Byungkook Lee, Ph. D.
Molecular Modeling and Bioinformatics Section

I. OVERVIEW

Scientific

We worked mainly on two specific problems in the past two years. Both are in the area of protein structure studies. One is on how to find internally symmetric proteins. The other is on how to break a protein structure into smaller pieces called domains. The two seem unrelated, but we are finding that they may be more related than one would suspect at first.

We began to write a better structure-based sequence alignment routine some time ago, which became the program SE for seed extension. The algorithm for this is based on my experience with manually aligning protein structures using an in-house graphics program GEMM. When SE became RSE (Recursive SE) by including iteration process, and after seeing the characteristics of RSE, I became convinced that it could be used to develop a neat internal symmetry detecting tool. After working on it for nearly one and a half years, the development is now nearly complete. And after examining some of the outputs that the program produces, I am convinced that it is qualitatively superior to anything that has been described in the literature so far.

The impact of the availability of such a program is difficult to gauge, but could be large. We now have a tool that one can use to divide protein structures into two groups, those symmetric and those that are not. This separation is useful because there are operations, domain parsing for example (see attached manuscript #2), that work well with non-symmetric proteins but poorly with the symmetric ones. We can also begin to catalogue and classify symmetric proteins, which is what we propose to do in the immediate future. We hope that this will substantially increase our understanding of the symmetric structures. As I argue in the main body of the report, symmetric proteins can offer some definite advantages in studying topics that are of importance today, such as protein-protein interaction. The availability of a pool of symmetric structures can facilitate such surrogate studies.

Switching to the other problem, domain parsing is an old problem that just will not go away. We absolutely need domains and domain parsing operation. But the problem has not been solved in a satisfactory fashion after some 30 years of effort. We think that at least part of this difficulty lies in the fact that domains have been defined ultimately by subjective criteria. Essentially all automatic procedures rely on recognizing groups of residues that are ‘separated’ from others either geometrically and/or energetically. But there is no precise, objective guidance on how to decide when groups are separated enough to be in separate domains. The criterion of recurrence is different at least in principle. Here, one defines a group of residues as a domain if they occur together in otherwise unrelated structures. The decision is made by observation rather than by judgment, in principle. Our ambition, when we started, was to develop a domain parsing

procedure using recurrence and thereby provide an authoritative voice and put a stop to all other definitions.

This of course did not happen but we made an unexpected discovery during this effort. As detailed in the attached manuscript and in the body of this report, we found that a large collection of locally similar structural pieces, which we named LSSPs, defines domains nearly as well as other programs that operate on the principle of separation. LSSPs are small pieces, typically covering only about 10% to 20% of the domain, but containing 3 or more secondary structure elements. We did not expect that a collection of such small pieces can define domains so well and we are still trying to figure out why. However, we are suddenly becoming aware of many occasions where LSSP-like pieces figure prominently. As described in the body of this report and in the attached manuscript, these include the fragments used in the highly successful fragment assembly method of protein structure prediction[1; 2; 3], the SSS (super secondary structure) library of Szustakovski et al.[4], which are built from a collection of what looks like our LSSPs and which one can use like a lego set to build most protein domains, the LSSP-like pieces of Petrey et al.[5] that seem to move through proteins of different folds carrying a common function, and finally the proposal by Lupas et al.[6] that domains originated from a conglomerate of LSSP-like ADSs (antecedent domain segments).

These separate reports seem to point to the common idea that domains are made of some non-random collection of small pieces that preserve their structure, and perhaps the sequence and function as well, across different protein folds. If so, symmetric proteins share some common features since they are also made of small units. One may think of the symmetric proteins as homomers of small units and the domains as heteromers of small units. These ideas also give us some confidence that domains can, or perhaps should, be defined by means of the LSSPs.

Administrative

We continue to collaborate with Dr. Peter Munson at the Center for Information Technology (CIT) of NIH and with Drs. Jean Garnier and Jean-Francois Gibrat of the Institut National de la Recherche Agronomique, Jouy-en-Josas, France on domain definition work.

Since 2007, a staff scientist has left for a new job and two post-doctoral fellows have left after completing their training. They are now the Director of the Bioinformatics Center at the University of Vermont, an Assistant professor at a University in Seoul, Korea, and a program director at a non-profit research institute in Seoul, Korea, respectively. We added one post-doctoral fellow nearly two years ago. She will be leaving in May. Anticipating this we recently hired two new Ph.D.s, who are currently receiving post-doctoral training.

II. SUMMARY OF ACTIONS TAKEN IN RESPONSE TO PREVIOUS SITE VISIT

The main criticism at the last site visit two years ago was that my program lacked focus and that the problems we were working on were unlikely to have a high impact. I took this criticism seriously and decided to concentrate. We have worked mainly on only two projects in the past two years. They are both in the area of protein structure, require some level of mathematics, and are not unrelated to each other.
I do not know if these will make a high impact, but I find them interesting and feel as if we are opening the door to a new world.

III. RESEARCH SUMMARY

PROJECT 1: PROTEIN STRUCTURE STUDIES –SYMMETRIES AND DOMAINS

Background

Three of the cancer gene products we studied, POTE, CAPC, and mesothelin, are predicted to contain structures that are made of repeating units arranged in a superhelical manner: POTE with the ankyrin repeats[7], CAPC with the leucine-rich repeats[8], and mesothelin with the probable ARM-type alpha/alpha superhelix repeats[9].

These are internal symmetries, which exist in monomeric proteins, as distinct from the symmetries of multimeric complexes formed by symmetrically arranging non-symmetric monomers. The internally symmetric proteins are interesting objects to study for a number of reasons. Firstly, one wonders what sequence features produce the symmetry. The repeating units often have high sequence similarity, which will make them to have a common structure, although there are many, including some alpha/alpha superhelices, for which the repeats are difficult to recognize from sequence alone. But what makes them to be arranged in a symmetric manner? The secret probably does not lie only in the linker region between the repeats since these are usually flexible structures. The repeating units themselves must have codes that dictate the type of symmetry, in addition to the code for folding of the individual units.

Most symmetric proteins have a relatively small core unit, which is repeated. These are simple structures compared to proteins that are not symmetric. Yet, they appear to be capable of carrying out all types of functions. Some are enzymes, others are carriers of proteins, still others are receptors, etc. Therefore, if one is interested in designing proteins *de novo* to perform a specific function, symmetric proteins are probably a good start. They should also be good molecules with which to study the sequence-structure-function relations because of their relative simplicity.

The evolutionary history of these proteins is also interesting. These proteins most probably arose by gene duplication and fusion[10]. Although mutation rate will be different depending on the requirement of symmetry for function, generally those that have highly

sequence similar repeats presumably arose late, compared to those for which the similarity is beginning to disappear. After sufficient time, the sequence similarity will disappear and structural symmetry will also be degraded. Thus, the symmetry should generally give an additional handle for following the evolution of these proteins.

The interest in symmetric structures seems to be rising; there were only a few reports on symmetry detection prior to 2008[11; 12; 13; 14; 15], but at least three different groups reported separate symmetry detection methods in the past two years[16; 17; 18; 19]. We will be reporting our own method soon (See the attached manuscript #1.).

