCLGPKCEYLHHIPDEEDIGKLALRTEVLDCFGRE KFADYREDMGGIGSFRKKNKTLYVGGIDGALNSKHLK PAQIESRIRFVFSRLGDIDRIRYVESKNCGFVKFKYQAN AEFAKEAMSNQTLLLPSDKEWDDRREGTGLLVKWAN	3TP2
KSHV PROTEASE	227	QGLYVGGFVDVVSCPKLEQELYLDPDQVTDYLPVTEP LPITIEHLPETEVGWTLGLFQVSHGIFCTGAITSPAFLEL ASRLADTSHVARAPVKNLPKEPLLEILHTWLPGLSLSSI HPRELSQPSGPVFQHVSLCALGRRRGTVAVYGHDAE WVVSRFSSVSKSERAHILQHVSSCRLEDLSTPNFVSPLE TLMAKAIDAGFIRDRLDLLKTDRGVASILSPVYLKA	2PBK
SIR2 HOMOLOG PROTEIN DEACETYLASE.	232	KPRVLVLTGAGISAESGIRTFRAADGLWEEHRVEDVA TPEGFDRDPELVQAFYNARRRQLQQPEIQPNAAHLAL AKLQDALGDRFLLVTQNIDNLHERAGNTNVIHMHGEL LKVRCSQSGQVLDWTGDVTPEDKCPLRPHVVWFGEM PLGMDEIYMALSMADIFIAIGTSGHVYPAAGFVHEAKL HGAHTVELNLEPSQEFAEKYYGPASQVVPEFVEKLLK GLKKGGARHR	1S5P
THERMOTOGA MARITIMA IMPASE TM1415	254	MDRLDFSIKLLRKVGHLLMIHWGRVDNVEKKTGFKDI VTEIDREAQRMIVDEIRKFFPDENIMAEEGIFEKGDRL WIIDPIDGTINFVHGLPNFSISLAYVENGEVKLGVVHAP ALNETLYAEEGSGAFFNGERIRVSENASLEECVGSTGS YVDFTGKFIERMEKRTRRIRILGSAALNAAYVGAGRV DFFVTWRINPWDIAAGLIIVKEAGGMVTDFSGKEANA FSKNFIFSNGLIHDEVVKVVNEVVEEIG	2P3V
METHYLTRANSFERAS E RSMH	283	TTVLLDEAVNGLNIRPDGIYIDGTFGRGGHSRLILSQLG EEGRLLAIDRDPQAIAVAKTIDDPRFSIIHGPFSALGEY VAERDLIGKIDGILLDLGVSSPQLDDAERGFSFMRDGP LDMRMDPTRGQSAAEWLQTAEEADIAWVLKTYGEER FAKRIARAIVERNREQPMTRTKELAEVVAAATPVKHP ATRTFQAVRIWVNSELEEIEQALKSSLNVLAPGGRLSII SFHSLEDRIVKRFMRENSRGRQLRALGKLMPGEEEVA ENPRARSSVLRIAERTNA	3TKA
PDE5A1-IBMX	311	EETRELQSLAAAVVPSAQTLKITDFSFSDFELSDLETAL CTIRMFTDLNLVQNFQMKHEVLCRWILSVKKNYRKN VAYHNWRHAFNTAQCMFAALKAGKIQNKLTDLEILA LLIAALSHDLDHRGVNNSYIQRSEHPLAQLYCHSIMEH HHFDQCLMILNSPGNQILSGLSIEEYKTTLKIIKQAILAT DLALYIKRRGEFFELIRKNQFNLEDPHQKELFLAMLMT ACDLSAITKPWPIQQRLAELVATEFFDQGDREKKNKIP SMQVGFIDAICLQLYEALTHVSEDCFPLLDGCRKNRQ KWQALAEQQ	1RKP
PERAKINE REDUCTASE,	311	MPRVKLGTQGLEVSKLGFGCMGLSGLPEEQGIAVIKE AFNCGITFFDTSDIYGENGSNEELLGKALKQLPREKIQV GTKFGIHEIGFSGVKAKGTPDYVRSCCEASLKRLDVDY IDLFYIHRIDTTVPIEITMGELKLVEEGKIKYVGLSEASP DTIRRAHAVHPVTALQIEYSLWTRDIEDEIVPLCRQLGI GIVPYSPIGRGLFAGKAIKESKNKQIYYRIEALSQKHGC TPVQLALAWVLHQGEDVVPIPGTTKIKNLHNNVGALK	3UYI

		VKLTKEDLKEISDAVPLDEVAGESIHEVIAVTNWKFAN TPPL	
B1B2 DOMAINS	315	QCNVPLGMESGRIANEQISASSTYSDGRWTPQQSRLH GDDNGWTPNLDSNKEYLQVDLRFLTMLTAIATQGAIS RETQNGYYVKSYKLEVSTNGEDWMVYRHGKNHKVF QANNDATEVVLNKLHAPLLTRFVRIRPQTWHSGIALR LELFGCRVTDAPCSNMLGMLSGLIADSQISASSTQELW SPSAARLVSSRSGWFPRIPQAQPGEEWLQVDLGTPKTV KGVIIQGARGGAVEARAFVRKFKVSYSLNGKDWEYIQ DPRTQQPKLFEGNMHYDTPDIRRFDPIPAQYVRVYPER WSPAGIGMRLEVLGCDWT	2QQJ
MACA WILD-TYPE OXIDIZED	320	EDVMKRAQGLFKPIPAKPPVMKDNPASPSRVELGRML FFDPRLSASHLISCNTCHNVGLGGTDILETSIGHGWQK GPRNSPTVLNAVYNIAQFWDGRAEDLAAQAKGPVQA SVEMNNKPENLVATLKSIPGYPPLFRKAFPGQGDPVTF DNVAKAIEVFEATLVTPDAPFDKYLKGNRKAISSTAEQ GLALFLDKGCAACHSGVNMGGTGYFPFGVREDPGPV DDTGRYKVTSTAADKYVFRSPSLRNVAITMPYFHSGK VWKLKDAVKIMGSAQLGISITDADADKIVTFLNTLTG AQPKVMHPVLPPNSDDTPRPVSN	4AAL
MUSCLE FRUCTOSE-1,6-BISPHOSPHATASE E69Q MUTANT	326	TDMLTLTRYVMEKGRQAKGTGELTQLLNSMLTAIKAI SSAVRKAGLAHLYGIAGSVNVDQVKKLDVLSNSLVIN MLQSSYSTCVLVSEENKDAIITAKEKRGKYVVCFDPLD GSSNIDCLASIGTIFAIYRKTSDEPSEKDALQCGRNIVA AGYALYGSATLVALSTGQGVDLFMLDPALGEFVLVE KDVKIKKKGKIYSLNEGYAKYFDAATTEYVQKKKFPE DGSAPYGARYVGSMVADVHRTLVYGGIFLYPANQKS PKGKLRLLYECNPVAYIIEQAGGLATTGTQPVLDVKPE AIHQRVPLILGSPEDVQEYLTCVQKNQA	3IFA
PRE-REACTIVE STATE OF PORCINE OAS	349	MELRHTPARDLDKFIEDHLLPNTCFRTQVKEAIDIVCR FLKERCFQGTADPVRVSKVVKGGSSGKGTTLRGRSDA DLVVFLTKLTSFEDQLRRRGEFIQEIRRQLEACQREQK FKVTFEVQSPRRENPRALSFVLSSPQLQQEVEFDVLPA FDALGQWTPGYKPNPEIYVQLIKECKSRGKEGEFSTCF TELQRDFLRNRPTKLKSLIRLVKHWYQTCKKTHGNKL PPQYALELLTVYAWEQGSRKTDFSTAQGFQTVLELVL KHQKLCIFWEAYYDFTNPVVGRCMLQQLKKPRPVILD PADPTGNVGGGDTHSWQRLAQEARVWLGYPCCKNL DGSLVGAWTMLQKI	4RWN
PHOSPHATIDYL MANNOSYLTRANSFERASE PIMA	359	MRIGMVCPYSFDVPGGVQSHVLQLAEVLRDAGHEVS VLAPASPHVKLPDYVVSGGKAVPIPYNGSVARLRFGP ATHRKVKKWIAEGDFDVLHIHEPNAPSLSMLALQAAE GPIVATFHTSTTKSLTLSVFQGILRPYHEKIIGRIAVSAV EIPNGVDVASFADAPLLDGYPREGRTVLFLGRYDEPRK GMAVLLAALPKLVARFPDVEILIVGRGDEDELREQAG DLAGHLRFLGQVDDATKASAMRSADVYCAPHLGGES FGIVLVEAMAAGTAVVASDLDAFRRVLADGDAGRLV PVDDADGMAAALIGILEDDQLRAGYVARASERVHRY DWSVVSAQIMRVYETVSGAGIKVQVS	4N9W
PROPLASMEPSIN	375	TEHLTLAFKIERPYDKVLKTISKKNLKNYIKETFNFFKS GYMKQNYLGSENDVIELDDVANIMFYGEGEVGDNHQ KFMLIFDTGSANLWVPSKKCNSSGCSIKNLYDSSKSKS YEKDGTKVDITYGSGTVKGFFSKDLVTLGHLSMPYKFI EVTDTDDLEPIYSSVEFDGILGLGWKDLSIGSIDPIVVEL KNQNKIDNALFTFYLPVHDVHAGYLTIGGIEEKFYEGN ITYEKLNHDLYWQIDLDVHFGKQTMEKANVIVDSGTT TITAPSEFLNKFFANLNVIKVPFLPFYVTTCDNKEMPTL EFKSANNTYTLEPEYYMNPILEVDDTLCMITMLPVDID SNTFILGDPFMRKYFTVFDYDKESVGFAIAKN	1MIQ
OVALBUMIN MUTANT R339T	381	GSIGAASMEFCFDVFKELKVHHANENIFYCPIAIMSAL AMVYLGAKDSTRTQINKVVRFDKLPGFGDSIEAQCGT SVNVHSSLRDILNQITKNDVYSFSLASRLYAEERYPILP	1JTI

		EYLQCVKELYRGGLEPINFQTAADQARELINSWVESQ TNGIIRNVLQPSSVDSQTAMVLVNAIVFKGLWEKTFK DEDTQAMPFRVTEQESKPVQMMYQIGLFRVASMASE KMKILELPFASGTMSMLVLLPDEVSGLEQLESIINFEKL TEWTSSNVMEERKIKVYLPRMKMEEKYNLTSVLMAM GITDVFSSSANLSGISSAESLKISQAVHAAHAEINEAGT EVVGSAEAGVDAAEEFRADHPFLFCIKHIATNAVLFFG RCVSP	
HPIN1 WW DOMAIN (5-39)	402	MEKLPPGWEKRMSRSSGRVYYFNHITNASQWERPSG KIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTV EHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQSG LLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEA LSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFN LQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNA GAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGETA MTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVG VLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDK PLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQ MSAFWYAVRTAVINAASGRQTVDEALKDAQT	5B3Z
GLYCOPROTEIN G ECTODOMAIN	413	KFTIVFPHNQKGNWKNVPSNYHYCPSSSDLNWHNDLI GTALQVKMPKSHKAIQADGWMCHASKWVTTCDFRW YGPKYITHSIRSFTPSVEQCKESIEQTKQGTWLNPGFPP QSCGYATVTDAEAVIVQVTPHHVLVDEYTGEWVDSQ FINGKCSNYICPTVHNSTTWHSDYKVKGLCDSNLISMD ITFFSEDGELSSLGKEGTGFRSNYFAYETGGKACKMQY CKHWGVRLPSGVWFEMADKDLFAAARFPECPEGSSIS APSQTSVDVSLIQDVERILDYSLCQETWSKIRAGLPISP VDLSYLAPKNPGTGPAFTIINGTLKYFETRYIRVDIAAPI LSRMVGMISGTTTERELWDDWAPYEDVEIGPNGVLRT SSGYKFPLYMIGHGMLDSDLHLSSKAQVFEHPHIQDA	5I2M
EDTA TREATED	421	YKPSGNKRVTFKDVGGAEEAIEELKEVVEFLKDPSKF NRIGARMPKGILLVGPPGTGKTLLARAVAGEANVPFF HISGSDFVELFVGVGAARVRDLFAQAKAHAPCIVFIDE IDAVGREQTLNQLLVEMDGFDSKEGIIVMAATNRPDIL DPALLRPGRFDKKIVVDPPDMLGRKKILEIHTRNKPLA EDVNLEIIAKRTPGFVGADLENLVNEAALLAAREGRD KITMKDFEEAIDRVILISPAEKRIIAYHEAGHAVVSTVV PNGEPVHRISIIPRGYKALGYTLHLPEEDKYLVSRNELL DKLTALLGGRAAEEVVFGDVTSGAANDIERATEIARN MVCQLGMSEELGPLAWGRNYSEEVASKIDEEVKKIVT NCYERAKEIIRKYRKQLDNIVEILLEKETIEGDELRRILS EEFE	2CE7
TMH1-LOCK MUTANT	465	DTLNDVIQDPTRRNKLINDNNLLKGIIMGRDGPVPSSR ELIVRPDTLRAIINNRATIETTTMEAEFTETLMESNYNS ASVKVSAPCITANSEYSESSSFKNTETEKSMYTSSRYLF PQGRIDFTTPDSGDVIKLSPQFTSGVQAALAKATGTEK REALQNLFQEYGCVFRTKVHIGGVLSAHTMETFSRSE NETEVKQDVKAGLEGAVKGWGGGATAGHGNTQGTI TTSQNRKLNVKYIVNGGDYTKIQNTEEWVASTNQSEH WRVIEVTEVTAVADLLPQPIRGQVKDLLKPLLGKWVD VEKVPGLESLPVSVYRPKGAIPAGWFWLGDTADASKA LLVKPTLPARSGRNPALTSLHQGSGMTEQPFVDLPQY QYLSTYFGSFAHDTPPGSTLRGLRPDHVLPGRYEMHG DTISTAVYVTRPVDVPFPEDEAFDLKSLVRVKLPGSGN PPKPRSALKKSMVLFD	4OV8
PNEUMOLYSIN D168A MUTANT.	487	AHHHHHHSSGLVPRGSMANKAVNDFILAMNYDKKKL LTHQGESIENRFIKEGNQLPDEFVVIERKKRSLSTNTSDI SVTATNDSRLYPGALLVVDETLLENNPTLLAVDRAPM TYSIDLPGLASSDSFLQVEDPSNSSVRGAVNDLLAKWH QDYGQVNNVPARMQYEKITAHSMEQLKVKFGSAFEK TGNSLDIDFNSVHSGEKQIQIVNFKQIYYTVSVDAVKN PGDVFQDTVTVEDLKQRGISAERPLVYISSVAYGRQV	5AOE

		YLKLETTSKSDEVEAAFEALIKGVKVAPQTEWKQILD NTEVKAVILGGDPSSGARVVTGKVDMVEDLIQEGSRF TADHPGLPISYTTSFLRDNVVATFQNSTDYVETKVTAY RNGDLLLDHSGAYVAQYYITWDELSYDHQGKEVLTP KAWDRNGQDLTAHFTTSIPLKGNVRNLSVKIRECTGL AWEWWRTVYEKTDLPLVRKRTISIWGTTLYPQVEDK VEND	
DIPHTHERIA TOXIN MUTANT CRM197	499	VVDSSKSFVMENFSSYHGTKPGYVDSIQKGIQKWKEF YSTDNKYDAAGYSVDNENPLSGKAGGVVKVTYPGLT KVLALKVDNAETIKKELGLSLTEPLMEQVGTEEFIKRF GDGASRVVLSLPFAEGSSSVEYINNWEQAKALSVELEI NFETRGKRGQDAMYEYMAQACACINLDWDVIRDKTK TKIESLKEHGPIKNKMSESPNKTVSEEKAKQYLEEFHQ TALEHPELSELKTVTGTNPVFAGANYAAWAVNVAQVI DSETADNLEKTTAALSILPGIGSVMGIADGAVHHNTEE IVAQSIALSSLMVAQAIPLVGIGFAAYNFVESIINLFQV VHNSYNRPAYSPGHKTQPFLHDGYAVSWNTVEDSIIR TGFQGESGHDIKITAENTPLPIAGVLLPTIPGKLDVNKS KTHISVNGRKIRMRCRAIDGDVTFCRPKSPVYVGNGV HANLHVAFHRSSSEKIHSNEISSDSIGVLGYQKHTKVN SKLSLFFEIKS	4AE0
HIV-1 REVERSE TRANSCRIPTASE	552	PISPIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICT EMEKEGKISKIGPENPYNTPVFAIKKKDSTKWRKLVDF RELNKRTQDFWEVQLGIPHPAGLKKKKSVTVLDVGD AYFSVPLDEDFRKYTAFTIPSINNETPGIRYQYNVLPQG WKGSPAIFQSSMTKILEPFRKQNPDIVIYQYMDDLYVG SDLEIGQHRTKIEELRQHLLRWGLTTPDKKHQKEPPFL WMGYELHPDKWTVQPIVLPEKDSWTVNDIQKLVGKL NWASQIYPGIKVRQLCKLLRGTKALTEVIPLTEEAELE LAENREILKEPVHGVYYDPSKDLIAEIQKQGQGQWTY QIYQEPFKNLKTGKYARMRGAHTNDVKQLTEAVQKIT TESIVIWGKTPKFKLPIQKETWETWWTEYWQATWIPE WEFVNTPPLVKLWYQLEKEPIVGAETFYVDGAANRET KLGKAGYVTNRGRQKVVTLTDTTNQKTELQAIYLAL QDSGLEVNIVTDSQYALGIIQAQPDQSESELVNQIIEQLI KKEKVYLAWVPAHKGIGGNEQVDKLV	3MEE
VIBRIO CHOLERAE CYTOLYSIN (HLYA) PRO-TOXIN	663	AIKYYNAADWQALPSLAELRDLVINQQKRVLVDFSQI SDAEGQAEMQAQFRKAYGVGFANQFIVITEHKGELLF TPFDRTEETNTLPHVAFYISVNRAISDEECTFNNSWLW KNEKGSRPFCKDANISLIYRVNLERSLQYGIVGSATPD AKIVRISLDDDSTGAGIHLNDQLGYRQFGASYTTLDAY FREWSTDAIAQDYRFVFNASNNKAQILKTFPVDNINEK FERKEVSGFELGVTGGVEVSGDGPKAKLEARASYTQS RWLTYNTQDYRIERNAKNAQAVSFTWNRQQYATAES LLNRSTDALWVNTYPVDVNRISPLSYASFVPKMDVIY KASATETGSTDFIIDSSVNIRPIYNGAYKHYYVVGAHQ SYHGFEDTPRRRITKSASFTVDWDHPVFTGGRPVNLQL ASFNNRCIQVDAQGRLTANMCDSQQSAQSFIYDQLGR YVSASNTKLCLDGAALDALQPCNQNLTQRWEWRKGT DELTNVYSGESLGHDKQTGELGLYASSNDAVSLRTIT AYTDVFNAQESSPILGYTQGKMNQQRVGQDNRLYVR AGAAIDALGSASDLLVGGNGGSLSSVDLSGVKSITATS GDFQYGGQQLVALTFTYQDGRQQTVGSKAYVTNAHE DRFDLPDAAKITQLKIWADDWLVKGVQFDLN	1XEZ