In collaboration with Drs. Garnier and Gibrat, both at INRA, France, and Dr. Munson at CIT, NIH, we have been studying the problems of automatic protein structure classification[20; 21] and domain parsing. Protein structures need to be broken into domains before they can be compared, classified, and their function understood in molecular terms. The problem of automatically breaking a known protein structure into domains has been studied at least from the late 1970s[22; 23]. However, this task remains a topic of active investigation today because it is important and still has not been satisfactorily resolved[24]. A large review on the topic has recently been published[25].

All known automatic methods basically work by finding geometrically and/or energetically separable modules in the structure. Defining domains in this manner is intuitively appealing, but the definition depends on what we consider 'separable' and becomes ultimately subjective. We sought to define domains more objectively as the group of residues that appear together in different protein structures. This is the principle of recurrence, which in principle is a more objective way of defining domains since it depends on whether the same (or substantially the same) group of residues appears in other structures or not, regardless of whether that group satisfies our pre-conceived notion of what a domain shall look like. The principle is used in manual domain parsing but not used in any automatic procedures, except one[26], in which the recurrence is used only to augment their otherwise geometric procedure.

While working on defining domains using this principle, we made an important, unexpected discovery. We found (1) that only relatively few other structures have the entire domain in a given query structure, as expected, and a relatively large number of other structures that contain parts of the domain or parts from two or more domains. The domain definition becomes difficult under such a circumstance, which probably explains why others have not used this principle in the past. However, we also found (2) that we can use a large number of locally similar structural pieces (LSSPs), which can collectively define domains. (See the attached manuscript #2.)

LSSPs are small sets of three or more secondary structural elements, which have become popular lately, as described in the Overview section above. Our finding that LSSPs can define domains appears to be consistent with these reports and adds weight to the importance of LSSPs in the structure, the evolutionary history, and the function of domains.

Specific Research Aims

Sub-project 1-1

To find internally symmetric proteins in the protein structure database and to study their structure, symmetry, sequence-structure relation, structure-function relation, and their evolutionary history.

Sub-project 1-2

To recognize and define protein structural domains using locally similar structural pieces (LSSP) and to find new relations between domains through shared LSSPs.

Accomplishments

Sub-project 1-1

As described in the last site visit report and now in a published report[27], we found that most structure-based sequence alignment programs, including our own program SHEBA[28], made a significant number of errors in the alignments they produced. The best among those we tested (DaliLite[29]) is also extremely slow. We examined SHEBA and found that the major problem, at least for this program, was in the dynamic programming procedure that was used to obtain the sequence alignment from the superposed structures. Therefore, we designed a new algorithm, which we call Seed Extension (SE), to do just this part of the operation. The algorithm was described in the last site visit report and has now been published[30]. SE works by finding seed alignments and extending them, does not use dynamic programming or a gap penalty, and is extremely fast. It produces significant improvements, making SHEBA comparable to DaliLite, in terms of the accuracy of the sequence alignments it produces, at a fraction of the computing time. We have since made a sequence alignment refinement package, which performs iterations of (1) structure superposition from a given sequence alignment and (2) sequence alignment from the new superposed structures using SE. This package, called RSE[31], can be run as a stand-alone post-processing procedure to improve the sequence alignment output from any structure comparison programs, including very fast ones such as FAST[32] and MATRAS[33], to the level of DaliLite, without adding a significant amount of computing time.

The availability of RSE was crucial for the development of the algorithm for detecting symmetric proteins. The algorithm, called SymD (for Symmetric protein Detection), works by aligning, using RSE, a protein structure to itself after circularly permuting the second copy by k residues for all k values from 1 to N-3 residues where N is the total number of residues of the protein. The input to each of the RSE procedure is the pair of sequences, the original and the one permuted by k residues. For each circular shift, we keep only one optimal, non-self structural alignment, fully allowing gaps and unaligned loops. We call this process the alignment scan. Non-symmetric structures should not give a high score, as measured by the number of residues aligned or something similar, at any

shift. But the symmetric ones should yield high score every *n*-th shift where *n* is the number of residues in one repeating unit. The details of this procedure are described in the manuscript #1 attached. Fig. I-1 and Fig. I-2 give some examples.

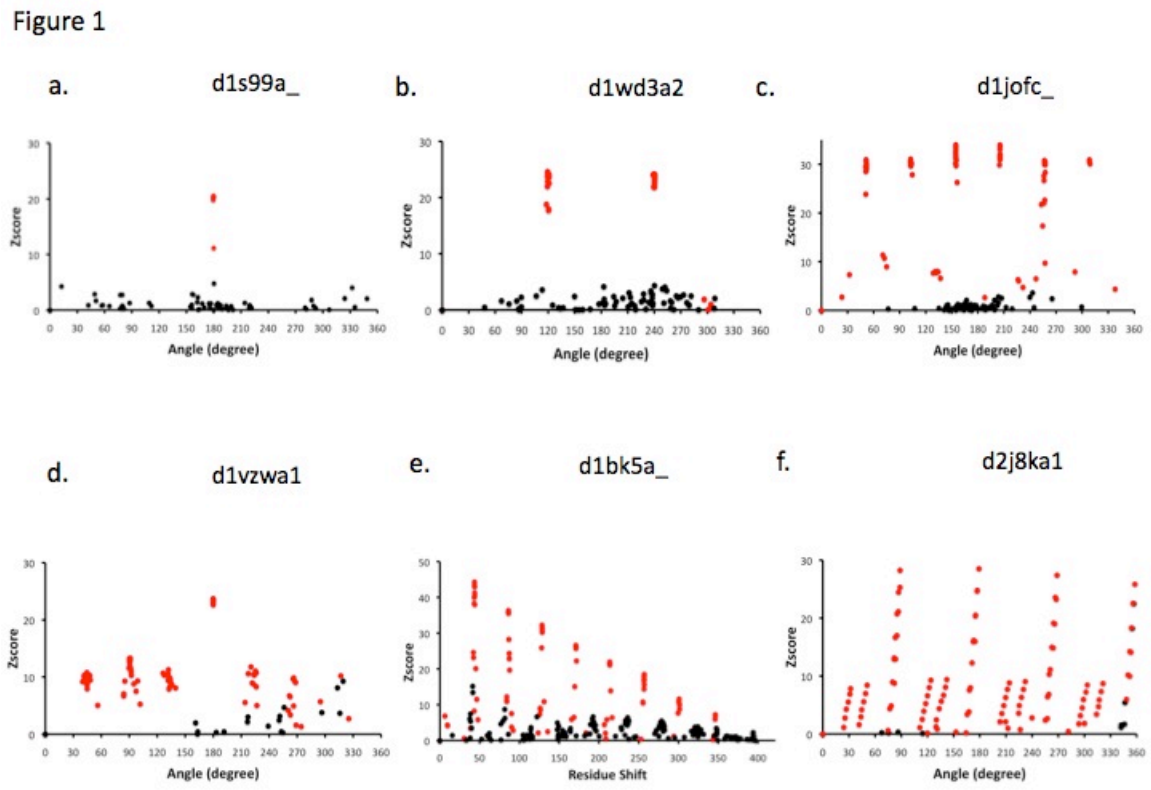

Figure I-1. The Z-score vs. rotation angle scatter plot for all alignments from the alignment scan for (a) d1s99a_, (b) d1wd3a2, (c) d1jofc_, (d) d1vzwa1, and (f) d2j8ka1. The red points are those whose rotation axis is within about 20° ($\cos\theta>0.95$) of that of the point with the highest Z-score. Others are black. Panel (e) for d1bk5a_, is an exception; here the Z-scores are plotted against the average alignment shift (average of the residue serial number difference between aligned residues).

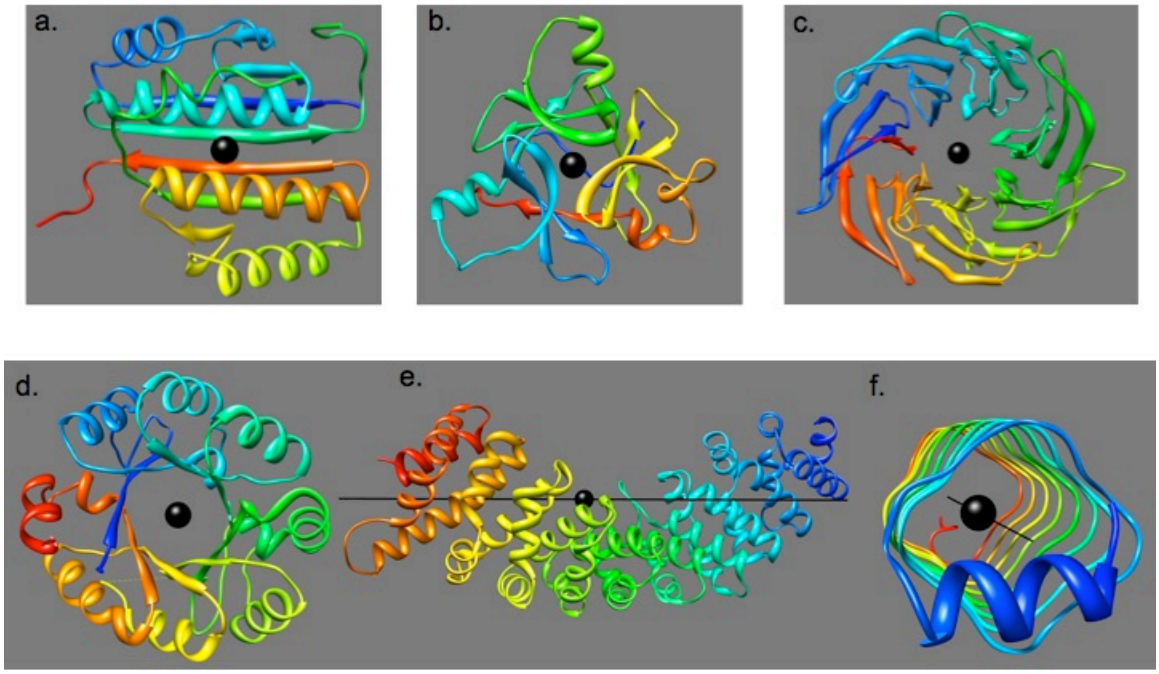

Figure I-2. The ribbon rendering of the structure of the proteins of Fig. I-1: (a) d1s99a_, 2-fold symmetric ferredoxin-like fold, (b) d1wd3a2, a 3-fold symmetric beta-trefoil, (c) d1jofc_, a 7-bladed beta-propeller, (d) d1vzwa1, a 2-fold symmetric TIM barrel, (e) d1bk5a_, an alpha/alpha superhelix, and (f) d2j8ka1, a right-handed beta-helix with square cross-section. The ribbons are colored in rainbow colors, starting from the blue N-terminus to the red C-terminus. The calculated symmetry axis is shown as a black rod with a ball at the center.

There are a handful of procedures that have been used to detect symmetric proteins[11; 12; 13; 14; 15; 16; 17; 18; 19]. The SymD procedure is superior to these methods in several aspects (see the attached manuscript #1 for details): (1) The procedure allows detection of symmetry even when the structure contains symmetry-breaking insertions or deletions either within or between the repeating units. This is because it uses SE, which allows gaps of unlimited size. Some of the other algorithms are forced to assume no or only small gaps. (2) The procedure depends and uses the symmetry of the molecule. Suppose the structure is made of two similar units A and B. If the second copy is circularly permuted by N/2 residues, it has the structure B-A, and if and only if the structure is symmetric, it will match the original structure A-B in its entirety. All other programs that we know of detect repeats rather than symmetry. (3) The procedure amplifies symmetric signal. Suppose the structure is 6-fold symmetric but one of the repeating units is somewhat different from the rest. If the difference is large, most straight repeat-detection algorithms might fail to recognize this unit. However, the SymD algorithm will still

recognize the 6-fold symmetry since the alignment scan will report 5 matches out of perfect 6 at every 60° rotation. (4) The procedure yields both the sequence and structural alignments after each symmetry operation. The sequence alignment will give information on the residues that make up the repeating units. The structural alignment, or the structure transformation matrix, contains the information on the direction and position of the symmetry axis, the rotation angle, and the pitch if the symmetry is that of a helix. Programs that detect repeats by, for example, a Fourier transform do not yield such information. (5) The procedure is capable of detecting more than one symmetry for a molecule as detailed in the attached manuscript using the examples of a 2-fold symmetric TIM barrel and a beta-helix structures. As far as we know, no other program has this capability.

SymD was run on all 9,479 domains in the SCOP1.73 ASTRAL 40% domain dataset[34]. It finds that symmetric domains make up between 10% and 15% of the dataset, depending on the cutoff value one uses on a variable (Z-score of a TM-like score) that measures the degree of perfection of the symmetry. Perhaps not coincidentally, this range brackets 14% found for proteins with repeating units in the whole protein sequence database[35]. Table I-1 gives the list of SCOP[36] folds that have at least 10 symmetric domains using the more generous cutoff value of 8, sorted in decreasing number of symmetric domains using the more stringent cutoff value of 10.