MEMBRANE PROTIEN DATA SET

NAME	N	SEQUENCE	PDB
AUTOPHAGIC SNARE COMPLE	64	DRVRNLQSEVEGVKNIMTQNVERILARGENLEHLRNK TEDLEATSEHFKTTSQKVARKFWWKNV	4WY4
SELENOPROTEIN S (VCP-INTERACTING MEMBRANE PROTEIN)	69	GSARLRALRQRQLDRAAAAVEPDVVVKRQEALAAAR LKQEELNAQVEKHKEKLKQLEEEKRRQKIEWDS	2Q2F
HCNK2-SAM/DHYP-SAM COMPLEX	74	EPVSKWSPSQVVDWKGLDDCLQQYIKNFEREKISGDQ LLRITHQELEDLGVSRIGHQELILEAVDLLCALNYGL	3BS5
HUMAN CD59	78	LQCYNCPNPTADCKTAVNCSSDFDACLITKAGLQVYN KCWKFEHCNFNDVTTRLRENELTYYCCKKDLCNFNE QLENC	2UWR
EXTRACELLULAR DOMAIN OF HUMAN RAMP2	78	VKNYETAVQFCWNHYKDQMDPIEKDWCDWAMISRP YSTLRDCLEHFAELFDLGFPNPLAERIIFETHQIHFANCS LVQ	2XVT
SYNTENIN PDZ2	82	GAMDPRTITMHKDSTGHVGFIFKNGKITSIVKDSSAAR NGLLTEHNICEINGQNVIGLKDSQIADILSTSGTVVTITI MPAF	1R6J
HUMAN SYNCYTIN 1	89	QFYYKLSQELNGDMERVADSLVTLQDQLNSLAAVVL QNRRALDLLTAERGGTCLFLGEECCYYVNQSGIVTEK VKEIRDRIQRRAEELR	6RX1
HUMAN SYNCYTIN 2	89	TYSQLSKEIANNIDTMAKALTTMQEQIDSLAAVVLQN RRGLDMLTAAQGGICLALDEKCCFWVNQSGKVQDNI RQLLNQASSLRERATQ	6RX3
T-CELL SURFACE GLYCOPROTEIN CD3 EPSILON CHAIN	91	QTPYKVSISGTTVILTCPQYPGSEILWQHNDKNIGGDE DDKNIGSDEDHLSLKEFSELEQSGYYVCYPRGSKPEDA NFYLYLRARVCENCM	1XIW
GLUTAMATE RECEPTOR INTERACTING PROTEIN-1	94	SRTVEVTLHKEGNTFGFVIRGGAHDDRSRPVVITSVRP GGPADREGTIKPGDRLLSVDGIRLLGTTHAEAMSILKQ CGQEAALLIEYDVSETAV	2JIL
COILED-COIL DOMAIN-CONTAINING PROTEIN 90B	94	DTHALVQDLETHGFDKTQAETIVSALTALSNVSLDTIY KEMVTQAQQEITVQQLMAHLDAIRKDMKQLEWKVEE LLSKVYHLENEVARLKKLVG	6H9M
PATCHED-1 ECTODOMAIN 2 (PTCH1-ECD2) IN COMPLEX WITH NANOBODY 75	94	KMWLHYFRDWLQGLQDAFDSDWETGKIMPNNYKNG SDDGVLAYKLLVQTGSRDKPIDISQLTKQRLVDADGII NPSAFYIYLTAWVSNDPVAYA	6RVC
PLK1 POLO-BOX DOMAIN IN COMPLEX WITH PL-2	96	EGVLYKWTNYLTGWQPRWFVLDNGILSYYDSQDKGS KGSIKAVCEIKVHSADNTRELIIPGEQHFYKAVNAAER QRWLVALGSSKASLTDTRLVPR	4RCP
HUMAN DISCS LARGE 1 PDZ2	97	RKPVSEKIMEIKLIKGPKGLGFSIAGGVGNQHIPGDNSI YVTKIIEGGAAHKDGKLQIGDKLLAVNNVCLEEVTHE EAVTALKNTSDFVYLKVAKPT	4G69
BETA-NERVE GROWTH FACTOR	99	EFSVCDSVSVWVGDKTTATDIKGKEVMVLGEVNINNS VFKQYFFETKCRDGCRGIDSKHWNSYCTTTHTFVKAL TMDGKQAAWRFIRIDTACVCVLSRK	1SG1
PDZ DOMAIN OF SYNAPTOJANIN-2 BINDING PROTEIN	100	SDYLVTEEEINLTRGPSGLGFNIVGGTDQQYVSNDSGI YVSRIKENGAAALDGRLQEGDKILSVNGQDLKNLLHQ DAVDLFRNAGYAVSLRVQHRLESSI	2JIK
GLUCAGON-LIKE PEPTIDE-1	100	TVSLWETVQKWREYRRQCQRSLTEDPPPATDLFCNRT FDEYACWPDGEPGSFVNVSCPWYLPWASSVPQGHVY RFCTAEGLWLQKDNSSLPWRDLSECEE	5E+94
PLEXIN A2 RBD	102	LIRQQIEYKTLILNCVNPDNENSPEIPVKVLNCDTITQV KEKILDAVYKQRPRAVDMDLEWRQGRIARVVLQDED ITTKIKRLNTLMHYQVSDRSVVALVPK	3Q3J

N-TERMINAL BETA-SHEET	107	AGAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYREN MHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTV TTTTKGENFTETDVKMMERVVEQMCITQYERESQA	4N9O
INNER DYSF DOMAIN OF HUMAN DYSFERLIN	109	DAGHLSFVEEVFENQTRLPGGQWIYMSDNYTDVNGE KVLPKDDIECPLGWKWEDEEWSTDLNRAVDEQGWEY SITIPPERKPKHWVPAEKMYYTHRRRRWVRLRRRDLS	4CAI
PGRMC1 CYTOCHROME B5-LIKE DOMAIN	112	SPEFDFTPAELRRFDGVQDPRILMAINGKVFDVTKGRK FYGPEGPYGVFAGRDASRGLATFCLDKEALKDEYDDL SDLTAAQQETLSDWESQFTFKYHHVGKLLKEGEEPTV	4X8Y
DISEASE MUTANT OF THE VOLTAGE-GATED SODIUM CHANNEL BETA 4	115	KATDIYAVNGTEILLPCTFSSAFGFEDLHFRWTYNSSD AFKILTEGTVKNEKSDPKVTLKDDDRITLVGSKMNNIS IVLRDLEFSDTGKYTCHVKNPKENNLQHHATIFLQVV DR	6VSV
PLECKSTRIN HOMOLOGY DOMAIN OF HUMAN SIN1	116	GAMATVQDMLSSHHYKSFKVSMIHRLRFTTDVQLGIS GDKVEIDPVTKFWIKQKPISIDSDLLCACDLAEEKSPSH AIFKLTYLSNHDYKHLYFESDAATVNEIVLKVNYILES RA	3VOQ
MPZL1 MUTANT	119	LEVYTPKEIFVANGTQGKLTCKFKSTTGGLTSVSWSFQ PEGADTTVGFFHYSQGQVYLGNYPPFKDRISWAGDLD KKDASINIENMQFIHNGTYICDVKNPPDIVGKTSHIRLY VVEKE	6IGW
IG V-SET DOMAIN OF HUMAN PAIRED IMMUNOGLOBULIN-LIKE TYPE 2 RECEPTOR ALPHA	120	MLYGVTQPKHLSASMGGSVEIPFSFYYPWELATAPDV RISWRRGHFHGQSFYSTRPPSIHKDYVNRLFLNWTEGQ KSGFLRISNLQKQDQSVYFCRVELDTRSSGRQQWQSIE GTKLSIT	3WUZ
VOLTAGE-GATED SODIUM CHANNEL BETA 2 SUBUNIT EXTRACELLULAR DOMAIN	122	SNAMEVTVPATLNVLNGSDARLPCTFNSAYTVNHKQF SLNWTYQECNNCSEEMFLQFRMKIINLKLERFQDRVE FSGNPSKYDVSVMLRNVQPEDEGIYNCYIMNPPDRHR GHGKIHLQVLM	5FEB
EEA1 HOMODIMER OF C-TERMINAL FYVE DOMAIN	123	QDERRALLERCLKGEGEIEKLQTKVLELQRKLDNTTA AVQELGRENQSLQIKHTQALNRKWAEDNEVQNCMAC GKGFSVTVRRHHCRQCGNIFCAECSAKNALTPSSKKP VRVCDACFNDLQG	1JOC
DYSFERLIN C2A VARIANT 1 (C2AV1)	127	IHMLACLLVRASNLPSAKKDRRSDPVASLTFRGVKKR TKVIKNSVNPVWNEGFEWDLKGIPLDQGSELHVVVKD HETMGRNRFLGEAKVPLREVLATPSLSASFNAPLLDTK KQPTGASLVLQVSYT	4IQH
PERIPHERAL MEMBRANE PROTEIN P2 FROM HUMAN MYELIN	132	GSNKFLGTWKLVSSENFDDYMKALGVGLATRKLGNL AKPTVIISKKGDIITIRTESTFKNTEISFKLGQEFEETTAD NRKTKSIVTLQRGSLNQVQRWDGKETTIKRKLVNGK MVAECKMKGVVCTRIYEKV	4BVM
FAM134B LIR FUSED TO HUMAN GABARAP	132	DDFELLDQSELDQIESMKSTYKEEHPFEKRRSEGEKIR KKYPDRVPVIVEKAPKARIGDLDKKKYLVPSDLTVGQ FYFLIRKRIHLRAEDALFFFVNNVIPPTSATMGQLYQEH HEEDFFLYIAYSDESVYG	7BRQ
SLMAP FHA DOMAIN IN COMPLEX WITH PMST2	134	PSALAIFTCRPNSHPFQERHVYLDEPIKIGRSVARCRPA QNNATFDCKVLSRNHALVWFDHKTGKFYLQDTKSSN GTFINSQRLSRGSEESPPCEILSGDIIQFGVDVTENTRKV THGCIVSTIKLFLPDGMEA	6AR2
MPGES-1	147	SLVMSSPALPAFLLCSTLLVIKMYVVAIITGQVRLRKK AFANPEDALRHGGPQYRSDPDVERCLRAHRNDMETIY PFLFLGFVYSFLGPNPFVAWMHFLVFLVGRVAHTVAY LGKLRAPIRSVTYTLAQLPCASMALQILWEAARHL	5TL9
LTC4S IN COMPLEX WITH AZ13690257	148	HHHHHKDEVALLAAVTLLGVLLQAYFSLQVISARRAF RVSPPLTTGPPEFERVYRAQVNCSEYFPLFLATLWVAG IFFHEGAAALCGLVYLFARLRYFQGYARSAQLRLAPL YASARALWLLVALAALGLLAHFLPAALRAALLGRLR	6R7D

ADP-RIBOSYLATION FACTOR BINDING PROTEIN GGA1	151	HHHHMELSLASITVPLESIKPSNILPVTVYDQHGFRILF HFARDPLPGRSDVLVVVVSMLSTAPQPIRNIVFQSAVP KVMKVKLQPPSGMELPAFNPIVHPSAITQVLLLANPQK EKVRLRYKLTFTMGDQTYNEMGDVDQFPPPETWGSL	1NA8
P-SELECTIN LECTIN/EGF DOMAINS	158	WTYHYSTKAYSWNISRKYCQNRYTDLVAIQNKNEIDY LNKVLPYYSSYYWIGIRKNNKTWTWVGTKKALTNEA ENWADNEPNNKRNNEDCVEIYIKSPSAPGKWNDEHCL KKKHALCYTASCQDMSCSKQGECLETIGNYTCSCYPG FYGPECEYVRD	1G1S
P115RHOGEF RGS DOMAIN	165	SIIGAEDEDFEFQSLEQVKRRPAHLMALLQHVALQFEP GPLLCCLHADMLGAKKAFLDFYHSFLEKTAVLRVPVP FVQEVVQSQQVAVGRQLEDFRSKRLMGMTPWEQELA QLERASYEARERHVAERLLMHLEEMQHTISTDEEKSA AVVNAIGLYMRHLGVRT	3AB3
C-TERMINAL PGN-BINDING DOMAIN	167	VCPNIIKRSAWEARETHCPKMNLPAKYVIIIHTAGTSCT VSTDCQTVVRNIQSFHMDTRNFCDIGYHFLVGQDGGV YEGVGWHIQGSHTYGFNDIALGIAFIGYFVEKPPNAAA LEAAQDLIQCAVVEGYLTPNYLLMGHSDVVNILSPGQ ALYNIISTWPHFKHAK	1TWQ
HUMAN FC GAMMA RECEPTOR	171	APPKAVLKLEPPWINVLQEDSVTLTCQGARSPESDSIQ WFHNGNLIPTHTQPSYRFKANNNDSGEYTCQTGQTSL SDPVHLTVLFEWLVLQTPHLEFQEGETIMLRCHSWKD KPLVKVTFFQNGKSQKFSHLDPTFSIPQANHSHSGDYH CTGNIGYTLFSSKPVTITVQV	1FCG
STING BOUND TO C-DI-GMP	173	SVAHGLAWSYYIGYLRLILPELQARIRTYNQHYNNLLR GAVSQRLYILLPLDCGVPDNLSADPNIRFLDKLPQDRV YSNSIYELLENGQRAGTCVLEYATPLQTLFASQYSQAG FSREDRLEQAKLFCRTLEDILADAPESQNNCRLIAYQE PADDSSFSLSQEVLRHLRQEE	4EMT
TRPML2 ELD	173	AFKEDNTVAFKHLFLKGYSGTDEDDYSCSVYTQEDAY ESIFFAINQYHQLKDITLGTLGYGENEDNRIGLKVCKQ HYKKGNDVELDCVQLDLQDLSKKPPDWKNSSFFRLEF YRLLQVEISFHLKGIDLQTPDCYVFQNTIIFDNKAHSGK IKIYFDSDAKIEECKDLNIFGS	6HRR
MACA WILD-TYPE FULLY REDUCED	176	LDMREIPKSSIKPEHFHLMYLLEQHSPYFIDAELTELRD SFQIHYDINDNHTPFDNIKSFTKNEKLRYLLNIKNLEEV NRTRYTFVLAPDELFFTRDGLPIAKTRGLQNVVDPLPV SEAEFLTRYKALVICAFNEKQSFDALVEGNLELHKGTP FETKVIEAATLDLLTAFLDEQY	4AAN
VON WILLEBRAND FACTOR A DOMAIN	177	RAFDLYFVLDKSGSVANNWIEIYNFVQQLAERFVSPE MRLSFIVFSSQATIILPLTGDRGKISKGLEDLKRVSPVG ETYIHEGLKLANEQIQKAGGLKTSSIIIALTDGKLDGLV PSYAEKEAKISRSLGASVYCVGVLDFEQAQLERIADSK EQVFPVKGGFQALKGIINSILAQS	1SHT
HLA DR52C	178	EEHVIIQAEFYLNPDQSGEFMFDFDGDEIFHVDMAKKE TVWRLEEFGRFASFEAQGALANIAVDKANLEIMTKRS NYTPITNVPPEVTVLTNSPVELREPNVLICFIDKFTPPVV NVTWLRNGKPVTTGVSETVFLPREDHLFRKFHYLPFLP STEDVYDCRVEHWGLDEPLLKHWEF	3C5J
KINASE DOMAIN OF MPP1/P55	180	FQGRKTLVLIGASGVGRSHIKNALLSQNPEKFVYPVPY TTRPPRKSEEDGKEYHFISTEEMTRNISANEFLEFGSYQ GNMFGTKFETVHQIHKQNKIAILDIEPQTLKIVRTAELS PFIVFIAPTDQGTQTEALQQLQKDSEAIRSQYAHYFDLS LVNNGVDETLKKLQEAFDQACSSPQ	3NEY
VON WILLEBRAND FACTOR A DOMAIN OF HUMAN CAPILLARY MORPHOGENESIS PROTEIN 2	181	SCRRAFDLYFVLDKSGSVANNWIEIYNFVQQLAERFVS PEMRLSFIVFSSQATIILPLTGDRGKISKGLEDLKRVSPV GETYIHEGLKLANEQIQKAGGLKTSSIIIALTDGKLDGL VPSYAEKEAKISRSLGASVYCVGVLDFEQAQLERIADS KEQVFPVKGGFQALKGIINSILAQSC	1SHU