Table I-1. SCOP Folds with 10 or more symmetric domains (at Z-score cutoff of 8)

Scop Id	*Zscore >=8[a]*	*Zscore >=10[b]*	*Total[c]*	*Fold Name*
c.1	268	223	322	TIM beta/alpha-barrel
a.118	51	45	94	alpha-alpha superhelix
b.42	41	39	41	beta-Trefoil
b.69	35	35	35	7-bladed beta-propeller
a.102	39	33	42	alpha/alpha toroid
a.25	34	29	43	Ferritin-like
b.80	27	27	29	Single-stranded right-handed beta-helix
b.68	27	27	27	6-bladed beta-propeller
c.10	24	21	25	Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)
f.4	19	18	20	Transmembrane beta-barrels
d.58	58	16	302	Ferredoxin-like
a.24	33	16	65	Four-helical up-and-down bundle
d.131	20	16	22	DNA clamp
d.211	15	14	17	beta-hairpin-alpha-hairpin repeat
b.82	13	13	82	Double-stranded beta-helix
c.94	18	12	52	Periplasmic binding protein-like II
a.2	18	12	40	Long alpha-hairpin
b.81	12	12	16	Single-stranded left-handed beta-helix
a.7	23	11	40	Spectrin repeat-like
c.93	12	11	15	Periplasmic binding protein-like I
d.126	12	11	12	Pentein, beta/alpha-propeller

d.19	10	10	15	MHC antigen-recognition domain
b.67	10	10	10	5-bladed beta-propeller
a.26	15	9	28	4-helical cytokines
a.47	11	9	11	STAT-like
b.50	10	9	15	Acid proteases
a.29	14	8	25	Bromodomain-like
d.157	12	8	21	Metallo-hydrolase/oxidoreductase
a.39	14	6	58	EF Hand-like
b.40	10	3	126	OB-fold
d.32	10	3	31	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase
b.1	33	2	369	Immunoglobulin-like beta-sandwich
c.2	15	2	193	NAD(P)-binding Rossmann-fold domains

[a]Number of domains with Z-score >= 8.
[b]Number of domains with Z-score >= 10.
[c]Total number of domains in the fold.

The symmetries observed are broadly of two types, closed and open. In symmetric closed structures, the N- and C-termini of the molecule come close together and the two ends of the molecule are 'stitched' together, often by using a set of hydrogen bonds (the Velcro joining). Most of these have 2- to 8-fold rotational symmetries, but the transmembrane beta-barrels can have higher symmetries and also the screw symmetries. In the symmetric open structures, the N- and C-termini are at the opposite ends of the molecule. All have a helical or a pure 2-fold rotational symmetry. A protein with a pure 2-fold rotational symmetry can have either a closed (intertwined) or an open structure.

We have yet to fully analyze the results, but more detail is available in the attached manuscript #1.

Sub-project 1-2

Domains are the essential basic units of protein structure. We need them for properly exploring the fold space and to understand and organize the relations between and among different protein structures. They are also essential for exploring the evolutionary history of protein structures. However, after nearly 40 years since the concept was introduced[37], domains are still difficult to define quantitatively[24; 25; 38; 39; 40; 41]: the same protein structure can be partitioned into different sets of domains by different people or programs and SCOP[36] and CATH[42], the two well known protein domain databases, have significant differences in their domain assignments.

Many automatic domain partition procedures have been reported (see a recent review by Veretnik et al.[28] and a Ph. D. thesis by Todd Taylor[43]) They all use the concept of a domain as a geometrically and/or energetically separable module of the protein structure. Since different methods that use this principle have so far failed to yield a consistent result, we and our collaborators (Jean Garnier and Jean-Francois Gibrat of INRA, France and Peter Munson of CIT, NIH) sought initially to define domains using the recurrence

principle, i.e. a domain is the region of the protein structure that exists in other proteins in a different context. Among the automatic domain partition procedures that we know of, only DDD (Dali Domain Database)[26] uses this principle to augment their otherwise geometrical/energetic PUU[44] domain partition procedure.

We used the structure comparison program VAST[20; 45] for this purpose because it produces multiple hits, or 'cliques', per protein pair, each representing a set of aligned residue pairs. We initially used only significant hits using what we deemed was a reasonable set of cutoff criteria. However, we basically failed in this attempt. We found that, for many of the domains we tested, target structures were found that matched only parts of a domain and others that matched parts of two or more domains. Number of such hits was not negligible compared to that of those that properly covered only the known domain, which made it difficult to determine precise domain boundaries and in many cases even the number of domains.

However, we saw domain boundaries emerge clearly when we (1) increased the number of cliques by accepting almost every clique that VAST produced and (2) padded small gaps within each LSSP (Locally Similar Structural Pieces = the query residues in the clique, see below) with the query sequence, wrote the position of the residues in the padded LSSPs (pLSSPs) on the query sequence as a binary matrix (matrix A), and then transformed it to produce the co-occurrence matrix ($N=A^T*A$). Fig. II-1 shows the unpadded, padded and sorted maps of LSSPs, as well as the N-matrix and the domain structure, for a sample test protein, 1jjcB (phenylalanyl-tRNA synthetase, B chain). We developed three different mathematical methods to extract the precise domain boundary information from the N-matrix. One (by JFG) uses the well known matrix factorization process called the Singular Vector Decomposition (SVD), another (by BL) uses a new Symmetric Matrix Factorization (SMF) procedure, and the third (Pair Correlation Method or PCM, by PM) uses a special weighting method from multivariate statistical correlation. All performed at a similar level and approximately as well as other existing domain partition programs when tested on a dataset from the literature[24] using the NDO score[46]. (See Fig. II-1, panel (e) for example.) More details can be found in the attached manuscript #2.

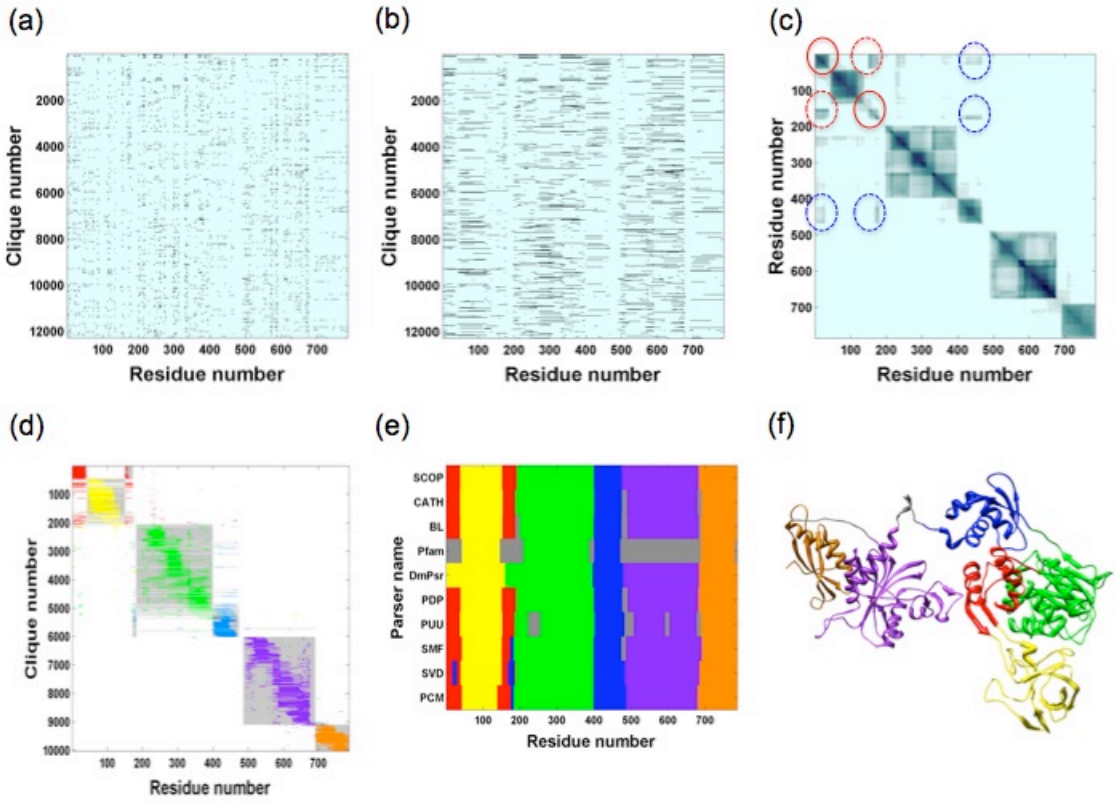

Fig. II-1. The six panels of this Figure all pertain to 1jjcB. (a) Location of LSSPs (Locally Similar Structural Pieces) on the sequence of 1jjcB. The Y-axis is the LSSP serial number and the X-axis is the residue number. Black lines indicate the residues in 1jjcB that are aligned in the VAST cliques produced by the target chains. There are 12,282 cliques for which the RMSD is less than 4 Å. (b) Location of pLSSPs (gap-filled or "padded" LSSP). This is also the *A*-matrix, which has the value 1 on the line segments and 0 outside. (c) The heat map of the co-occurrence *N*-matrix of 1jjcB. The X and Y-axes are the residue numbers of 1jjcB. The pixel intensity indicates the value of the matrix element, which is the number of pLSSPs that contain both of the residues represented by the pixel position. The two squares along the diagonal, circled in red, indicate two segments of a segmented domain, which also produces the off-diagonal intensities, indicated by the red dotted circles. This is the domain indicated in red color in panels d, e, and f. The off-diagonal intensities dot-circled in blue indicate that there are pLSSPs that span the red domain and another domain, which is colored in blue in panels d, e, and f. (d) A sorted map of pLSSPs, colored according to the domain assignments made by SMF. The pLSSPs were sorted in ascending order of the mean of the serial numbers of the residues in the pLSSP. The gray shading indicates the boundaries of the domains. (e) The ribbon bar chart shows the domain boundaries of 1jjcB according to SCOP, CATH, visual inspection by one of the authors (BL), Pfam, and different domain partition programs indicated by their name (DmPsr for DomainParser). Residues in the

same domain, which are sometimes separated, are colored in the same color. The grey areas indicate the residues that do not belong to any domain or, for Pfam, any protein family. (f) The structure of 1jjcB colored according to the CATH domains. Same colors are used for the same domains consistently in panels d, e, and f.

Thus, we are led to the conclusion that domains can be defined using many cliques from VAST. We examined the pLSSPs and found that most are short and heavily padded. For example, the pLSSPs that cover a 207 residue domain of 1jjcB are on average 56 residues long, about half of which are padded residues. This means that the target structure has only about 28 residues on average that match the query domain of 207 residues. Thus, the target structures contain locally similar structural pieces (LSSPs), but most do not resemble the target domain as a whole. But there are very many such pieces. The number of LSSPs that cover the particular domain mentioned above is 3633 from 2244 target structures (each target generates multiple cliques) out of the 6373 total number of target structures in our target database. For the whole test dataset, the average number of LSSPs is over 6,000 per domain. (See the attached manuscript #2 for more detail.)

VAST cliques typically contain 3 or more secondary structural elements and do not contain unaligned residues in the loops between them. It was totally unexpected, and we find it quite striking, that such short bits and pieces of locally similar sub-structures can collectively define domains. Fig. II-2 shows a couple of examples of LSSPs that contribute to the definition of the domain of 1jjcB mentioned above.

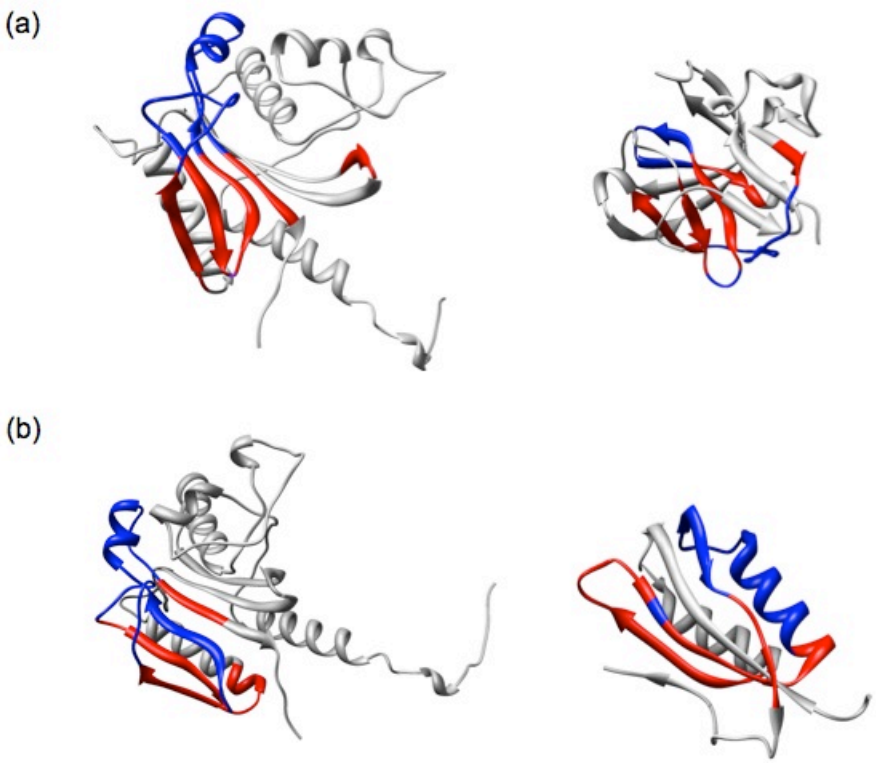

Fig. II-2. Two examples of typical LSSPs and target structures for the 1jjcB domain of residues 475 – 681, colored purple in Figs. 1d, 1e, and 1f. Both examples have the pLSSP length of 56 residues, the mean of all pLSSPs for this domain, and the number of aligned residues of 26, the average for this domain. Aligned residues are colored red, padded residues blue, and others grey. In each panel, the left structure is the 1jjcB purple domain and the right one the target. (a) 1urrA residues 34 – 38, 39 – 46, 48-56 and 81 – 84 align to 1jjcB residues 617 – 621, 622-629, 637-645 and 669-672 (b) 2hrvA residues 38-41, 56-61, 108-113, 115-120, 130-133 align to 1jjcB residues 576 – 579, 623-328, 640-645, 647-652, and 671-674.