PRY-SPRY DOMAIN OF HUMAN TRIM72	193	RKMFRALMPALEELTFDPSSAHPSLVVSSSGRRVECSE QKAPPAGEDPRQFDKAVAVVAHQQLSEGEHYWEVDV GDKPRWALGVIAAEAPRRGRLHAVPSQGLWLLGLRE GKILEAHVEAKEPRALRSPERRPTRIGLYLSFGDGVLSF YDASDADALVPLFAFHERLPRPVYPFFDVCWHDKGK NAQPLLLV	3KB5
GLIOMA PATHOGENESIS-RELATED PROTEIN 1	193	ANILPDIENEDFIKDCVRIHNKFRSEVKPTASDMLYMT WDPALAQIAKAWASNCQFSHNTRLKPPHKLHPNFTSL GENIWTGSVPIFSVSSAITNWYDEIQDYDFKTRICKKVC GHYTQVVWADSYKVGCAVQFCPKVSGFDALSNGAHF ICNYGPGGNYPTWPYKRGATCSACPNNDKCLDNLCV NRQRDQV	3Q2U
BECLIN 1	195	KSVENQMRYAQTQLDKLVFNATFHIWHSGQFGTINNF RLGRLPSVPVEWNEINAAWGQTVLLLHALANKMGLK FQRYRLVPYGNHSYLESLTDKSKELPLYCSGGLRFFW DNKFDHAMVAFLDCVQQFKEEVEKGTRFCLPYRMDV EKGKIEDTGGSGGSYSIKTQFNSEEQWTKALKFMLTN LKWGLAWVSSQF	4DDP
GRASP55 GRASP DOMAIN (RESIDUES 7-208)	200	VEIPGGGTEGYHVLRVQENSPGHRAGLEPFFDFIVSING SRLNKDNDTLKDLLKANVEKPVKLIYSSKTLELRETSV TPSNLWGGQGLLGVSIRFCSFDGANENVWHVLEVESN SPAALAGLRPHSDYIIGADTVNESEDLFSLIETHEAKPL KLYVYNTDTDNCREVIITPNSAWGGEGSLGCGIGYGY LHRIPTRPFE	3RLE
FOLR1	204	RTELLNVCMNAKHHKEKPGPEDKLHEQCRPWRKNAC CSTNTSQEAHKDVSYLYRFNWNHCGEMAPACKRHFI QDTCLYECSPNLGPWIQQVDQSWRKERVLNVPLCKED CEQWWEDCRTSYTCKSNWHKGWNWTSGFNKCAVG AACQPFHFYFPTPTVLCNEIWTHSYKVSNYSRGSGRCI QMWFDPAQGNPNEEVARFYAAAM	4KM6
KDEL RECEPTOR	207	MNIFRLTGDLSHLAAIIILLLKIWKSRSCAGISGKSQLLF ALVFTTRYLDLFTSFISLYNTSMKLIYIACSYATVYLIY MKFKATYDGNHDTFRVEFLIVPVGGLSFLVNHDFSPLE ILWTFSIYLESVAILPQLFMISKTGEAETITTHYLFFLGL YRALYLVNWIWRYYFEGFFDLIAVVAGVVQTVLYCD FFYLYVTKVLKGKK	6I6H
ADHESION MOLECULE-1	209	VKPLQVEPPEPVVAVALGASRQLTCRLACADRGASVQ WRGLDTSLGAVQSDTGRSVLTVRNASLSAAGTRVCV GSCGGRTFQHTVQLLVYAFPNQLTVSPAALVPGDPEV ACTAHKVTPVDPNALSFSLLVGGQELEGAQALGPEVQ EEEEEPQGDEDVLFRVTERWRLPPLGTPVPPALYCQAT MRLPGLELSHRQAIPVLHSPTSPE	1BQS
OXIDISED CLIC1	213	GNCPFSQRLFMVLWLKGVTFNVTTVDTKRRTETVQK LCPGGQLPFLLYGTEVHTDTNKIEEFLEAVLCPPRYPK LAALNPESNTAGLDIFAKFSAYIKNSNPALNDNLEKGL LKALKVLDNYLTSPLPEEVDETSAEDEGVSQRKFLDG NELTLADCNLLPKLHIVQVVCKKYRGFTIPEAFRGVHR YLSNAYAREEFASTCPDDEEIELAYE	1RK4
ADVANCED GLYCOSYLATION END PRODUCT-SPECIFIC RECEPTOR	219	MAQNITARIGEPLVLKCKGAPKKPPQRLEWKLNTGRT EAWKVLSPQGGGPWDSVARVLPNGSLFLPAVGIQDEG IFRCQAMNRNGKETKSNYRVRVYQIPGKPEIVDSASEL TAGVPNKVGTCVSEGSYPAGTLSWHLDGKPLVPNEK GVSVKEQTRRHPETGLFTLQSELMVTPARGGDPRPTFS CSFSPGLPRHRALRTAPIQPRVWEPVPLEEVQL	3CJJ
AQUAPORIN 4	223	QAFWKAVTAEFLAMLIFVLLSLGSTINWGGTEKPLPV DMVLISLCFGLSIATMVQCFGHISGGHINPAVTVAMVC TRKISIAKSVFYIAAQCLGAIIGAGILYLVTPPSVVGGLG VTMVHGNLTAGHGLLVELIITFQLVFTIFASCDSKRTD VTGSIALAIGFSVAIGHLFAINYTGASMNPARSFGPAVI MGNWENHWIYWVGPIIGAVLAGGLYEYVFCP	3GD8
LAP1	224	VETTAVQEFQNQMNQLKNKYQGQDEKLWKRSQTFLE KHLNSSHPRSQPAILLLTAARDAEEALRCLSEQIADAY	4TVS

		SSFRSVRAIRIDGTDKATQDSDTVKLEVDQELSNGFKN GQNAAVVHRFESFPAGSTLIFYKYCDHENAAFKDVAL VLTVLLEEETLGTSLGLKEVEEKVRDFLKVKFTNSNTP NSYNHMDPDKLNGLWSRISHLVLPVQPENALKRGICL	
ANKYRIN REPEAT DOMAIN OF HUMAN TRPV2	244	PNRFDRDRLFNAVSRGVPEDLAGLPEYLSKTSKYLTDS EYTEGSTGKTCLKAVLNLKDGVNACILPLLQIDRDSGN PQPLVNAQCTDDYYRGHSALHIAIEKRSLQCVKLLVE NGANVHARACGRFFQKGGTCFYFGELPLSLAACTKQ WDVVSYLLENPHQPASLQATDSQGNTVLHALVISDNS AENIALVTSYDGLLQAGARLCPTVQLEDIRNLQDLTPL KLAAKEGKIEIFRHILQREF	2F37
HUMAN AQUAPORIN 7	247	MVREFLAEFMSTYVMMVFGLGSVAHMVLNKKYGSY LGVNLGFGFGVTMGVHVAGRISGAHMNAAVTFANCA LGRVPWRKFPVYVLGQFLGSFLAAATIYSLFYTAILHF SGGQLMVTGPVATAGIFATYLPDHMTLWRGFLNEAW LTGMLQLCLFAITDQENNPALPGTEALVIGILVVIIGVS LGMNTGYAINPSRDLPPRIFTFIAGWGKQVFSNGENW WWVPVVAPLLGAYLGGIIYLVFIGST	6QZI
GLUTAMATE RECEPTOR, IONOTROPIC KAINATE 1	256	ANRTLIVTTILEEPYVMYRKSDKPLYGNDRFEGYCLDL LKELSNILGFIYDVKLVPDGKYGAQNDKGEWNGMVK ELIDHRADLAVAPLTITYVREKVIDFSKPFMTLGISILY RKGTPIDSADDLAKQTKIEYGAVRDGSTMTFFKKSKIS TYEKMWAFMSSRQQTALVRNSDEGIQRVLTTDYALL MESTSIEYVTQRNCNLTQIGGLIDSKGYGVGTPIGSPYR DKITIAILQLQEEGKLHMMKEKWWRGNGCP	3FVO
MAIT TCR	262	RTHSLRYFRLGVSDPIGVPEFISVGYVDSHPITTYDSVT RQKEPRAPWMAENLAPDHWERYTQLLRGWQQMFKV ELKRLQRHYNHSGSHTYQRMIGCELLEDGSTTGFLQY AYDGQDFLIFNKDTLSWLAVDNVAHTIKQAWEANQH ELLYQKNWLEEECIAWLKRFLEYGKDTLQRTEPPLVR VNRKETFPGVTALFCKAHGFYPPEIYMTWMKNGEEIV QEIDYGDILPSGDGTYQAWASIELLYSCHVEHSGVHM VLQV	4L4V
F508A NBD1	267	STTEVVMENVTAFWEEGFGELFEKSFSNFSLLGTPVLK DINFKIERGQLLAVAGSTGAGKTSLLMMIMGELEPSEG KIKHSGRISFCSQFSWIMPGTIKENIIAGVSYDEYRYRS VIKACQLEEDISKFAEKDNIVLGEGGITLSGGQRARISL ARAVYKDADLYLLDSPFGYLDVLTEKEIFESCVCKLM ANKTRILVTSKMEHLKKADKILILHEGSSYFYGTFSEL QNLQPDFSSKLMGCDSFDQFSAERRNSILTETLRRFSL	1XMI
HLA-B46	274	GSHSMRYFYTAMSRPGRGEPRFIAVGYVDDTQFVRFD SDAASPRMAPRAPWIEQEGPEYWDRETQKYKRQAQT DRVSLRNLRGYYNQSEAGSHTLQRMYGCDVGPDGRL LRGHDQSAYDGKDYIALNEDLSSWTAADTAAQITQRK WEAAREAEQWRAYLEGLCVEWLRRYLENGKETLQR ADPPKTHVTHHPISDHEATLRCWALGFYPAEITLTWQ RDGEDQTQDTELVETRPAGDRTFQKWAAVVVPSGEE QRYTCHVQHEGLPKPLTLRW	4LCY
DDR2 KINASE DOMAIN	275	GPRVDFPRKLLTFKEKLGEGFGEVHLCEVEGMSANQP VLVAVKMLRADANKNARNDFLKEIKIMSRLKDPNIIH LLAVCITDDPLCMITEYMENGDLNQFLSRHQPTVSYTN LKFMATQIASGMKYLSSLNFVHRDLATRNCLVGKNYT IKIADFGMSRNLYSGDYYRIQGRAVLPIRWMSWESILL GKFTTASDVWAFGVTLWETFTFCQEQPYSQLSDEQVI ENTGEFFRDQGRQTYLPQPAICPDSVYELMLSCWRRD TKNRPSFQEIHRLL	7AYM
ANION EXCHANGER 1	276	VYVELQELVMDEKNQELRWMEAARWVQLEENLGEN GAWGRPHLSHLTFWSLLELRRVFTKGTVLLDLQETSL AGVANQLLDRFIFEDQIRPQDREELLRALLLKHSHAGE LEALGGVKPAVLTRPSQPLLPQHSSLETQLFCEEKIPPD SEATLVLVGRADFLEQPVLGFVRLQEAAELEAVELPVP IRFLFVLLGPEAPHIDYTQLGRAAATLMSERVFRIDAY	4KY9

		MAQSRGELLHSLEGFLDCSLVLPPTDAPSEQALLSLVP VQRELLRRRYQSS	
PATCHED-1 (PTCH1) ECTODOMAIN	277	EEAMFNPQLMIQTPKEEGANVLTTEALLQHLDSALQA SRVHVYMYNRQWKLEHLCYKSGELITETGYMDQIIEY LYPCLIITPLDCFWEGAKLQSGTAYLLGKPPLRWTNFD PLEFLEELKKINYQVDSWEEMLNKAEVGHGYMDRPC LNPADPDCPATAPNKNSTKPLDMALVLNGGCHGLSRK YMHWQEELIVGGTVKNSTGKLVSAHALQTMFQLMTP KQMYEHFKGYEYVSHINWNEDKAAAILEAWQRTYVE VVHQSVAQNSTQKVLSFTGT	6RTW
PLASMODIUM VIVAX DUFFY BINDING PROTEIN (PVDBP)	282	SNTVMKNCNYKRKRRERDWDCNTKKDVCIPDRRYQL CMKELTNLVITFRKLYLKRKLIYDAAVEGDLLLKLNN YRYNKDFCKDIRWSLGDFGDIIMGTDMEGIGYSKVVE NNLRSIFGTDEKAQQRRKQWWNESKAQIWTAMMYS VKKRLKICKLNVAVNIEPQIYRWIREWGRDYVSELPTE VQKLKEKCDGKINYTDKKVCKVPPCQNACKSYDQWI TRKKNQWDVLSNKFISVKNAEQTAGIVTPYDILKQEL DEFNEVAFENEINKRDGAYIELCVCS	4NUU
ADIPONECTIN RECEPTOR 2	282	EEGRWRVIPHDVLPDWLKDNDFLLHGHRPPMPSFRAC FKSIFRIHTETGNIWTHLLGCVFFLCLGIFYMFRPNISFV APLQEKVVFGLFFLGAILCLSFSWLFHTVYCHSEGVSR LFSKLDYSGIALLIMGSFVPWLYYSFYCNPQPCFIYLIVI CVLGIAAIIVSQWDMFATPQYRGVRAGVFLGLGLSGII PTLHYVISEGFLKAATIGQIGWLMLMASLYITGAALYA ARIPERFFPGKCDIWFHSHQLFHIFVVAGAFVHFHGVS NLQEFRFMIGGGC	5LX9
MOESIN FERM	289	TISVRVTTDAELEFAIQPNTTGKQLFDQVVKTIGLREV WFFGLQYQDTKGFSTWLKLNKKVTAQDVRKESPLLF KFRAKFYPEDVSEELIQDITQRLFFLQVKEGILNDDIYC PPETAVLLASYAVQSKYGDFNKEVHKSGYLAGDKLLP QRVLEQHKLNKDQWEERIQVWHEEHRGLREDAVLEY LKIAQDLEYGVNYFSIKNKKGSELWLGVDALGLNIYE QNDRLTPKIGFPWSEIRNISFNDKKFVIKPIDKKAPDFV FYAPRLRINKRILALCGNHELYRRRKP	1EF1
EXTENDED-SYNAPTOTAGMIN 2	292	NRITVPLVSEVQIAQLRFPVPKGVLRIHFIEAQDLQGKD TYLKGLVKGKSDPYGIIRVGNQIFQSRVIKENLSPKWN EVYEALVYEHPGQELEIELFDEDPDKDDFLGSLMIDLIE VEKERLLDEWFTLDEVPKGKLHLRLEWLTLMPNASNL DKVLTDIKADKDQANDGLSSALLILYLDSARNLPSGNP NPVVQMSVGHKAQESKIRYKTNEPVWEENFTFFIHNP KRQDLEVEVRDEQHQCSLGNLKVPLSQLLTSEDMTVS QRFQLSNSGPNSTIKMKIALRVLHLEK	4NPJ
OREXIN-1	301	YAWVLIAAYVAVFVVALVGNTLVCLAVWRNHHMRT VTNYFLVNLSLADVLATAICLPASLLVDITESWLFGHA LCKVIPYLQTVSVSVAVLTLSFIALDRWYAICHPLLFKS TARRALGSILGIWAVSLAIMVPQAAVMECSSVLPELAA RTRAFSVCDERWADDLAPKIYHSCFFIVTYLAPLGLM AMAYFQIFRKLWGRQIPGTTSEVKQMRARRKTAKML MVVVLVFALCYLPISVLNVLKRVFGMFRQASDREAVY AAFTFSHWLVYANSAANPIIYNFLSGKFREQFKAAFSW WLP	6TOD
CYSTEINYL LEUKOTRIENE RECEPTOR 2	365	RNCTIENFKREFFPIVYLIIFFVGVLGNGLSIYVFLQPYK KSTSVNVFMLNLAISNLLFISTLPFRADYYLRGSNWIFG DLACRIMSYSLYVNMYSSIYFLTVLSVVRYLAMVHPF RVTSIRSAWILCGIIWILIMASSIMLLDSEQNGSVTSCLE LNLYKIAKLQTMNYIALVVGCLLPFFTLSICYLLIIRVL LKVEADLEDNWETLNDNLKVIEKAAAQVKDALTKMR AAALDAQMKDFRHGFDILVGQIDDALKLVKEAQAAA EQLKTTRNAYIQKYLVSHRKALTTIIITLIIFFLCFLPYH TLRTVHLTTWKVGLCKDRLHKALVITLALAAANACF NPLLYYFAGENFKDRLKSALRK	6RZ6

DHODH	367	MATGDERFYAEHLMPTLQGLLDPESAHRLAVRFTSLG LLPRARFQDSDMLEVRVLGHKFRNPVGIAAGFDKHGE AVDGLYKMGFGFVEIGSVTPKPQEGNPRPRVFRLPED QAVINRYGFNSHGLSVVEHRLRARQQKQAKLTEDGLP LGVNLGKNKTSVDAAEDYAEGVRVLGPLADYLVVNV SSPNTAGLRSLQGKAELRRLLTKVLQERDGLRRVHRP AVLVKIAPDLTSQDKEDIASVVKELGIDGLIVTNTTVSR PAGLQGALRSETGGLSGKPLRDLSTQTIREMYALTQG RVPIIGVGGVSSGQDALEKIRAGASLVQLYTALTFWGP PVVGKVKRELEALLKEQGFGGVTDAIGADHRR	6IDJ
IQGAP1	369	ASKEKREKLEAYQHLFYLLQTNPTYLAKLIFQPQNKST KFDSVIFTLYNYASNQREEYLLLRLFKTALQEEIKSKV DQIQEIVTGNPTVIKVVSFNRGARGQNALRQILAPVVK EIDDKSLNIKTDPVDIYKSWVNQESQTGEASKLPYDVT PEQALAHEEVKTRLDSSIRNRAVTDKFLSAIVSSVDKIP YGRFIAKVLKDSLHEKFPDAGEDELLKIIGNLLYYRYN PAIVAPDAFDIIDLSAGGQLTTDQRRNLGSIAKLQHAA SNKFLGDNAHLSIINEYLSQSYQKFRRFFQTACDVPEL QDKFNVDEYSDLVTLTKPVIYISIGEIINTHTLLLDHQD AIAPEHNDPIHELLDDLGEVPTIES	3FAY
C5A	374	LEDNWETLNDNLKVIEKADNAAQVKDALTKMRAAA LDAKDFRHGFDILVGQIDDALKLANEGKVKEAQAAAE QLKTTRNAYSNTLRVPDILALVIFAVVFLVGVLGNALV VWVTAFEAKRTINAIWFLNLAVADFLSCLALPILFTSIV QHHHWPFGGAACSILPSLILLNMYASILLLATISADRFL LVFKPIWQNFRGAGLAWIACAVAWGLALLLTIPSFLY RVVREEYFPPKVLCGVDYSHDKRRERAVAIVRLVLGF LWPLLTLTICYTFILLRTWSRRSTKTLKVVVAVVASFFI FWLPYQVTGIMMSFLEPSSPTFLLLKKLDSLCVSFAYIN CCINPIIYVVAGQGFKSLPSLLRNVLTEESFPWR	6C1R
APO-FRIZZLED4	379	CHSVGTNSDQYIWVKRSLNCVLKCGYDAGLYSRSAK EFTDIWMAVWASLCFISTAFTVLTFLIDSSRFSYPERPII FLSMCYNIYSIAYIVRLTVGRERISCDFEEAAEPVLIQE GLKNTGCAIIFLLLYFFGMASSIWWVILTLTWFLAAGL KWGHEAIEMHSSYFHIAAWAIPAVKTIVILIMRLVDAD ELTGLCYVGNQNLDALTGFVVAPLFTYLVIGTLFIAAG LVALFKIRSNMKKYTCTVCGYIYNPEDGDPDNGVNPG TDFKDIPDDWVCPLCGVGKDQFEEVEEDKLERLMVKI GVFSVLYTVPATIVIACYFYEISNWALFRYSADDSNMA VEMLKIFMSLLVGITSGMWIWSAKTLHTWQKFYNRL VN	6BD4
M2 MUSCARINIC ACETYLCHOLINE RECEPTOR	384	SPYKTFEVVFIVLVAGSLSLVTIIGNILVMVSIKVNRHL QTVNNYFLFSLACADLIIGVFSMNLYTLYTVIGYWPLG PVVCDLWLALDYVVSNARVMNLLIISFDRYFCVTKPL TYPVKRTTKMAGMMIAAAWVLSFILWAPAILFWQFIV GVRTVEDGECYIQFFSNAAVTFGTAIAAFYLPVIIMTVL YWHISRASKADLEDNWETLNDNLKVIEKADNAAQVK DALTKMRAAALDAQKATPPKLEDKSPDSPEMKDFRH GFDILVGQIDDALKLANEGKVKEAQAAAEQLKTTRNA YIQKYLSREKKVTRTILAILLAFIITWAPYNVMVLINTF CAPCIPNTVWTIGYWLCYINSTINPACYALCNATFKKT FKHLLMCH	5ZKC
A2A ADENOSINE RECEPTOR A2AR-STAR2-BRIL	387	GPPIMGSSVYITVELAIAVLAILGNVLVCWAVWLNSNL QNVTNYFVVSLAAADILVGVLAIPFAITISTGFCAACH GCLFIACFVLVLAQSSIFSLLAIAIDRYIAIAIPLRYNGLV TGTRAAGIIAICWVLSFAIGLTPMLGWNNCGQPKEGK AHSQGCGEGQVACLFEDVVPMNYMVYFNFFACVLVP LLLMLGVYLRIFAAARRQLADLEDNWETLNDNLKVIE KADNAAQVKDALTKMRAAALDAQKMKDFRHGFDIL VGQIDDALKLANEGKVKEAQAAAEQLKTTRNAYIQK YLERARSTLQKEVHAAKSAAIIAGLFALCWLPLHIINCF	5IU4

		TFFCPDCSHAPLWLMYLAIVLAHTNSVVNPFIYAYRIR EFRQTFRKIIRS	
JAGN1	397	MSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDA TNGKLTLKFICTTGKLPVPWPTLVTTLVQCFSRYPDHM KRHDFFKSAMPEGYVQERTISFKDDGTYKTRAEVKFE GDTLVNRIELKGIDFKEDGNILGHKLEYNMASRQHRE RVAMHYQMSVTLKYEIKKLIYVHLVIWLLLVAKMSV GHLRLLSHDQVAMPYQWEYPYLLSILPSLLGLLSFPRN NISYLVLSMISMGLFSIAPLIYGSMEMFPAAQQLYRHG KAYRFLFGFSAVSIMYLVLVLAVQVHAWQLYYSKKL LDSWFTSTQEKKHKNSHNVYITADKQKNGIKANFKIR HNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQS VLSKDPNEKRDHMVLLEFVTAAGITH	6WVD
RHESUS GLYCOPROTEIN RHCG	403	SAWNTNLRWRLPLTCLLLQVIMVILFGVFVRYDFENE FYYRYPSFQDVHVMVFVGFGFLMTFLQRYGFSAVGFN FLLAAFGIQWALLMQGWFHFLQDRYIVVGVENLINAD FCVASVCVAFGAVLGKVSPIQLLIMTFFQVTLFAVNEFI LLNLLKVKDAGGSMTIHTFGAYFGLTVTRILYRRNLE QSKERQNSVYQSDLFAMIGTLFLWMYWPSFNSAISYH GDSQHRAAINTYCSLAACVLTSVAISSALHKKGKLDM VHIQNATLAGGVAVGTAAEMMLMPYGALIIGFVCGIIS TLGFVYLTPFLESRLHIQDTCGINNLHGIPGIIGGIVGAV TAASDWTARTQGKFQIYGLLVTLAMALMGGIIVGLIL RLPFWGQPSDENCFEDAVYWEMPEGNS	3HD6
MGLUR5	409	SPVQYLRWGDPAPIAAVVFACLGLLATLFVTVVFIIYR DTPVVKSSSRELCYIILAGICLGYLCTFLIAKPKQIYCYL QRIGIGLSPAMSYSALVTKTYRAARILAMSKKNIFEML RIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSEL DKAIGRNTNGVITKDEAEKLFNQDVDAAVRGILRNAK LKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLR MLQQKRWDEAAVNLAKSRWYNQTPNRAKRVITTFRT GTWDAYKISACAQLVIAFILICIQLGIIVALFIMEPPDIM EVYLICNTTNLGVVAPLGYNGLLILACTFYAFKTRNVP ANFNEAKYIAFTMYTTCIIWLAFVPIYFGSNYKIITMCF SVSLSATVALGCMFVPKVYIILAKPERN	6FFI
GLYPICAN-1 CORE	411	PASKSRSCGEVRQIYGAKGFSLSDVPQAEISGEHLRICP QGYTCCTSEMEENLANRSHAELETALRDSSRVLQAML ATQLRSFDDHFQHLLNDSERTLQATFPGAFGELYTQN ARAFRDLYSELRLYYRGANLHLEETLAEFWARLLERL FKQLHPQLLLPDGKQAEALRPFGEAPRELRLRATRAFV AARSFVQGLGVASDVVRKVAQVPLGPECSRAVMKLV YCAHCLGVPGARPCPDYCRNVLKGCLANQADLDAEW RNLLDSMVLITDKFWGTSGVESVIGSVHTWLAEAINA LQDNRDTLTAKVPRERPPSGTLEKLVSEAKAQLRDVQ DFWISLPGTLCSEKMADRCWNGMARGRYLPEVMGDG LANQINNPEVEVDITKPDMTIRQQIMQLKIMTNRLRSA YNG	4YWT
MITOFUSIN2 (MFN2)	428	VNASPLKHFVTAKKKINGIFEQLGAYIQESATFLEDTY RNAELDPVTTEEQVLDVKGYLSKVRGISEVLARRHMK VAFFGRTSNGKSTVINAMLWDKVLPSGIGHTTNCFLR VEGTDGHEAFLLTEGSEEKRSAKTVNQLAHALHQDK QLHAGSLVSVMWPNSKCPLLKDDLVLMDSPGIDVTTE LDSWIDKFCLDADVFVLVANSESTLMQTEKHFFHKVS ERLSRPNIFILNNRWDASASEPEYMEEVRRQHMERCTS FLVDELGVVDRSQAGDRIFFVSAKEVLNARIQKAQGM PEGGGALAEGFQVRMFEFQNFERRFEECISQSAVKTKF EQHTVRAKQIAEAVRLIMDSLHMAAREQQVVYCEEMR EERQDRTRENLEQEIAAMNKKIEVLDSLQSKAKLLRN KAGWLDSELNMFTHQYLQPS	6JFK
GPR52	441	VSERHSCPLGFGHYSVVDVCIFETVVIVLLTFLIIAGNL TVIFVFHCAPLLHHYTTSYFIQTMAYADLFVGVSCLVP TLSLLHYSTGVHESLTCQVFGYIISVLKSVSMWCLACIS	6LI0

		VDRYLAITKPLSYNQLVTPCRLRICIILIWIYSCLIFLPSF FGWGKPGYHGDIFEWCATSWLTSAYFTGFIVCLLYAP AAFVVCFTYFHIFKICRQHTKEAKALIVYGSTTGNTEY TAETIARELADAGYEVDSRDAASVEAGGLFEGFDLVL LGCSTWGDDSIELQDDFIPLFDSLEETGAQGRKVACFG CGDSSWEYFCGAVDAIEEKLKNLGAEIVQDGLRIDGD PRAARDDIVGWAHDVRGAIRRYLMVLFRITSVFYMLQ LPYIIYFLLESSRVLDNPTLSFLTTWLAISNSFCNPVIYA LSDSTFRLGLRRLSETMCTS	
LGR4	443	MDPLCAAPCSCDGDRRVDCSGKGLTAVPEGLSAFTQA LDISMNNITQLPEGAFKNFPFLEELQLAGNDLSFIHPKA LSGLKELKVLTLQNNQLKTVPSEAIRGLSALQSLRLDA NHITSVPEDSFEGLVQLRHLWLDDNSLTEVPVHPLSNL PTLQALTLALNKISSIPDFAFTNLSSLVVLHLHNNKIRS LSQHCFDGLDNLETLDLNYNNLGEFPQAIKALPSLKEL GFHSNSISVIPDGAFDGNPLLRTIHLYDNPLSFVGNSAF HNLSDLHSLVIRGASMVQQFPNLTGTVHLESLTLTGTK ISSIPNNLCQEQKMLRTLDLSYNNIRDLPSFNGCHALEE ISLQRNQIYQIKEGTFQGLISLRILDLASNQLKSVPDGIF DRLTSLQKIWLHTNPWDCSCPRIDYLSRWLNKNSQKE QGSAKCSGSGKPVRSIICPTL	4QXE
ZMPSTE24	444	LWEMPAEKRIFGAVLLFSWTVYLWETFLAQRQRRIYK TTTHVPPELGQIMDSETFEKSRLYQLDKSTFSFWSGLY SETEGTLILLFGGIPYLWRLSGRFCGYAGFGPEYEITQS LVFLLLATLFSALTGLPWSLYNTFVIEEKHGFNQQTLG FFMKDAIKKFVVTQCILLPVSSLLLYIIKIGGDYFFIYA WLFTLVVSLVLVTIYADYIAPLFDKFTPLPEGKLKEEIE VMAKSIDFPLTKVYVVEGSKRSSHSNKRIVLFDTLLEE YSVLNKEEIKAKVKNKKQGCKNEEVLAVLGHELGHW KLGHTVKNIIISQMNSFLCFFLFAVLIGRKELFAAFGFY DSQPTLIGLLIIFQFIFSPYNEVLSFCLTVLSRRFEFQADA FAKKLGKAKDLYSALIKLNKDNLGFPVSDWLFSMWH YSHPPLLERLQALKTMKQSGLEVL	5SYT
GLUT1 (SLC2A1)	448	LTGRLMLAVGGAVLGSLQFGYNTGVINAPQKVIEEFY NQTWVHRYGESILPTTLTTLWSLSVAIFSVGGMIGSFS VGLFVNRFGRRNSMLMMNLLAFVSAVLMGFSKLGKS FEMLILGRFIIGVYCGLTTGFVPMYVGEVSPTALRGAL GTLHQLGIVVGILIAQVFGLDSIMGNKDLWPLLLSIIFIP ALLQCIVLPFCPESPRFLLINRNEENRAKSVLKKLRGTA DVTHDLQEMKEESRQMMREKKVTILELFRSPAYRQPI LIAVVLQLSQQLSGINAVFYYSTSIFEKAGVQQPVYATI GSGIVNTAFTVVSLFVVERAGRRTLHLIGLAGMAGCAI LMTIALALLEQLPWMSYLSIVAIFGFVAFFEVGPGPIPW FIVAELFSQGPRPAAIAVAGFSNWTSNFIVGMCFQYVE QLCGPYVFIIFTVLLVLFFIFTYFKVPET	6THA
NEUROKININ 1 RECEPTOR	483	QPAWQIVLWAAAYTVIVVTSVVGNVVVMWIILAHKR MRTVTNYFLVNAAFAEASMAAFNTVVNFTYAVHNE WYYGLFYCKFHNFFPIAAIFASIYSMTAVAFDRYMAII HPLQPRLSLTATKVVICVIWVLALLLAFPQGYYSTTET MPSRVVCKIEWPEHPNKIYEKVYHICVTVLIYFLPLLVI GYLYTVVGITLRASGIDSFWNESYLTGSRDERKKSLLS KFGMDEGVTFMFIGRFDRGQKGVDVLLKAIEILSSKKE FQEMRFIIIGKGDPELEGWARSLEEKHGNVKVITEMLS REFVRELYGSVDFVIIPSYFEPFGLVALEAMCLGAIPIAS AVGGLRDIITNETGILVKAGDPGELANAILKALELSRSD LSKFRENCKKRAMSFSEQVSAARKVVKMMIVVVCTF AICWLPFHIFFLLPYINPDLLKKFIQQVYLAIMWLAMSS TMYNPIIYCCLNDRFRLGFKHAFRCPFIS	6HLP
ADAM22	486	NVEEETKYIELMIVNDHLMFKKHRLSVVHTNTYAKSV VNMADLIYKDQLKTRIVLVAMETWATDNKFAISENPL ITLREFMKYRRDFIKEKSDAVHLFSGSQFESSRSGAAYI GGICSLLKGGGVNEFGKTDLMAVTLAQSLAHNIGIISD	3G5C

		KRKLASGECKCEDTWSGCIMGDTGYYLPKKFTQCNIE EYHDFLNSGGGACLFNKPSKLLDPPECGNGFIETGEEC DCGTPAECVLEGAECCKKCTLTQDSQCSDGLCCKKCK FQPMGTVCREAVNDCDIRETCSGNSSQCAPNIHKMDG YSCDGVQGICFGGRCKTRDRQCKYIWGQKVTASDKY CYEKLNIEGTEKGNCGKDKDTWIQCNKRDVLCGYLL CTNIGNIPRLGELDGEITSTLVVQQGRTLNCSGGHVKL EEDVDLGYVEDGTPCGPQMMCLEHRCLPVASFNFSTC LSSKEGTICSGNGVCSNELKCVCNRHWIGSDCNTYFPH N	
RECEPTOR ECTODOMAIN HETERODIMER	577	SAKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKV TVEHPDKLEEKFPQVAATGDGPDIIFWAHDRFGGYAQ SGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAV EALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALM FNLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDN AGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKGET AMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFV GVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKD KPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIP QMSAFWYAVRTAVINAASGRQTVDEALKDAQTNAA AEFTTACQEANYGALLRELCLTQFQVDMEAVGETLW CDWGRTIRSYRELADCTWHMAEKLGCFWPNAEVDRF FLAVHGRYFRSCPISGRLGVTRNKIMTAQYECYQKIM QDPIQQAEGVYCQRTWDGWLCWNDVAAGTESMQLC PDYFQDFDPSEKVTKICDQDGNWFRHPASQRTWTNYT QCNVNTHEKVKTALNLFYLHHHHHH	6ZHO
L390P MUTANT	594	KHIVVCGHITLESVSNFLKDFLHKDRDDVNVEIVFLHN ISPNLELEAPFKRHFTQVEFYQGSVLNPHDLARVKIESA DACLILANKYCADPDAEDASNIMRVISIKNYHPKIRIIT QMLQYHNKAHLLNIPSWNWKEGDDAICLAELKLGFIA QSCLAQGLSTMLANLFSMRSFIKIEEDTWQKYYLEGV SNEMYTEYLSSAFVGLSFPTVCELCFVKLKLLMIAIEYS RILINPGNHLKIQEGTLGFFIASDAKEVKRAFFYCKDSN VKKYDSTGMFHWCAPKEIEKVILTRSEAAMTVLSGHV VVCIFGDVSSALIGLRNLVMPLRASNFHYHELKHIVFV GSIEYLKREWETLHNFPKVSILPGTPLSRADLRAVNINL CDMCVILSANQDKECILASLNIKSMQFDSITTGVNIPIIT ELVNDTNVQFLDQDDDDDPDTELYLTQPFACGTAFAV SVLDSLMSATYFNDNILTLIRTLVTGGAEALIAEENAL RGGYSTPQTLANRDRCRVAQLALLDGPFADLGDGGC YGDLFCKALKTYNMLCFGIYRLRDAPSQCTKRYVITN PPYEFELVPTDLIFCLMQFDSNSL	6v5A

Table 3.2 Summary of mean v_{model} in the protein sequence sets.

Set	Number	v_{model} [a]	*vs* Test Set		*vs* previous set	
			t-test [b]	*U*-test [b]	*t*-test [b]	*U*-test [b]
PS Test Set	**224**	0.542 ± 0.020	-	-	-	-
ID Null	**23**	0.558 ± 0.019	$3.8E^{-4}$	$2.4E^{-4}$	-	-
Folded	**433**	0.537 ± 0.008	$1.1E^{-3}$	$1.3E^{-4}$	-	-
Previous Folded Set	**82**	0.536 ± 0.008	$2.5E^{-04}$	$7.8E^{-03}$	-	-
Human	**122**	0.536 ± 0.007	$2.0E^{-04}$	$2.2E^{-03}$	0.40	0.32
Small-to-large	**32**	0.537 ± 0.009	$9.2E^{-03}$	$6.2E^{-02}$	0.36	0.41
Extremophile	**54**	0.542± 0.011	$3.4E^{-01}$	$3.3E^{-01}$	$1.2E^{-4}$	$2.4E^{-4}$
Membrane	**90**	0.537 ± 0.006	$1.3E^{-03}$	$2.0E^{-02}$	0.17	0.21
Metamorphic	**53**	0.537± 0.006	$3.4E^{-03}$	$5.0E^{-02}$	0.15	0.18

[a] Mean ± standard deviation.
[b] One-tail *p*-value, where values <0.05 indicate the two compared sets are statistically different in their means.