We are intrigued by this finding not only because it provides a new avenue of defining domains, but also because it is consistent with a number of reports that indicate the important role that small structural pieces play in the structure, function and etiology of protein domains. Thus, Lucas et al.[6] proposed that domains originated from a conglomerate of polypeptide pieces that they call antecedent domain segments (ADS). Their reasoning is basically two-fold. First, they noted that the modern proteins with repeating units must have arisen from ancient proteins that had only one of these units, which must have functioned by forming homo-oligomeric complexes. If functioning homo-oligomeric complexes existed, it is not unreasonable to imagine that similarly functioning hetero-oligomeric complexes might also have existed, which in time became single chain proteins through the genetic process of gene fusion. Secondly, they could find a large number of short sequence motifs each of which exist in a common structural

form in many protein domains with different structural types and to which a common function could be associated. These are the ADSs whose sequence signature was preserved presumably because of their association with function. It is tempting to imagine that their ADSs are part of our LSSPs. They also note the success of fragment assembly method in protein structure prediction[1; 2; 3]. The fact that the fragment assembly method works indicates that protein domains are made of fragments that exist in other proteins which do not necessarily have the overall structure of the protein of interest. Szustakowski et al.[4] reported that they could construct a set of super-secondary structural pieces (SSSs) from current structural database, which can describe the structure of a substantial fraction of all known folds. They prefer to explain the existence of common SSSs among different structural folds as the result of convergent evolution towards structural attractors, similar to those described by Holm and Sander[47] for entire domain structures. Recently, Petrey et al.[5] also reported that fragments that contain 3 or more secondary structural elements with common function could be found in proteins with different folds. In view of these reports, our finding might have been expected; defining domains by their constituent structural pieces could, in fact, be a most natural way of defining domains.

Current Research and Future Plans

1. <u>Development of a new structure comparison/alignment program</u>

Our experiences with SymD for finding and characterizing symmetric proteins show that the alignment scan procedure is a powerful and efficient technique for sampling and locating the local minima in the space of relative position/orientation between a protein and its permuted self. I expect that it will be equally powerful and efficient in finding local structural matches between two different protein structures. I propose that we test this idea.

The procedure is a simple modification of SymD: Call the larger of the two proteins A and the smaller one B. Run RSE starting from the initial alignment in which a small number of the N-terminal residues of A are aligned to the same number of the C-terminal residues of B. Repeat the procedure after shifting the position of the smaller protein by one residue to the right so that the number of residues aligned in the initial alignment is increased by one at each iteration, until whole B is completely embedded in A. Continue to repeat the procedure so that position of B is moved across A, then the C-terminal of B begins to slip off of the C-terminal of A, until finally B is completely off of A. We keep the highest Z-score alignment at each step and report all distinct alignments with Z-score higher than a cutoff value. In order to gain in speed, one can do the scan by shifting, for example, 5 residues at a time rather than 1; then, once a good Z-score step is identified, one can do a finer search around the initial alignment that produced the good alignment.

The virtue of this new structure alignment procedure will be that it will produce highly accurate structure-based sequence alignments, since it uses SE and RSE. Another is that it will naturally report all local alignments, not just one globally optimal alignment, each with a simple quality index in the form of the Z-score. In addition, it will be very fast.

From our point of view, the fourth advantage is that it would be easy for us to develop this procedure. What is uncertain at this stage is how completely the procedure covers the search space, particularly when two large multi-domain proteins are compared.

2. Development of a full domain partition program

The domain partition routines we wrote so far are for exploring the possibility of using structural recurrence to define domains. As such, they use only, or essentially only, the recurrence, ignoring obvious geometrical separation, structural symmetry/repetition, or any sequence similarity. Also, we relied entirely on VAST program to produce the locally similar structural pieces, mainly because it was the only readily available program that produced such pieces (in the form of the cliques).

Since we now know that domains can be defined using the LSSPs, we would like to develop a full-fledged domain partition program, which is based mainly on the collection of LSSPs. We will take the SMF procedure and improve it by implementing the following features:
(1) Use geometry to influence the SMF procedure in two places, (a) in deciding which two domains to join in the bottom-up process and (b) in calculating the score function, which determines the final solution among the 12 candidate solutions.
(2) Use another LSSP finding program in addition to VAST in order to avoid biases introduced by using only one program. The program based upon alignment scan described above, after suitable modification, is a strong candidate.
(3) Use a custom-made clustering routine instead of the one provided by MATLAB. All currently available clustering procedures do partition, i.e. they cluster a set of objects into mutually exclusive and exhaustive set of clusters, meaning that each object belongs to one and only one cluster. However, in a protein, there can be residues that do not properly belong to any domain because they are in the linker between two domains. If these residues are forced to be included into a domain, the domain sometimes becomes prone to combine with other unrelated domains, at least in the SMF algorithm. There are also examples of proteins in which two clearly recognizable domains share a common secondary structural element, for example, a helix. We will design a new clustering scheme that allows one or both of these features (un-clustered residues and overlapping clusters). We will be guided in this process by the novel and simple clustering procedure developed by Peter Munson for the PCM process in our collaborative work.

The new program will be an improvement over the current version, which is already comparable to other programs in terms of the NDO score. More importantly, it will prove more convincingly that LSSPs can indeed define domains and domain boundaries. Our ultimate hope is that this will define domains with ‘authority’ in the sense that domains can be defined in evolutionary terms, using LSSPs, with least number of sensitive artificial parameters.

3. Relation between LSSPs and domains

Why is it possible to define domains by means of the locally similar structural pieces? Is it really because domains originally started from such pieces to begin with (divergent evolution), as the theory of Lucas et al.[6] would suggest? Or is it because domains are made of attractor SSSs, as Szustakowski et al.[4] would suggest (convergent evolution), whereas inter-domain regions are not? These questions are difficult to answer. But they inspire us to propose doing a couple of simple explorations.

One is to see if the pairs of aligned LSSPs are similar in sequence. If the similarity is low, it would not disprove the divergence theory since the event described is very ancient and LSSPs had plenty of time to mutate except perhaps those that are directly involved in the biological activity. However, it would also suggest that the theory, although possibly correct, is not terribly useful since the effect from the event has now dissipated and not traceable. If the sequence similarity is generally high, it would be a supporting evidence for the divergence theory, although one cannot rule out the possibility of convergent evolution toward a common sequence in this case, among the pieces that have a similar structure.

Another simple exploration is to generate locally aligned pieces, not by comparing structures, but by comparing sequences. There are of course many conserved sequence compilations, pfam[48] being among them. However, pfam is not particularly good at defining domains in our test set. For example, 1oy8A is considered as one-domain protein in pfam whereas SCOP, CATH and visual inspection all consider it to be made of 8 domains. (See the attached manuscript #2) This is somewhat reminiscent of our experience with structural comparisons; when we used only the significant hits, seeking those that cover the whole domain, we failed at recognizing domains. Therefore, we would instead accept many locally aligned sequences that are at or even below the significance level. We would collect these 'junk' alignments and subject them to the same mathematical procedures that we used to define domains from LSSPs. The large number of these pieces from the sub-standard alignments will swamp out the few significant alignments, which, in the case of 1oy8A, cover the whole chain. This will perhaps make it possible for the underlying domain structure to emerge. If domains are the result of a divergent evolution, many of these pieces may be the current descendents of the ancient ADSs. If they arose by the convergent evolution, the more frequently observed pieces among them may represent attractants. In either case, these pieces may collectively define the domains.

4. <u>Classification and characterization of internally symmetric proteins</u>

Using SymD, we now have a large collection of internally symmetric proteins (those that are symmetric in the monomeric form), collected from nearly 10,000 ASTRAL40 domain database. So far, we have examined individual symmetric structures manually and could see that there were broadly two types, the closed and open, as already described in the "Accomplishment" section. We are currently coding an algorithm for determining the number of repeating units from the output of SymD as a part of the process for automatically determining the type of symmetry. Once we have the procedure

established, we will prepare a database of internally symmetric proteins grouped into different symmetry types in a completely automatic fashion.