Table 3.3 Summary of mean β-turn propensity in the protein sequence sets.

Set	Number	β -turn propensity [a]	*vs* Test Set		*vs* previous set	
			t-test [b]	*U*-test [b]	*t*-test [b]	*U*-test [b]
PS Test Set	**224**	1.152 ± 0.087	-	-	-	-
ID Null	**23**	1.062 ± 0.082	$1.4E^{-5}$	$9.7E^{-7}$	-	-
Folded	**433**	0.971 ± 0.040	$1.9E^{-33}$	$1.6E^{-90}$	-	-
Previous Folded Set	**82**	0.969 ± 0.039	$8.0E^{-31}$	$1.7E^{-38}$	-	-
Human	**122**	0.980 ± 0.039	$6.2E^{-31}$	$3.0E^{-48}$	0.03	0.07
Small-to-large	**32**	0.968 ± 0.027	$1.3E^{-29}$	$3.1E^{-19}$	0.42	0.34
Extremophile	**54**	0.983 ± 0.030	$1.3E^{-29}$	$7.5E^{-28}$	0.01	0.03
Membrane	**90**	0.956 ± 0.046	$2.7E^{-31}$	$2.5E^{-41}$	0.02	0.02
Metamorphic	**53**	0.972 ± 0.040	$1.9E^{-28}$	$2.3E^{-27}$	0.30	0.48

[a] Mean ± standard deviation.
[b] One-tail *p*-value, where values <0.05 indicate the two compared sets are statistically different in their means.

Table 3.4 Summary of top 5% of the Amino Acid index scales.

Name	Separation	Amino Acid Index Accession Number	References
Sheet	0.434	QIAN880116	(100)
Turn	0.339	TANS770110	(140)
Hydrophobicity (Structure)	0.320	NADH010106	(101)
Non-sheet	0.320	CHOP780209	(141)
Turn	0.310	CHOP780101	(141)
Helix	0.292	FINA770101	(31)
Turn	0.230	PALJ810106	(142)
Turn	0.226	PALJ810105	(142)
Hydrophobicity (Structure)	0.225	NADH010105	(101)
Aperiodic	0.216	GEIM800108	(143)
Hydrophobicity (Structure)	0.207	PONP800107	(144)
Turn	0.205	CHOP780216	(141)
Turn	0.202	CHOP780203	(141)
Coil	0.187	CHAM830101	(145)
Turn	0.180	PALJ810114	(142)
Coil	0.174	QIAN880131	(100)
Coil	0.161	ROBB760112	(146)
Coil	0.158	QIAN880133	(100)
Hydrophobicity (Structure)	0.145	RACS770101	(147)
Aperiodic	0.132	GEIM800111	(143)
Sheet	0.130	QIAN880115	(100)
Sheet	0.129	QIAN880126	(100)
Coil	0.128	NAGK730103	(148)
Non-sheet	0.124	CHOP780210	(141)
Turn	0.123	PALJ810115	(142)
Flexibility	0.119	KARP850101	(149)
β-turn (Levitt)	0.110	LEVM780103	(15)

Table 3.5 Summary of mean AA properties in the protein sequence sets.

AA property	**Folded set**	**Test set**	**Null set**
α-helix	0.971± 0.025	0.889 ± 0.053	0.943 ± 0.025
β-turn	0.971 ± 0.040	1.152 ± 0.087	1.062 ± 0.082
β-Sheet	-0.068 ± 0.023	0.007 ± 0.041	-0.043 ± 0.024
Φ (structure)	17.83 ± 7.13	13.7 ± 13.11	3.22 ±11.8

Mean ± standard deviation calculated for scale property

Table 3.6 Percent composition of the classical HDX proteins.

NAME	*PDB*	*N*	*Sequence*	*%F*	*%D*	*%P*
Ribonuclease A	1RBX	124	KETAAAKFERQHMDSSTSA ASSSNYCNQMMKSRNLTK DRCKPVNTFVHESLADVQA VCSQKNVACKNGQTNCYQ SYSTMSITDCRETGSSKYPN CAYKTTQANKHIIVACEGN PYVPVHFDASV	73	15	12
Barnase	1A2P	108	VINTFDGVADYLQTYHKLP DNYITKSEAQALGWVASK GNLADVAPGKSIGGDIFSNR EGKLPGKSGRTWREADINY TSGFRNSDRILYSSDWLIYK TTDHYQTFTKIR	68	2	30
Cytochrome c	1HRC	104	GDVEKGKKIFVQKCAQCHT VEKGGKHKTGPNLHGLFG RKTGQAPGFTYTDANKNK GITWKEETLMEYLENPKKY IPGTKMIFAGIKKKTEREDLI AYLKKATNE	47	43	10
Staphylococcal nuclease	1STN	136	KLHKEPATLIKAIDGDTVKL MYKGQPMTFRLLLVDTPET KHPKKGVEKYGPEASAFTK KMVENAKKIEVEFDKGQRT DKYGRGLAYIYADGKMVN EALVRQGLAKVAYVYKPN NTHEQHLRKSEAQAKKEKL NIWS	62	37	1

Table 3.7 Protection factors for the classical HDX proteins.

NAME	# PF residues [a]	# F residues [b]	#ID residues [c]	PF value (F) [d]	PF value (ID) [e]	Mean± σ F residues [f]	Mean± σ ID residues [g]
Ribonuclease A	25	16	9	94.6	46.3	5.9 ± 1.0	5.1 ±1.2
Barnase	40	26	14	158.9	89.6	6.1 ± 1.0	6.4 ± 1.0
Cytochrome c	42	28	14	176.8	76.0	6.3 ±1.4	5.1 ± 1.1
Staphylococcal nuclease	92	64	28	142.9	52.7	4.1 ± 0.4	3.7 ± 0.4

[a] Total number of residues with resolved protection factor (PF) data
[b] Total number of positions in sequence classified as F by ParSe with PF value
[c] Total number of positions in sequence classified as ID (D or P) by ParSe with PF value
[d] Sum of logarithmic PF for protected residues
[e] Sum of logarithmic PF for unprotected residues
[f] Mean ± σ of protected residues
[g] Mean ± σ of unprotected residues

Table 3.8 Percent composition of folded, ID, and PS-ID HDX proteins.

NAME	N	Sequence	%F	%D	%P
Apo-Myoglobin	152	VLSEGEWQLVLHVWAKVEADVAGHG QDILIRLFKSHPELEKFDRFKHLKTEAE MKASEDLKKHGVTVLTALGAILKKKG HHEAELKPLAQSHATKHKIPIKYLEFIS EAIIHVLHSRHPGDFGADAQGAMNKA LELFRKDIAAKYKELGY	95	5	0
T4 Lysozyme	164	MNIFEMLRIDEGLRLKIYKDTEGYYTI GIGHLLTKSPSLNAAKSELDKAIGRNC NGVITKDEAEKLFNQDVDAAVRGILR NAKLKPVYDSLDAVRRCALINMVFQM GETGVAGFTNSLRMLQQKRWDEAAV NLAKSRWYNQTPNRAKRVITTFRTGT WDAYKNL	76	3	1
Chymotrypsin Inhibitor 2	65	NLKTEWPELVGKSVEEAKKVILQDKP EAQIIVLPVGTIVTMEYRIDRVRLFVDK LDNIAEVPRVG	97	3	0
Bovine Pancreatic Trypsin Inhibitor	58	RPDFCLEPPYTGPCKARIIRYFYNAKA GLCQTFVYGGCRAKRNNFKSAEDCMR TCGGA	43	57	0
Hen Egg-White Lysozyme	130	KVFGRCELAAAMKRHGLDNYRGYSL GNWVCAAKFESNFNTQATNRNTDGST DYGILQINSRWWCNDGRTPGSRNLCNI PCSALLSSDITASVNCAKKIVSDGNGM NAWVAWRNRCKGTDVQAWIRGCRL	73	2	25
Ribonuclease H	155	MLKQVEIFTDGSCLGNPGPGGYGAILR YRGREKTFSAGYTRTTNNRMELMAAI VALEALKEHCEVILSTDSQYVRQGITQ WIHNWKKRGWKTADKKPVKNVDLW QRLDAALGQHQIKWEWVKGHAGHPE NERCDELARAAAMNPTLEDTGYQVEV	49	33	18
Che-y	128	ADKELKFLVVDDFSTMRRIVRNLLKEL GFNNVEEAEDGVDALNKLQAGGYGF VISDWNMPNMDGLELLKTIRADGAMS ALPVLMVTAEAKKENIIAAAQAGASG YVVKPFTAATLEEKLNKIFEKLGM	89	8	2
α-Lactalbumin (Guinea Pig)	123	KQLTKCALSHELNDLAGYRDITLPEWL CIIFHISGYDTQAIVKNSDHKEYGLFQI NDKDFCESSTTVQSRNICDISCDKLLD DDLTDDIMCVKKILDIKGIDYWLAHKP LCSDKLEQWYCEAQ	89	11	0
Ubiquitin	76	MQIFVKTLTGKTITLEVEPSDTIENVKA KIQDKEGIPPDQQRLIFAGKQLEDGRTL SDYNIQKESTLHLVLRLRGG	68	29	3
Tendamistat	74	DTTVSEPAPSCVTLYQSWRYSQADNG CAETVTVKVVYEDDTEGLCYAVAPGQ ITTVGDGYIGSHGHARYLARCL	47	45	8
Carbon-Monoxy Myoglobin	153	VLSEGEWQLVLHVWAKVEADVAGHG QDILIRLFKSHPETLEKFDRFKHLKTEA EMKASEDLKKHGVTVLTALGAILKKK GHHEAELKPLAQSHATKHKIPIKYLEFI SEAIIHVLHSRHPGDFGADAQGAMNK ALELFRKDIAAKYKELGYQG	95	5	0

Bovine β-Lactoglobulin	162	LIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI	75	25	0
Equine Lysozyme	129	KVFSKCELAHKLKAQEMDGFGGYSLANWVCMAEYESNFNTRAFNGKNANGSSDYGLFQLNNKWWCKDNKRSSSNACNIMCSKLLDENIDDDISCAKRVVRDPKGMSAWKAWVKHCKDKDLSEYLASCNL	76	9	15
α-Lactalbumin	120	MQFTKCELSQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDITDDIMCAKKILDIKGIDYWLAHKALCTEKLEQWLC	79	21	0

Table 3.9 Protection factor values for HDX proteins.

NAME	# Protection factor positions [a]	# F residues [b]	# ID residues [c]	Protection Factor (F) [d]	Protection Factor (ID)[e]	Mean± σ F residues [f]	Mean± σ ID residues [g]
Apo-Myoglobin	39	38	1	126.1	3.0	3.4 ± 1.2	-
T4 Lysozyme	40	29	11	161.5	58.2	5.6 ± 0.8	5.3 ± 0.6
Chymotrypsin Inhibitor 2	37	35	2	19.2	3.4	0.5 ± 1.2	1.7 ± 1.8
Bovine Pancreatic Trypsin Inhibitor	10	4.0	6	7.9	11.9	1.9 ± 0.7	1.9 ± 2.2
Hen Egg-White Lysozyme	60	45	15	74.0	20.0	1.6 ± 1.0	1.4 ± 0.8
Ribonuclease H	31	20	11	23.6	0.0	1.7 ± 0.5	-
Che-y	37	32	4	129.0	16.0	4.0 ± 0.4	4.0 ± 0.4
α-Lactalbumin (Guinea Pig)	41	26	15	27.3	4.0	1.6 ± 0.7	1.1 ± 0.2
Ubiquitin	41	26	15	83.6	50.7	3.2 ± 1.1	3.4 ± 1.2
Tendamistat	50	23	27	63.0	67.5	2.7 ± 0.8	2.5 ± 1.1
Carbon-Monoxy Myoglobin	38	37	1	126.6	3.0	3.4 ± 1.2	-
Bovine β-Lactoglobulin	79	55	24	183.1	79.2	3.3 ± 0.9	3.4 ± 1.0
Equine Lysozyme	67	52	15	215.5	50	4.1 ± 1.0	3.4 ± 0.9
α-Lactalbumin	45	33	12	149.2	5.2	4.4 ± 0.7	4.6 ± 0.5

[a] Total number of residues with resolved protection factor (PF) data
[b] Total number of positions in sequence classified as F by ParSe with PF value
[c] Total number of positions in sequence classified as ID (D or P) by ParSe with PF value
[d] Sum of logarithmic PF for protected residues
[e] Sum of logarithmic PF for unprotected residues
[f] Mean ± σ of protected residues
[g] Mean ± σ of unprotected residues

Figures.

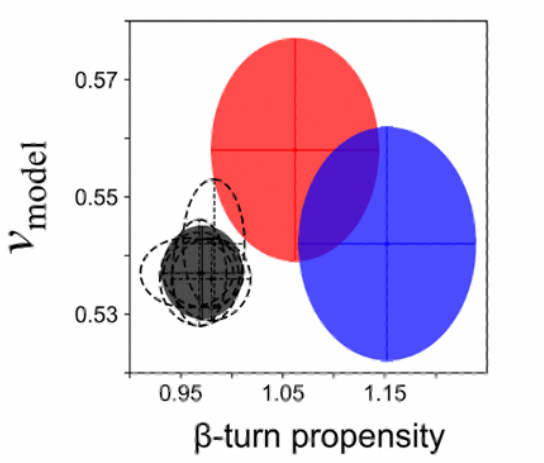

Figure 3.1 Mean values of β-turn propensity and ν_{model} by protein class. Sequence calculated mean ± σ values of added folded sequence sets plotted with previously determined mean ± σ values of the previous folded, null, and testing set. Data points and error bars represent mean and standard deviation values for each set. The mean ± σ of the folded set is represented in black, the ID-null sequence set is represented in red, and the PS-ID test set is represented in blue.

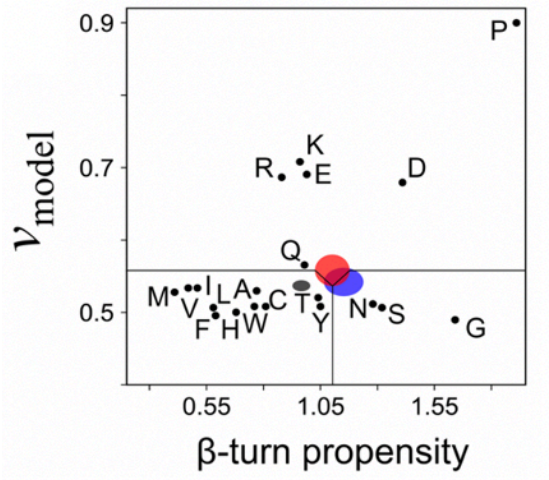

Figure 3.2 Mean values of β-turn propensity and ν_{model} for the homopolymers of the 20 common amino acids. Sector boundaries are defined by mean and standard deviation values of the null set (X =1.062 ± 0.082, Y= 0.558 ± 0.019). Amino acid residues classify into order promoting, disorder promoting, or phase separation promoting sectors based on calculated mean values of β-turn propensity and ν_{model} for each residue.

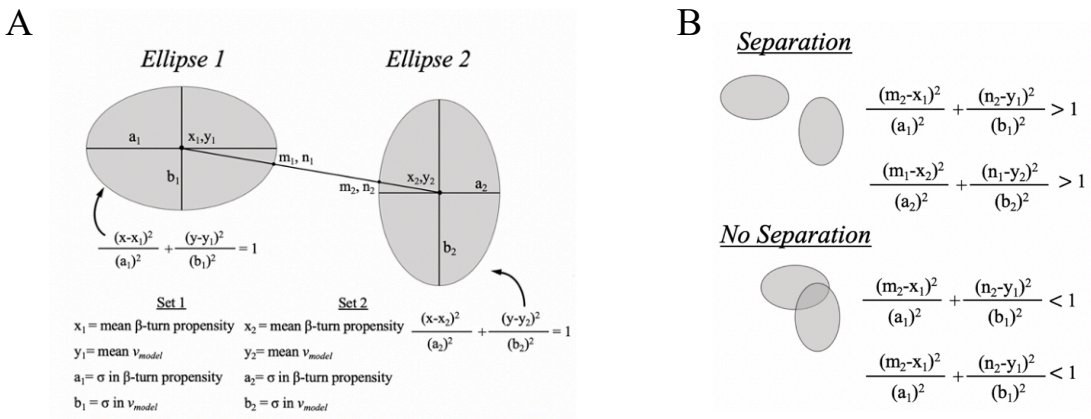

Figure 3.3 Calculating separation from mean ± σ of protein sets. *A*, Calculated mean ± σ values of protein sets (i.e., Folded, ID, PS-ID) are plotted on β-turn propensity versus ν_{model} plot so that mean and standard deviation values define boundaries to create an ellipse. *B,* Separation of ellipses is determined by the distance between each origin of each ellipse (i.e., x1, y1, x2, y2) minus the vector length between each ellipse (i.e., m1, n1).

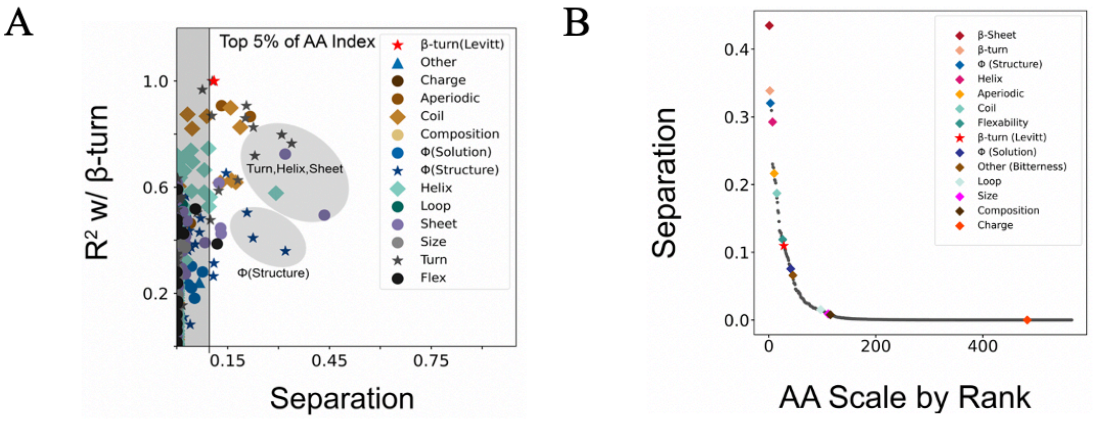

Figure 3.4 Separation in the three sequence sets calculated for ν_{model} paired with each amino acid scale. *A*, Correlation, R^2, of each amino acid scale to the Levitt β-turn propensity scale is plotted against separation, where the separation was calculated for the three sequence sets, PS-IDR test, IDR null, and folded. In general, amino acid scales based upon conformational propensities showed the greatest separation between folded, ID, and PS-ID protein regions. *B,* Separation plotted *versus* rank order in separation, which demonstrates that only a small subset of the scales produce relatively large separation in the three sequence sets.

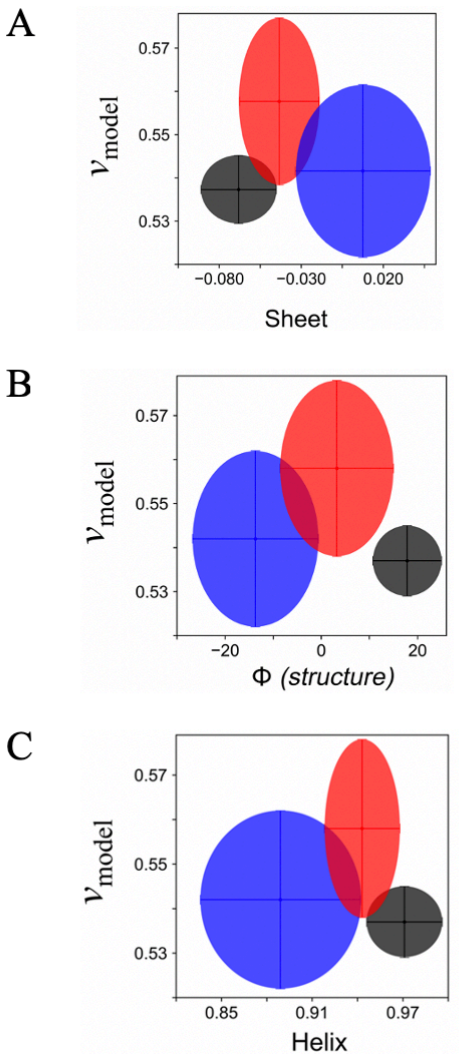

Figure 3.5 Mean values of β-sheet, α-helix, and hydrophobicity and ν_{model} by protein class. The amino acid scale properties were applied to the folded, ID-null, PS-ID test set to determine the mean ± σ of each sequence set using each amino acid property. Ellipses are defined by mean ± σ values of each sequence set. Black ellipses represent the folded set, red ellipse represent ID-null set, and blue ellipse represents PS-ID test set.

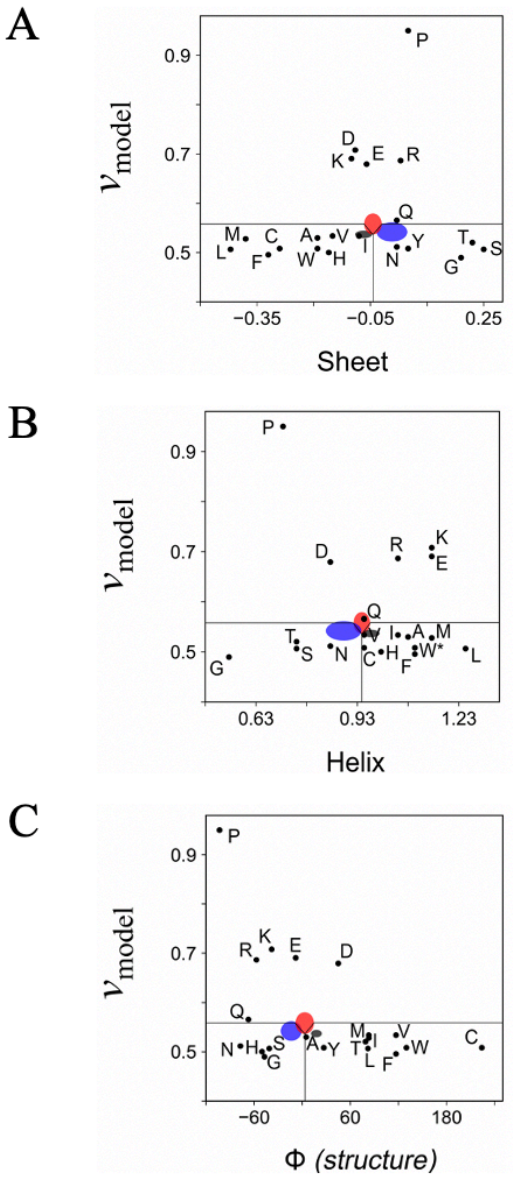

Figure 3.6 Mean values of β-sheet, α-helix, and Φ properties and ν_{model} for the homopolymers of the 20 common amino acids. The 20 common amino acids were evaluated by applying each amino acid scale property to determine mean β-sheet, α-helix, and Φ and mean ν_{model} values for each amino acid residue. Folded, ID, and PS-ID sectors are defined by the mean ± σ values of the ID-null set. Black, red, and blue ellipses are plotted to indicate the mean values of folded, ID, or PS-ID for each sequence set.

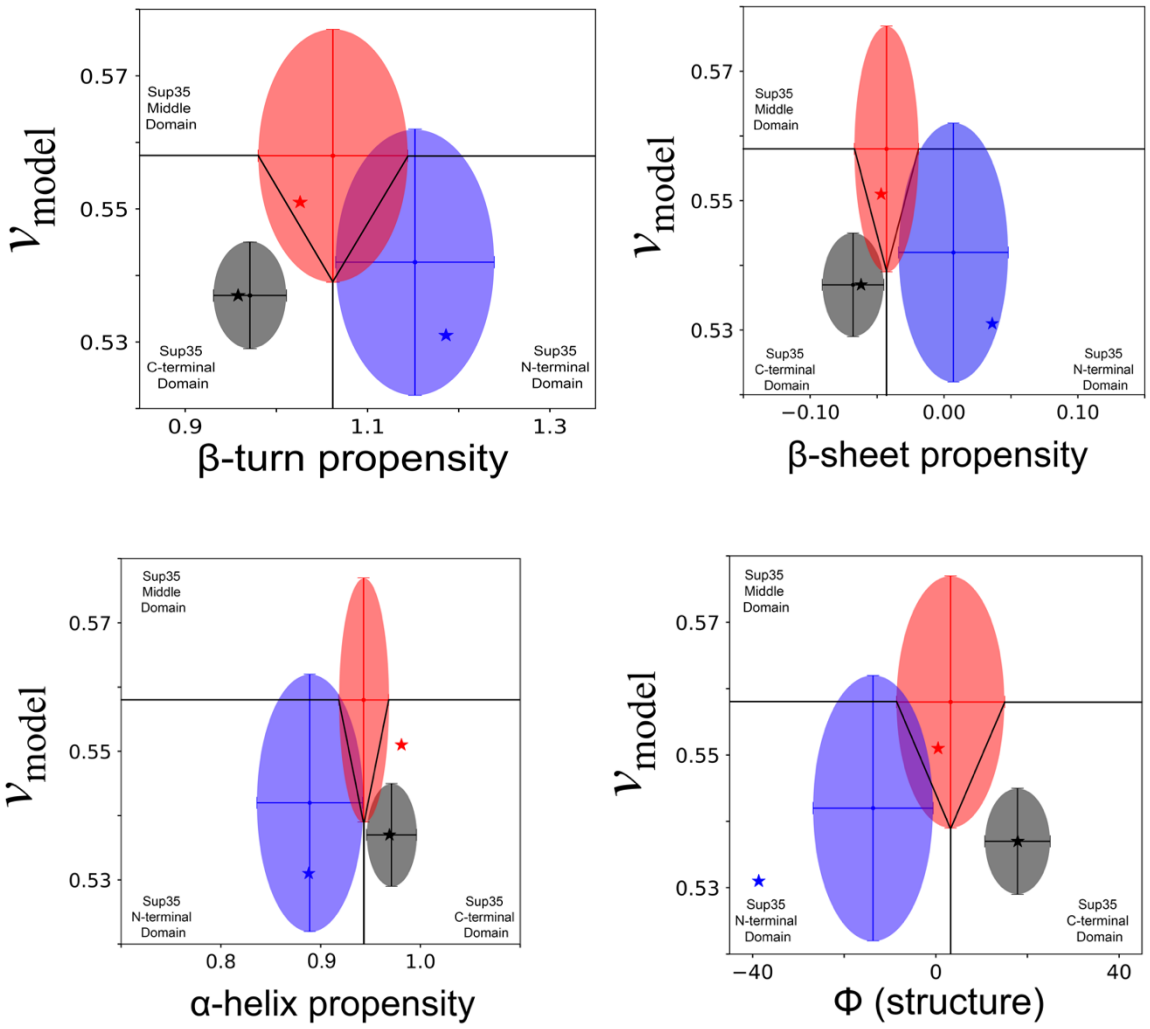

Figure 3.7 Sequence calculated means of β-turn, β-sheet propensity, α-helix propensity, Φ, and ν_{model} identify folded, ID, and PS-IDR in Sup35. *A-D*, Amino acid sequence properties are paired with ν_{model} to calculate mean values of known folded (C-terminal domain), ID (middle domain), and PS-ID (N-terminal) regions of Sup35. Mean values for each domain are indicated by stars in each plot. Ellipses represent mean ± σ values of the folded (black), ID-null (red), and PS-ID test (blue) sequence sets.

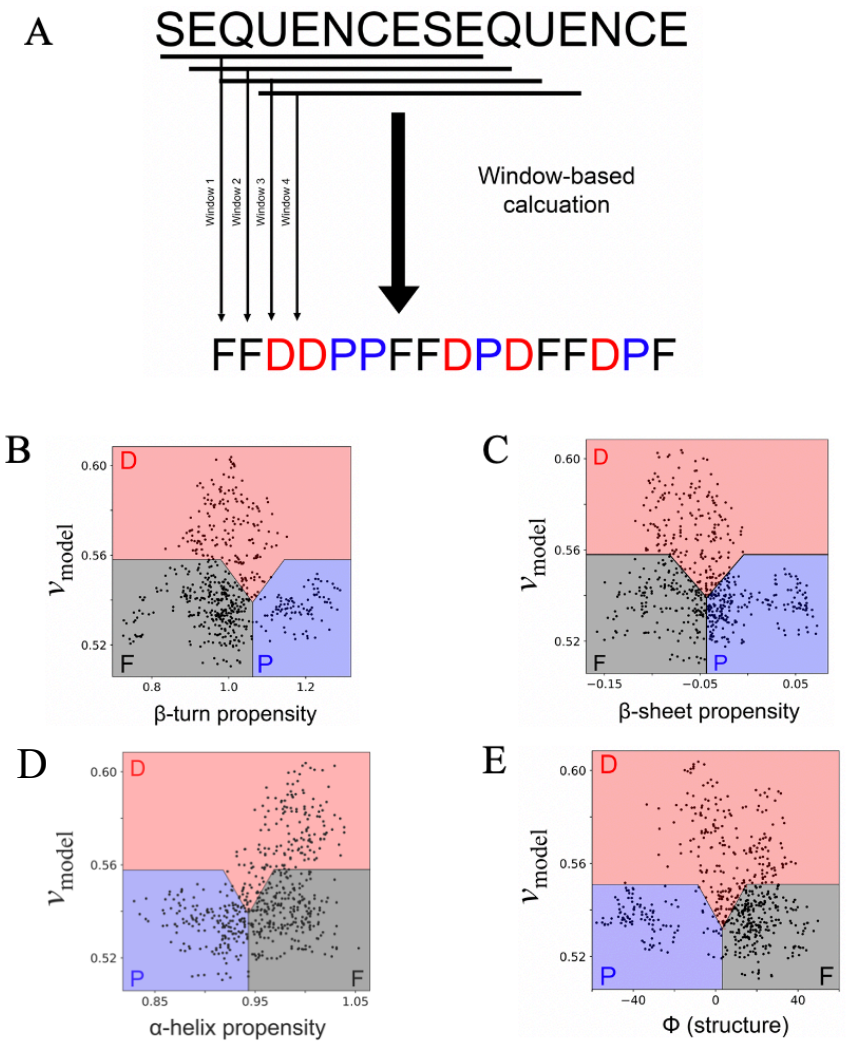

Figure 3.8 Window based calculations using top performing amino acid scales paired with ν_{model} classify folded, IDR, or PS-IDR residues in Sup35. A, a sliding window algorithm is used to identify from sequence regions within a protein that match the PS ID, ID, and folded classes. *B,* β-turn propensity, β-sheet propensity, α-helix propensity , Hydrophobicity (ϕ), and ν_{model} are calculated for each contiguous stretch of 25-residues, or "window", in the primary sequence of Sup35. Each window is assigned a label of F, D, or P based on if mean ν_{model} and mean amino acid scale placed it in the PS-ID, ID, or folded sector, respectively, of the scale versus ν_{model} plot.

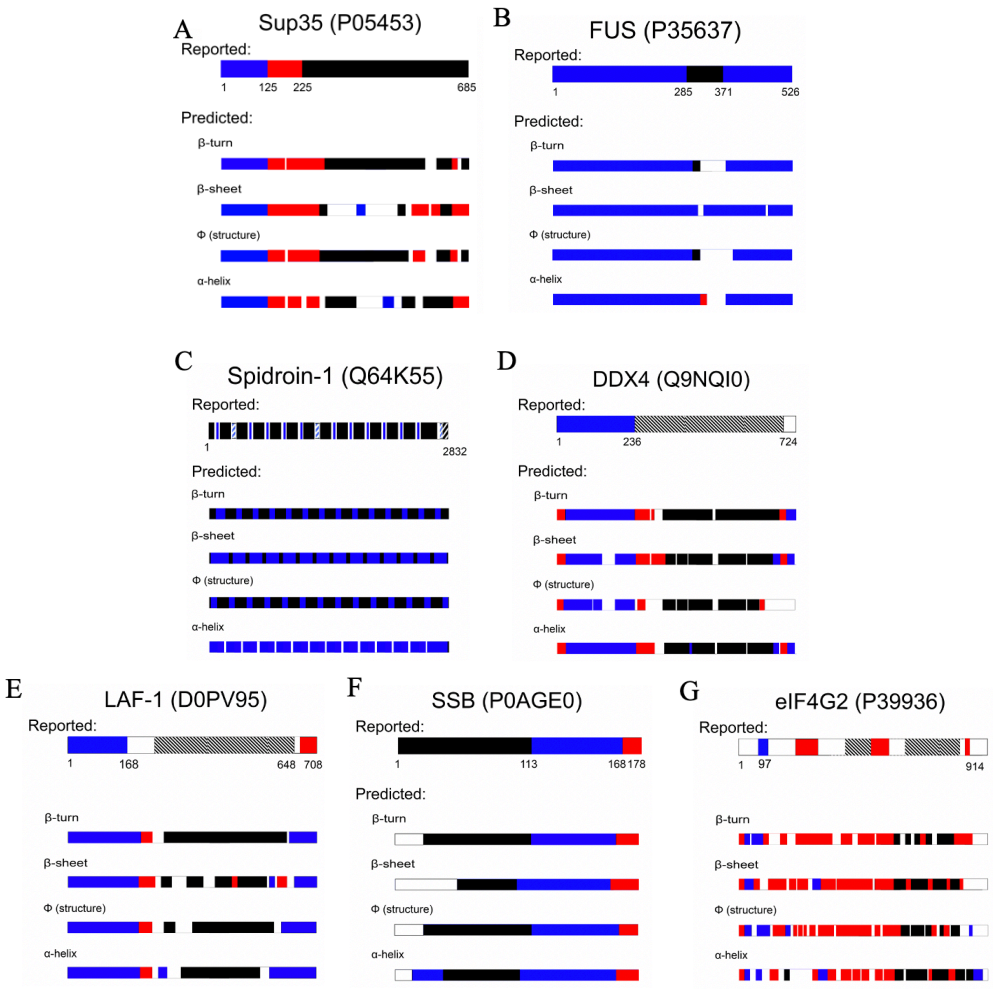

Figure 3.9 Sliding window calculations applied to verified *in vitro* sufficient LLPS proteins. *A-E*, For each 25-residue window, our algorithm assigns an F, D, or P based on the means determined by v_{model} paired with each sequence property (β -turn propensity, β-sheet propensity, α-helix propensity, Φ property). Folded (black), intrinsically disordered (red), and PS-ID (blue) regions. Calculations were compared to previously reported (15) structural data for each protein, identified by name and UniProt accession number. Striped represents ≥50% identity to a known LLPS IDR (blue) or folded protein (black).