We would like to study the symmetric aspect of these structures and explore any relations between the symmetry type and their sequence, function and evolutionary history. We will probably select a small number of interesting symmetry classes and study them in great detail.

For example, symmetric proteins are made of repeating units and the interaction between them can be viewed as a simpler model system for the internal packing of non-symmetric globular proteins or for the interaction between (non-symmetric) domains and proteins. With this in mind, we can compare these three types of interactions in terms of the secondary structural types involved and the packing density. It is also possible that different symmetry types exert different degree of strains to the interface. It will, therefore, be interesting to compare the same quantities among different symmetry types. It is also probable that these interactions largely determine the symmetry of arrangement of the repeating units. We will try to discern specific features of this interaction that determine the symmetry of the whole protein. We realize that many symmetric structures have already been studied in detail and some of the above information is already available for them. We can use this body of information and add new data to it.

PROJECT 2: MISCELLANEOUS

Background

These are short projects, which are less related to above projects, and more adventurous in nature. I undertake these projects sporadically when opportunities arise, often in collaboration with other scientists. The reason for engaging in these activities is that they might provide a new perspective to the main projects or open up a new area of interest and significance.

Specific Research Aims

Sub-project 2-1.

To help understand the possible origin of the ORFan genes in prokaryotic species through their composition bias profile.

Sub-project 2-2.

To find functionally complementary gene pairs in the yeast genome through a pair-wise property of the protein-protein interaction network.

Accomplishments

Sub-project 2-1.

ORFan genes are those which either do not have a homologue in any other organism or which exist in only closely related organisms. The existence of such genes raises the question of how they arose. Assuming that viable ORFan genes do exist, and convincing evidence exists that they do, there seem to be only three possible explanations. One is that they arose *de novo* from previously non-coding sequences. The second is that they arose by duplication of an existing gene, but that they mutated so fast that the ancestor gene can no longer be detected. The third is that they came from foreign sources, possibly viruses and phages. This third source, however, seems to be unlikely in view of a recent survey by Yin and Fischer[49] that the number of ORFan gene homologs found in the public viral gene database is small, substantially smaller than that for the normal genes.

It is well known that proteins in an organism have a composition bias characteristic to that organism. Therefore, the simple idea was to compute and compare the composition biases of the protein products of the ORFan and normal genes, and of the artificial protein sequences produced by translating non-coding and random sequences where the stop codons were replaced by another codon.

The amino acid composition of each protein was expressed as a vector with 20 components. The average vector over all normal proteins was used as the reference. The composition bias of a protein is then defined as the sum of the absolute value differences between its composition vector and the reference vector, measured in units of the standard deviation of each component of the reference composition. (Equation 1 of the attached manuscript #3)

Fig. III-1 shows the histogram of the composition biases for 6 selected organisms. The red curves are for the normal proteins and are closest to the origin (zero composition bias). This is expected since the reference composition is the average over the normal proteins. The blue curves are for the artificial proteins translated from the random sequences using the A/T/G/C ratios of nucleotides of the whole genome. They are shifted to the right, indicating that the random sequences have greater biases than the normal proteins. This is of course also an expected result. Although not shown in this Figure, we also computed the compositional biases for the artificial proteins translated from the non-coding antisense and intergenic regions. Rather surprisingly, these curves are to the right of the curves for random, indicating that they have more compositional biases than random sequences.

The green curves are for the ORFan proteins. They are between the normal and random curves, and their position varies depending on the organism. After some trials, we found that the position of the ORFan compositional bias correlated with the relative age of the organism as determined from a phylogenetic tree; the older the organism, the more the ORFan compositional bias moved to the left, away from random and toward the normal biases. When the distance of the average compositional bias of the ORFan proteins from that of the random proteins was plotted against the relative age of the organism for the 47 organisms in our dataset, the correlation coefficient was 0.59. (See manuscript #3 for more details.)