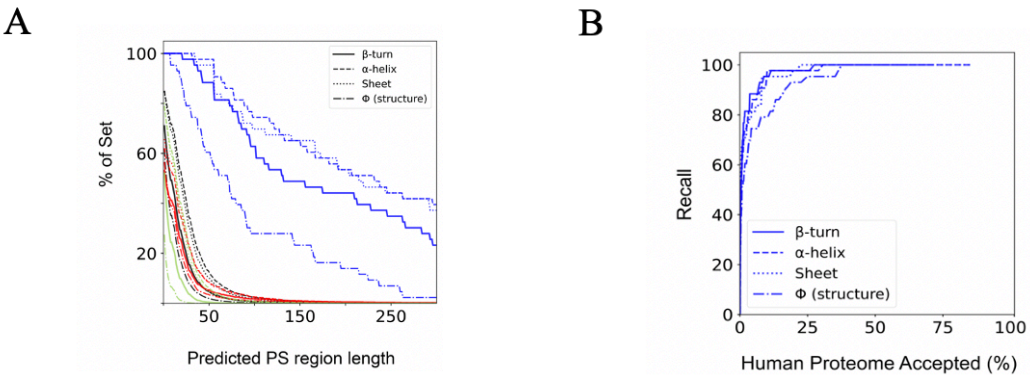

Figure 3.10 Long regions matching the LLPS IDR class are rare in the human proteome, the DisProt database, and folded proteins. *A*, ParSe (Solid line), ParSe α-helix propensity(dot-dash), ParSe β-sheet propensity (dotted line), and ParSe Φ (stippled line) were used to identify regions in proteins that were ≥ 90% labeled P, which are referred to as phase-separating, PS, regions. Shown by the y-axis is the percent of proteins in a set with PS regions at least as long as the length indicated by the x-axis. The human proteome (UniProt reference proteome UP000005640) is given by black lines; DisProt (minus LLPS annotated entries) by red lines; SCOPe (version 2.07) by green lines; a set of in vitro sufficient homotypic LLPS proteins by blue lines. *B*, Recall plot was produced by comparing our calculations for the human proteome to the *in vitro* sufficient homotypic LLPS proteins for β-turn propensity, β-sheet propensity, α-helix propensity, Φ property, and ν_{model}.

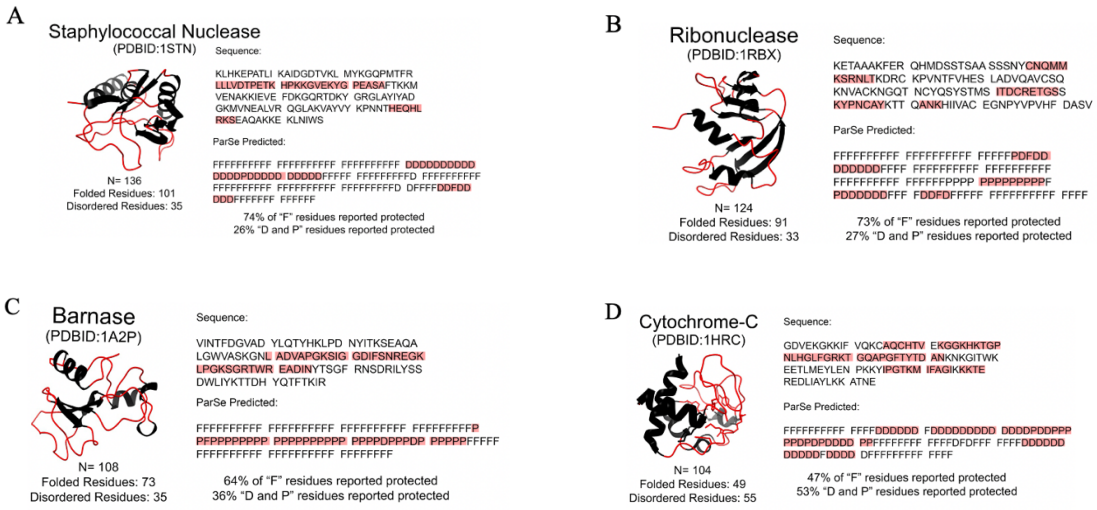

Figure 3.11 ParSe identifies regions of disorder in classically tested hydrogen deuterium proteins to be less protected. Window-based calculations using β-turn propensity and v_{model} were applied to *A* staphylococcal nuclease (PDBID:1STN), *B* ribonuclease (PDBID:1RBX), *C* barnase (PDBID:1A2P), and *D* cytochrome-C (PDBID:1HRC) to identify structured and disordered regions from sequence. Regions labeled "D" or "P" correspond to disordered (highlighted red), while regions labeled "F" correspond to structured regions. Regions labeled "D" or "P" correspond to regions resulting in lower protection factor values when compared to regions that were classified as "F" by ParSe.

REFERENCES

1. Murray, J. E., Laurieri, N., and Delgoda, R. (2017) Proteins. in *Pharmacognosy*, pp. 477–494, Elsevier, 10.1016/B978-0-12-802104-0.00024-X
2. Quick, M., and Javitch, J. A. (2007) Monitoring the function of membrane transport proteins in detergent-solubilized form. *Proc. Natl. Acad. Sci.* **104**, 3603–3608
3. Buccitelli, C., and Selbach, M. (2020) mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644
4. Englander, S. W., Mayne, L., Kan, Z.-Y., and Hu, W. (2016) Protein Folding—How and Why: By Hydrogen Exchange, Fragment Separation, and Mass Spectrometry. *Annu. Rev. Biophys.* **45**, 135–152
5. Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., and Zhou, Y. (2016) Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Brief. Bioinform.* 10.1093/bib/bbw129
6. Forman-Kay, J. D., and Mittag, T. (2013) From Sequence and Forces to Structure, Function, and Evolution of Intrinsically Disordered Proteins. *Structure*. **21**, 1492–1499

7. Cournia, Z., Allen, T. W., Andricioaei, I., Antonny, B., Baum, D., Brannigan, G., Buchete, N.-V., Deckman, J. T., Delemotte, L., del Val, C., Friedman, R., Gkeka, P., Hege, H.-C., Hénin, J., Kasimova, M. A., Kolocouris, A., Klein, M. L., Khalid, S., Lemieux, M. J., Lindow, N., Roy, M., Selent, J., Tarek, M., Tofoleanu, F., Vanni, S., Urban, S., Wales, D. J., Smith, J. C., and Bondar, A.-N. (2015) Membrane Protein Structure, Function, and Dynamics: a Perspective from Experiments and Theory. *J. Membr. Biol.* **248**, 611–640
8. Guzzo, A. V. (1965) The Influence of Amino Acid Sequence on Protein Structure. *Biophys. J.* **5**, 809–822
9. Wu, G. (2010) Functional Amino Acids in Growth, Reproduction, and Health. *Adv. Nutr.* **1**, 31–37
10. Campen, A., Williams, R., Brown, C., Meng, J., Uversky, V., and Dunker, A. (2008) TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. *Protein Pept. Lett.* **15**, 956–963
11. Lazar, G. A., and Handel, T. M. (1998) Hydrophobic core packing and protein design. *Curr. Opin. Chem. Biol.* **2**, 675–679
12. Munson, M., Balasubramanian, S., Fleming, K. G., Nagi, A. D., O'Brien, R., Sturtevant, J. M., and Regan, L. (1996) What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. *Protein Sci.* **5**, 1584–1593
13. Englander, S. W., and Mayne, L. (2014) The nature of protein folding pathways. *Proc. Natl. Acad. Sci.* **111**, 15873–15880

14. Wang, S., Gu, J., Larson, S. A., Whitten, S. T., and Hilser, V. J. (2008) Denatured-State Energy Landscapes of a Protein Structural Database Reveal the Energetic Determinants of a Framework Model for Folding. *J. Mol. Biol.* **381**, 1184–1201

15. Levitt, M. (1978) Conformational preferences of amino acids in globular proteins. *Biochemistry*. **17**, 4277–4285

16. Eisenberg, D. (2003) The discovery of the -helix and -sheet, the principal structural features of proteins. *Proc. Natl. Acad. Sci.* **100**, 11207–11210

17. Marcelino, A. M. C., and Gierasch, L. M. (2008) Roles of β-turns in protein folding: From peptide models to protein engineering. *Biopolymers*. **89**, 380–391

18. Reeb, J., and Rost, B. (2019) Secondary Structure Prediction. in *Encyclopedia of Bioinformatics and Computational Biology*, pp. 488–496, Elsevier, 10.1016/B978-0-12-809633-8.20267-7

19. Rudra, J. S., and Collier, J. H. (2011) Self-Assembling Biomaterials. in *Comprehensive Biomaterials*, pp. 77–94, Elsevier, 10.1016/B978-0-08-055294-1.00063-5

20. Robinson, S. W., Afzal, A. M., and Leader, D. P. (2014) Bioinformatics: Concepts, Methods, and Data. in *Handbook of Pharmacogenomics and Stratified Medicine*, pp. 259–287, Elsevier, 10.1016/B978-0-12-386882-4.00013-X

21. Müller, M. M. (2018) Post-Translational Modifications of Protein Backbones: Unique Functions, Mechanisms, and Challenges. *Biochemistry*. **57**, 177–185

22. Modi, V., and Dunbrack, R. L. (2019) Defining a new nomenclature for the structures of active and inactive kinases. *Proc. Natl. Acad. Sci.* **116**, 6818–6827

23. Podtelezhnikov, A. A., and Wild, D. L. (2009) Reconstruction and Stability of Secondary Structure Elements in the Context of Protein Structure Prediction. *Biophys. J.* **96**, 4399–4408

24. Mallamace, D., Fazio, E., Mallamace, F., and Corsaro, C. (2018) The Role of Hydrogen Bonding in the Folding/Unfolding Process of Hydrated Lysozyme: A Review of Recent NMR and FTIR Results. *Int. J. Mol. Sci.* **19**, 3825

25. Doig, A. J., Andrew, C. D., Hughes, E., Penel, S., Sun, J. K., Stapley, B. J., Clarke, D., and Jones, G. R. Structure, stability and folding of the ⌴-helix

26. Jacob, J., Duclohier, H., and Cafiso, D. S. (1999) The Role of Proline and Glycine in Determining the Backbone Flexibility of a Channel-Forming Peptide. *Biophys. J.* **76**, 1367–1376

27. Imai, K., and Mitaku, S. (2005) Mechanisms of secondary structure breakers in soluble proteins. *BIOPHYSICS.* **1**, 55–65

28. Javadpour, M. M., Eilers, M., Groesbeek, M., and Smith, S. O. (1999) Helix Packing in Polytopic Membrane Proteins: Role of Glycine in Transmembrane Helix Association. *Biophys. J.* **77**, 1609–1618

29. Holehouse, S., and Pappu, V. Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins

30. Deller, M. C., Kong, L., and Rupp, B. (2016) Protein stability: a crystallographer's perspective. *Acta Crystallogr. Sect. F Struct. Biol. Commun.* **72**, 72–95

31. Finkelstein, A. V., and Badretdinov, A. Y. (1997) Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold. Des.* **2**, 115–121

32. Luke, K. A., Higgins, C. L., and Wittung-Stafshede, P. (2007) Thermodynamic stability and folding of proteins from hyperthermophilic organisms: Thermodynamic stability and folding of proteins. *FEBS J.* **274**, 4023–4033

33. Zwanzig, R. (1997) Two-state models of protein folding kinetics. *Proc. Natl. Acad. Sci.* **94**, 148–150

34. Nick Pace, C., Scholtz, J. M., and Grimsley, G. R. (2014) Forces stabilizing proteins. *FEBS Lett.* **588**, 2177–2184

35. Guo, M., Xu, Y., and Gruebele, M. (2012) Temperature dependence of protein folding kinetics in living cells. *Proc. Natl. Acad. Sci.* **109**, 17863–17867

36. Tompa, P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.* **37**, 509–516

37. Berlow, R. B., Dyson, H. J., and Wright, P. E. (2018) Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation. *J. Mol. Biol.* **430**, 2309–2320

38. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331

39. Wright, P. E., and Dyson, H. J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29

40. Harmon, T. S., Holehouse, A. S., Rosen, M. K., and Pappu, R. V. (2017) Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *eLife*. **6**, e30294

41. Uversky, V. N. (2019) Intrinsically Disordered Proteins and Their “Mysterious” (Meta)Physics. *Front. Phys.* **7**, 10

42. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., and Babu, M. M. (2014) Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114**, 6589–6631

43. Rabouille, C. (2019) Membraneless organelles in cell biology. *Traffic*. **20**, 885–886

44. Babinchak, W. M., and Surewicz, W. K. (2020) Liquid–Liquid Phase Separation and Its Mechanistic Role in Pathological Protein Aggregation. *J. Mol. Biol.* **432**, 1910–1925

45. Banani, S. F., Lee, H. O., Hyman, A. A., and Rosen, M. K. (2017) Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298

46. Sawyer, I. A., Sturgill, D., Sung, M.-H., Hager, G. L., and Dundr, M. (2016) Cajal body function in genome organization and transcriptome diversity. *BioEssays*. **38**, 1197–1208

47. Brangwynne, C. P., Eckmann, C. R., Courson, D. S., Rybarska, A., Hoege, C., Gharakhani, J., Julicher, F., and Hyman, A. A. (2009) Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science*. **324**, 1729–1732

48. Yamashita, Y. M. (2018) Subcellular specialization and organelle behavior in germ cells. *Genetics*. **208**, 19–51

49. Martinelli, A., Lopes, F., John, E., Carlini, C., and Ligabue-Braun, R. (2019) Modulation of Disordered Proteins with a Focus on Neurodegenerative Diseases and Other Pathologies. *Int. J. Mol. Sci.* **20**, 1322

50. Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A. P., Kim, H. J., Mittag, T., and Taylor, J. P. (2015) Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. *Cell*. **163**, 123–133

51. Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., and Mittag, T. (2020) Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*. **367**, 694–699

52. Choi, J.-M., and Pappu, R. V. (2020) The Stickers and Spacers Framework for Describing Phase Behavior of Multivalent Intrinsically Disordered Proteins. *Biophys. J.* **118**, 492a

53. Brangwynne, C. P., Eckmann, C. R., Courson, D. S., Rybarska, A., Hoege, C., Gharakhani, J., Julicher, F., and Hyman, A. A. (2009) Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science*. **324**, 1729–1732

54. Banani, S. F., Rice, A. M., Peeples, W. B., Lin, Y., Jain, S., Parker, R., and Rosen, M. K. (2016) Compositional Control of Phase-Separated Cellular Bodies. *Cell*. **166**, 651–663

55. Yang, Y., Jones, H. B., Dao, T. P., and Castañeda, C. A. (2019) Single Amino Acid Substitutions in Stickers, but Not Spacers, Substantially Alter UBQLN2 Phase Transitions and Dense Phase Material Properties. *J. Phys. Chem. B*. **123**, 3618–3629

56. Chiu, Y.-P., Sun, Y.-C., Qiu, D.-C., Lin, Y.-H., Chen, Y.-Q., Kuo, J.-C., and Huang, J. (2020) Liquid-liquid phase separation and extracellular multivalent interactions in the tale of galectin-3. *Nat. Commun.* **11**, 1229

57. Baul, U., Chakraborty, D., Mugnai, M. L., Straub, J. E., and Thirumalai, D. (2019) Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins. *J. Phys. Chem. B*. **123**, 3462–3474

58. English, L. R., Tilton, E. C., Ricard, B. J., and Whitten, S. T. (2017) Intrinsic α helix propensities compact hydrodynamic radii in intrinsically disordered proteins: α Helix Propensities Compact IDP Structures. *Proteins Struct. Funct. Bioinforma.* **85**, 296–311

59. Tomasso, M. E., Tarver, M. J., Devarajan, D., and Whitten, S. T. (2016) Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. *PLOS Comput. Biol.* **12**, e1004686

60. Gabryelczyk, B., Cai, H., Shi, X., Sun, Y., Swinkels, P. J. M., Salentinig, S., Pervushin, K., and Miserez, A. (2019) Hydrogen bond guidance and aromatic stacking drive liquid-liquid phase separation of intrinsically disordered histidine-rich peptides. *Nat. Commun.* **10**, 5465

61. Chiu, Y.-P., Sun, Y.-C., Qiu, D.-C., Lin, Y.-H., Chen, Y.-Q., Kuo, J.-C., and Huang, J. (2020) Liquid-liquid phase separation and extracellular multivalent interactions in the tale of galectin-3. *Nat. Commun.* **11**, 1229

62. Paiz, E. A., Allen, J. H., Correia, J. J., Fitzkee, N. C., Hough, L. E., and Whitten, S. T. (2021) Beta turn propensity and a model polymer scaling exponent identify intrinsically disordered phase-separating proteins. *J. Biol. Chem.* **297**, 101343

63. Panja, A. S., Maiti, S., and Bandyopadhyay, B. (2020) Protein stability governed by its structural plasticity is inferred by physicochemical factors and salt bridges. *Sci. Rep.* **10**, 1822

64. Casares, D., Escribá, P. V., and Rosselló, C. A. (2019) Membrane Lipid Composition: Effect on Membrane and Organelle Structure, Function and Compartmentalization and Therapeutic Avenues. *Int. J. Mol. Sci.* **20**, 2167

65. Fitzkee, N. C., and Rose, G. D. (2004) Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci.* **101**, 12497–12502

66. Porter, L. L., and Looger, L. L. (2018) Extant fold-switching proteins are widespread. *Proc. Natl. Acad. Sci.* **115**, 5968–5973

67. Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., Lin, H., and Forman-Kay, J. D. (2018) Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *eLife*. **7**, e31486

68. Mészáros, B., Erdős, G., Szabó, B., Schád, É., Tantos, Á., Abukhairan, R., Horváth, T., Murvai, N., Kovács, O. P., Kovács, M., Tosatto, S. C. E., Tompa, P., Dosztányi, Z., and Pancsa, R. (2019) PhaSePro: the database of proteins driving liquid–liquid phase separation. *Nucleic Acids Res.* 10.1093/nar/gkz848

69. Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G. I., Bevilacqua, M., Chasapi, A., Chemes, L., Davey, N. E., Davidović, R., Dunker, A. K., Elofsson, A., Gobeill, J., Foutel, N. S. G., Sudha, G., Guharoy, M., Horvath, T., Iglesias, V., Kajava, A. V., Kovacs, O. P., Lamb, J., Lambrughi, M., Lazar, T., Leclercq, J. Y., Leonardi, E., Macedo-Ribeiro, S., Macossay-Castillo, M., Maiani, E., Manso, J. A., Marino-Buslje, C., Martínez-Pérez, E., Mészáros, B., Mičetić, I., Minervini, G., Murvai, N., Necci, M., Ouzounis, C. A., Pajkos, M., Paladin, L., Pancsa, R., Papaleo, E., Parisi, G., Pasche, E., Barbosa Pereira, P. J., Promponas, V. J., Pujols, J., Quaglia, F., Ruch, P., Salvatore, M., Schad, E., Szabo, B., Szaniszló, T., Tamana, S., Tantos, A., Veljkovic, N., Ventura, S., Vranken, W., Dosztányi, Z., Tompa, P., Tosatto, S. C. E., and Piovesan, D. (2019) DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* 10.1093/nar/gkz975

70. Perez, R. B., Tischer, A., Auton, M., and Whitten, S. T. (2014) Alanine and proline content modulate global sensitivity to discrete perturbations in disordered proteins: Alanine and Proline Effects on IDP Structure. *Proteins Struct. Funct. Bioinforma.* **82**, 3373–3384

71. Haaning, S., Radutoiu, S., Hoffmann, S. V., Dittmer, J., Giehm, L., Otzen, D. E., and Stougaard, J. (2008) An Unusual Intrinsically Disordered Protein from the Model Legume Lotus japonicus Stabilizes Proteins in Vitro. *J. Biol. Chem.* **283**, 31142–31152

72. Choi, U. B., McCann, J. J., Weninger, K. R., and Bowen, M. E. (2011) Beyond the Random Coil: Stochastic Conformational Switching in Intrinsically Disordered Proteins. *Structure*. **19**, 566–576

73. Sánchez-Puig, N., Veprintsev, D. B., and Fersht, A. R. (2005) Binding of Natively Unfolded HIF-1α ODD Domain to p53. *Mol. Cell*. **17**, 11–21

74. Uversky, V. N., Permyakov, S. E., Zagranichny, V. E., Rodionov, I. L., Fink, A. L., Cherskaya, A. M., Lyubov A.Wasserman, and, and Permyakov, E. A. (2002) Effect of Zinc and Temperature on the Conformation of the γ Subunit of Retinal Phosphodiesterase: A Natively Unfolded Protein. *J. Proteome Res.* **1**, 149–159

75. Sánchez-Puig, N., Veprintsev, D. B., and Fersht, A. R. (2009) Human full-length Securin is a natively unfolded protein. *Protein Sci.* **14**, 1410–1418

76. Campbell, K. M., Terrell, A. R., Laybourn, P. J., and Lumb, K. J. (2000) Intrinsic Structural Disorder of the C-Terminal Activation Domain from the bZIP Transcription Factor Fos. *Biochemistry*. **39**, 2708–2713

77. Adkins, J. N., and Lumb, K. J. (2002) Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57Kip2. *Proteins Struct. Funct. Genet.* **46**, 1–7

78. Lowry, D. F., Stancik, A., Shrestha, R. M., and Daughdrill, G. W. (2008) Modeling the accessible conformations of the intrinsically unstructured transactivation domain of p53. *Proteins Struct. Funct. Bioinforma.* **71**, 587–598

79. Permyakov, S. E., Millett, I. S., Doniach, S., Permyakov, E. A., and Uversky, V. N. (2003) Natively unfolded C-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin. *Proteins Struct. Funct. Genet.* **53**, 855-Na

80. Paleologou, K. E., Schmid, A. W., Rospigliosi, C. C., Kim, H.-Y., Lamberto, G. R., Fredenburg, R. A., Lansbury, P. T., Fernandez, C. O., Eliezer, D., Zweckstetter, M., and Lashuel, H. A. (2008) Phosphorylation at Ser-129 but Not the Phosphomimics S129E/D Inhibits the Fibrillation of α-Synuclein. *J. Biol. Chem.* **283**, 16895–16905

81. Baker, J. M. R. Structural Characterization and Interactions of the CFTR Regulatory Region

82. Soragni, A., Zambelli, B., Mukrasch, M. D., Biernat, J., Jeganathan, S., Griesinger, C., Ciurli, S., Mandelkow, E., and Zweckstetter, M. (2008) Structural Characterization of Binding of Cu(II) to Tau Protein. *Biochemistry*. **47**, 10841–10851

83. Wu, T.-J., Monokian, G., Mark, D. F., and Wobbe, C. R. (1994) Transcriptional Activation by Herpes Simplex Virus Type 1 VP16 In Vitro and Its Inhibition by Oligopeptides. *MOL CELL BIOL*. **14**, 10

84. Danielsson, J., Jarvet, J., Damberg, P., and Gräslund, A. (2002) Translational diffusion measured by PFG-NMR on full length and fragments of the Alzheimer Aβ(1-40) peptide. Determination of hydrodynamic radii of random coil peptides of varying length: Diffusion of Alzheimer peptides. *Magn. Reson. Chem.* **40**, S89–S97

85. Kawashima, S., Ogata, H., and Kanehisa, M. AAindex: Amino Acid Index Database

86. Dannenhoffer-Lafage, T., and Best, R. B. (2021) A Data-Driven Hydrophobicity Scale for Predicting Liquid–Liquid Phase Separation of Proteins. *J. Phys. Chem. B*. **125**, 4046–4056

87. Das, S., Lin, Y.-H., Vernon, R. M., Forman-Kay, J. D., and Chan, H. S. (2020) Comparative roles of charge, π , and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl. Acad. Sci.* **117**, 28795–28805

88. Loh, S. N., Prehoda, K. E., Wang, J., and Markley, J. L. (1993) Hydrogen exchange in unligated and ligated staphylococcal nuclease. *Biochemistry*. **32**, 11022–11028

89. Sauder, J. M., and Roder, H. (1998) Amide protection in an early folding intermediate of cytochrome c. *Fold. Des.* **3**, 293–301

90. Bhuyan, A. K., and Udgaonkar, J. B. Two Structural Subdomains of Barstar Detected by Rapid Mixing NMR Measurement of Amide Hydrogen Exchange

91. Wlodawer, A., and Sjolin, L. (1982) Hydrogen exchange in RNase A: neutron diffraction study. *Proc. Natl. Acad. Sci.* **79**, 1418–1422

92. Zardecki, C., Dutta, S., Goodsell, D. S., Voigt, M., and Burley, S. K. (2016) RCSB Protein Data Bank: A Resource for Chemical, Biochemical, and Structural Explorations of Large and Small Biomolecules. *J. Chem. Educ.* **93**, 569–575

93. Wang, G., and Dunbrack, R. L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*. **19**, 1589–1591

94. Flory, P. J. (1949) The Configuration of Real Polymer Chains. *J. Chem. Phys.* **17**, 303–310

95. Flory, Paul. J., and Volkenstein, M. (1969) Statistical mechanics of chain molecules. *Biopolymers*. **8**, 699–700

96. English, L. R., Voss, S. M., Tilton, E. C., Paiz, E. A., So, S., Parra, G. L., and Whitten, S. T. (2019) Impact of Heat on Coil Hydrodynamic Size Yields the Energetics of Denatured State Conformational Bias. *J. Phys. Chem. B*. **123**, 10014–10024

97. Langridge, T. D., Tarver, M. J., and Whitten, S. T. (2014) Temperature effects on the hydrodynamic radius of the intrinsically disordered N-terminal region of the p53 protein: Heat Effects on IDP Structure. *Proteins Struct. Funct. Bioinforma.* **82**, 668–678

98. Austin Elam, W., Schrank, T. P., Campagnolo, A. J., and Hilser, V. J. (2013) Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites: Polyproline II Bias and Function in the Proteome. *Protein Sci.* **22**, 405–417

99. Riback, J. A., Bowman, M. A., Zmyslowski, A. M., Knoverek, C. R., Jumper, J. M., Hinshaw, J. R., Kaye, E. B., Freed, K. F., Clark, P. L., and Sosnick, T. R. (2017) Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science*. **358**, 238–241

100. Ibrahim, A. Y., Khaodeuanepheng, N. P., Amarasekara, D. L., Correia, J. J., Lewis, K. A., Fitzkee, N. C., Hough, L. E., and Whitten, S. T. (2022) *Intrinsically disordered regions that drive phase separation form a robustly distinct protein class*, *BioRxiv*, 10.1101/2022.08.04.502866

101. Qian, N., and Sejnowski, T. J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865–884

102. Naderi-Manesh, H., Sadeghi, M., Arab, S., and Moosavi Movahedi, A. A. (2001) Prediction of protein surface accessibility with information theory. *Proteins Struct. Funct. Bioinforma.* **42**, 452–459

103. Pancsa, R., Varadi, M., Tompa, P., and Vranken, W. F. (2016) Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability. *Nucleic Acids Res.* **44**, D429–D434

104. Youn, J.-Y., Dyakov, B. J. A., Zhang, J., Knight, J. D. R., Vernon, R. M., Forman-Kay, J. D., and Gingras, A.-C. (2019) Properties of Stress Granule and P-Body Proteomes. *Mol. Cell*. **76**, 286–294

105. Li, Q., Peng, X., Li, Y., Tang, W., Zhu, J., Huang, J., Qi, Y., and Zhang, Z. (2020) LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro. *Nucleic Acids Res.* **48**, D320–D327

106. Alberti, S., Gladfelter, A., and Mittag, T. (2019) Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell*. **176**, 419–434

107. Banani, S. F., Rice, A. M., Peeples, W. B., Lin, Y., Jain, S., Parker, R., and Rosen, M. K. (2016) Compositional Control of Phase-Separated Cellular Bodies. *Cell*. **166**, 651–663

108. Boeynaems, S., Alberti, S., Fawzi, N. L., Mittag, T., Polymenidou, M., Rousseau, F., Schymkowitz, J., Shorter, J., Wolozin, B., Van Den Bosch, L., Tompa, P., and Fuxreiter, M. (2018) Protein Phase Separation: A New Phase in Cell Biology. *Trends Cell Biol.* **28**, 420–435

109. Lancaster, A. K., Nutter-Upham, A., Lindquist, S., and King, O. D. (2014) PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics*. **30**, 2501–2502

110. Orlando, G., Raimondi, D., Tabaro, F., Codicè, F., Moreau, Y., and Vranken, W. F. (2019) Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates. *Bioinformatics*. **35**, 4617–4623

111. Chu, X., Sun, T., Li, Q., Xu, Y., Zhang, Z., Lai, L., and Pei, J. (2022) Prediction of liquid–liquid phase separating proteins using machine learning. *BMC Bioinformatics*. **23**, 72

112. Ibrahim, A. Y. SEQUENCE-BASED PROPERTIES THAT IDENTIFY INTRINSICALLY DISORDERED PHASE-SEPARATING PROTEIN REGIONS(Master's thesis).Dep.Chem.Biochem.Grad.Coll.Sci.Eng.Texas State univ.**2022**

113. Yuan, C., Levin, A., Chen, W., Xing, R., Zou, Q., Herling, T. W., Challa, P. K., Knowles, T. P. J., and Yan, X. (2019) Nucleation and Growth of Amino Acid and Peptide Supramolecular Polymers through Liquid–Liquid Phase Separation. *Angew. Chem. Int. Ed.* **58**, 18116–18123

114. Goswami, D., Devarakonda, S., Chalmers, M. J., Pascal, B. D., Spiegelman, B. M., and Griffin, P. R. (2013) Time Window Expansion for HDX Analysis of an Intrinsically Disordered Protein. *J. Am. Soc. Mass Spectrom.* **24**, 1584–1592

115. Balasubramaniam, D., and Komives, E. A. (2013) Hydrogen-exchange mass spectrometry for the study of intrinsic disorder in proteins. *Biochim. Biophys. Acta BBA - Proteins Proteomics*. **1834**, 1202–1209

116. Masson, G. R., Burke, J. E., Ahn, N. G., Anand, G. S., Borchers, C., Brier, S., Bou-Assaf, G. M., Engen, J. R., Englander, S. W., Faber, J., Garlish, R., Griffin, P. R., Gross, M. L., Guttman, M., Hamuro, Y., Heck, A. J. R., Houde, D., Iacob, R. E., Jørgensen, T. J. D., Kaltashov, I. A., Klinman, J. P., Konermann, L., Man, P., Mayne, L., Pascal, B. D., Reichmann, D., Skehel, M., Snijder, J., Strutzenberg, T. S., Underbakke, E. S., Wagner, C., Wales, T. E., Walters, B. T., Weis, D. D., Wilson, D. J., Wintrode, P. L., Zhang, Z., Zheng, J., Schriemer, D. C., and Rand, K. D. (2019) Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat. Methods*. **16**, 595–602

117. Perrett, S., Clarke, J., Hounslow, A. M., and Fersht, A. R. (1995) Relationship between Equilibrium Amide Proton Exchange Behavior and the Folding Pathway of Barnase. *Biochemistry*. **34**, 9288–9298

118. Kern, G., Handel, T., and Marqusee, S. (1998) Characterization of a folding intermediate from HIV-1 ribonuclease H. *Protein Sci.* **7**, 2164–2174

119. Delacre, M., Lakens, D., and Leys, C. (2017) Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *Int. Rev. Soc. Psychol.* **30**, 92

120. Hart, A. (2001) Mann-Whitney test is not just a test of medians: differences in spread can be important. *BMJ*. **323**, 391–393

121. Ptitsyn, O. B., and Finkelstein, A. V. (1983) Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*. **22**, 15–25

122. Franzmann, T. M., Jahnel, M., Pozniakovsky, A., Mahamid, J., Holehouse, A. S., Nüske, E., Richter, D., Baumeister, W., Grill, S. W., Pappu, R. V., Hyman, A. A., and Alberti, S. (2018) Phase separation of a yeast prion protein promotes cellular fitness. *Science*. **359**, eaao5654

123. Wang, J., Choi, J.-M., Holehouse, A. S., Lee, H. O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., Poser, I., Pappu, R. V., Alberti, S., and Hyman, A. A. (2018) A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell*. **174**, 688-699.e16

124. Muiznieks, L. D., and Keeley, F. W. (2016) Phase separation and mechanical properties of an elastomeric biomaterial from spider wrapping silk and elastin block copolymers: Spider Wrapping Silk and Elastin Block Copolymers. *Biopolymers*. **105**, 693–703

125. Tremblay, M.-L., Xu, L., Lefèvre, T., Sarker, M., Orrell, K. E., Leclerc, J., Meng, Q., Pézolet, M., Auger, M., Liu, X.-Q., and Rainey, J. K. (2015) Spider wrapping silk fibre architecture arising from its modular soluble protein precursor. *Sci. Rep.* **5**,11502

126. Brady, J. P., Farber, P. J., Sekhar, A., Lin, Y.-H., Huang, R., Bah, A., Nott, T. J., Chan, H. S., Baldwin, A. J., Forman-Kay, J. D., and Kay, L. E. (2017) Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc. Natl. Acad. Sci.* 10.1073/pnas.1706197114

127. Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C. C.-H., Eckmann, C. R., Myong, S., and Brangwynne, C. P. (2015) The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl. Acad. Sci.* **112**, 7189–7194

128. Kim, Y., and Myong, S. (2016) RNA Remodeling Activity of DEAD Box Proteins Tuned by Protein Concentration, RNA Length, and ATP. *Mol. Cell*. **63**, 865–876

129. Matsumoto, T., Morimoto, Y., Shibata, N., Kinebuchi, T., Shimamoto, N., Tsukihara, T., and Yasuoka, N. (2000) Roles of Functional Loops and the CD-Terminal Segment of a Single-Stranded DNA Binding Protein Elucidated by X-Ray Structure Analysis. *J. Biochem. (Tokyo)*. **127**, 329–335

130. Lin, Y., Protter, D. S. W., Rosen, M. K., and Parker, R. (2015) Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol. Cell*. **60**, 208–219

131. Chandonia, J.-M., Fox, N. K., and Brenner, S. E. (2019) SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Res.* **47**, D475–D481

132. Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., Aspromonte, M. C., Davey, N. E., Davidović, R., Dosztányi, Z., Elofsson, A., Gasparini, A., Hatos, A., Kajava, A. V., Kalmar, L., Leonardi, E., Lazar, T., Macedo-Ribeiro, S., Macossay-Castillo, M., Meszaros, A., Minervini, G., Murvai, N., Pujols, J., Roche, D. B., Salladini, E., Schad, E., Schramm, A., Szabo, B., Tantos, A., Tonello, F., Tsirigos, K. D., Veljković, N., Ventura, S., Vranken, W., Warholm, P., Uversky, V. N., Dunker, A. K., Longhi, S., Tompa, P., and Tosatto, S. C. E. (2017) DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227

133. Cai, H., Vernon, R. M., and Forman-Kay, J. D. (2022) An interpretable machine-learning algorithm to predict disordered protein phase separation based on biophysical interactions. *Biomolecules*. **12**, 1131

134. Chen, Z., Hou, C., Wang, L., Yu, C., Chen, T., Shen, B., Hou, Y., Li, P., and Li, T. (2022) Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proc. Natl. Acad. Sci.* **119**, e2115369119

135. Saar, K. L., Morgunov, A. S., Qi, R., Arter, W. E., Krainer, G., Lee, A. A., and Knowles, T. P. J. (2021) Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc. Natl. Acad. Sci.* **118**, e2019053118

136. Balasubramaniam, D., and Komives, E. A. (2013) Hydrogen-exchange mass spectrometry for the study of intrinsic disorder in proteins. *Biochim. Biophys. Acta BBA - Proteins Proteomics*. **1834**, 1202–1209

137. Bai, Y., Milne, J. S., Mayne, L., and Englander, S. W. (1993) Primary structure effects on peptide group hydrogen exchange. *Proteins Struct. Funct. Genet.* **17**, 75–86

138. Englander, S. W. (2006) Hydrogen exchange and mass spectrometry: A historical perspective. *J. Am. Soc. Mass Spectrom.* **17**, 1481–1489

139. Gallagher, W., Tao, F., and Woodward, C. (1992) Comparison of hydrogen exchange rates for bovine pancreatic trypsin inhibitor in crystals and in solution. *Biochemistry*. **31**, 4673–4680

140. Karplus, P. A., and Schulz, G. E. (1985) Prediction of chain flexibility in proteins: A tool for the selection of peptide antigens. *Naturwissenschaften*. **72**, 212–213