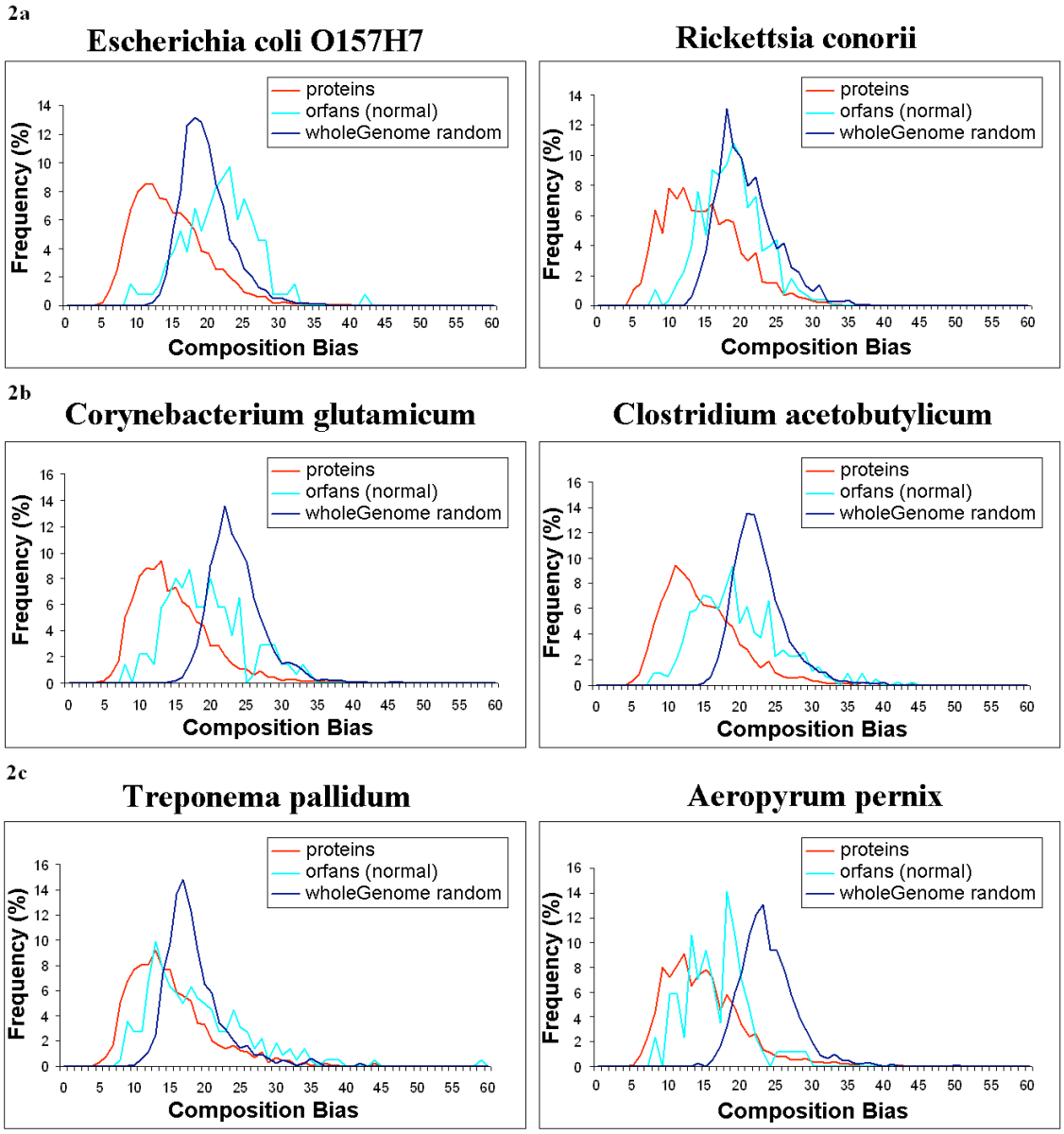

Figure III-1: Histograms of the composition bias of the set of ORFan proteins are compared
with the composition bias of all proteins and of random proteins for six organisms. Since there are fewer ORFan proteins, their histograms were scaled up accordingly (the results were validated to ensure that they are not due to sampling effects). In the two examples in the left panel (2a), the ORFan proteins behave like random proteins; in the two examples in the right panel (2c), the ORFans behave like the real proteins; and the behavior of the examples in the middle panel (2b) is intermediate.

As explained in the attached manuscript, these data are inconsistent with the hypothesis that the ORFan genes arose *de novo* from non-coding sequences, since the non-coding sequences have even more bias than the random sequences. The data are more consistent with the hypothesis that the ORFan gene arose from a gene that mutated rapidly under positive selection[50], then gradually changed its composition afterwards to that typical for the organism.

Sub-project 2-2.

It is well known that many genes in a given genome appear to be non-essential since knocking them out does not produce a phenotype. In the case of *S. cerevisiae*, a large scale study[51] showed that only about 19% of the genes were essential for growth on rich media. Part of the reason for this robustness is thought to be due to the presence of a backup copy, which performs a similar function when the first gene is knocked out. Indeed, another large scale study[52] reported that when a set of 132 single knockouts without a phenotype was crossed with all viable single knockouts, about 1% of the double knockouts were now lethal. Such pairs, or more precisely, pairs for which the effect of the double knockout is either more or less than the sum of the effects of single knockouts, are said to be in genetic interaction.

The question was whether there was any property of the protein-protein interaction network (PPIN) that can correlate with such mutually complementing backup pairs or genetically interacting pairs. PPINs are easier to construct experimentally and more widely available than genetic interaction networks, but of course do not explicitly carry any backup information.

Since genetic interaction involves pairs of genes, we also need a property of the pairs of nodes in PPIN. We reasoned that a node in PPIN was defined by its neighbors and, therefore, that a functionally similar pair in PPIN would be defined as two nodes that have a similar set of neighbors. Therefore we picked the number of common neighbors between two nodes, which we call NO for neighbor overlap, and decided to investigate how well this simplest of pair-wise measures correlates with possible backup function or genetic interaction between the pair.

Using the yeast PPIN downloaded from the DIP database, we could show: (See manuscript #4 for details.)

(1) Yeast PPIN is enriched with high NO pairs compared to carefully constructed random networks. This is consistent with the notion that biological systems are enriched with backup pairs for robustness, as well as possibly for other purposes. Yeast PPIN contains connections from complexes, which the random networks do not. Proteins in complexes interact with one another and increase the NO count. Therefore we removed interactions between proteins within the same complexes using a couple of different reported complex sets. The yeast PPIN was still enriched with high NO pairs after this removal; the signal was reduced by only about ¼ to 1/3 when the interactions within complexes were removed.

(2) The sequences are clearly more similar between high NO pairs than between low NO pairs.
(3) Clearly more pairs have the same GO (Gene Ontology) terms among the high NO group than among the low NO group. This is true for each of the three categories of GO terms, Component, Function, and Process. This indicates that the pairs with high NO tend to have a similar function.
(4) Genetic interaction is clearly stronger on average between a high NO pair than between a low NO pair.

We have examined many pairs with high NO. Some of them are from protein complexes. Most others seem to perform the same basic function but with nuances and small variations.

S. cerevisiae underwent an ancient whole genome duplication[53]. It is possible that high NO pairs are enriched in this organism because of this event and that other organisms will not show enrichment over random distribution. In any case, we have shown that NO, although simple and straightforward, is an informative property of a PPI network. We believe that the usefulness of the NO measure will be maintained when applied to PPIN of other organisms.

Current Research and Future Plans

Sub-project 2-1.

No follow up study is planned for this project.

Sub-project 2-2.

Protein-protein interaction network is something that can be constructed in the laboratory by following a rather well established experimental procedure, whether one uses the yeast two-hybrid system or a pull-down essay. Of course, there is the issue of false positives and false negatives, but still there is a clear path for obtaining PPIN. This is not true for function determination. When one has a new gene for which the only information is the sequence, there is no clear path toward determination of its function. Therefore, it would be highly useful if one can use PPIN to obtain some information on the function of the gene. The NO measure can possibly serve this function.

Many high NO pairs are clearly paralogous and have high sequence similarity. However, there are others that do not have high sequence similarity. If the function is known for one member of such a pair, one may infer that the other one has a similar function. In the future, we would like to conduct a proof of principle type of study to see if this idea is useful.

REFERENCES

1. Bystroff, C. & Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol* **281**, 565-77.
2. Bystroff, C., Thorsson, V. & Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol* **301**, 173-90.
3. Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**, 82-95.
4. Szustakowski, J. D., Kasif, S. & Weng, Z. (2005). Less is more: towards an optimal universal description of protein folds. *Bioinformatics* **21 Suppl 2**, ii66-71.
5. Petrey, D., Fischer, M. & Honig, B. (2009). Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc Natl Acad Sci U S A* **106**, 17377-82.
6. Lupas, A. N., Ponting, C. P. & Russell, R. B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol* **134**, 191-203.
7. Bera, T. K., Zimonjic, D. B., Popescu, N. C., Sathyanarayana, B. K., Kumar, V., Lee, B. & Pastan, I. (2002). POTE, a highly homologous gene family located on numerous chromosomes and expressed in prostate, ovary, testis, placenta, and prostate cancer. *Proc Natl Acad Sci U S A* **99**, 16975-80.
8. Egland, K. A., Liu, X. F., Squires, S., Nagata, S., Man, Y. G., Bera, T. K., Onda, M., Vincent, J. J., Strausberg, R. L., Lee, B. & Pastan, I. (2006). High expression of a cytokeratin-associated protein in many cancers. *Proc Natl Acad Sci U S A* **103**, 5929-34.
9. Sathyanarayana, B. K., Hahn, Y., Patankar, M. S., Pastan, I. & Lee, B. (2009). Mesothelin, Stereocilin, and Otoancorin are predicted to have superhelical structures with ARM-type repeats. *BMC Struct Biol* **9**, 1.
10. Street, T. O., Rose, G. D. & Barrick, D. (2006). The role of introns in repeat protein gene formation. *J Mol Biol* **360**, 258-66.
11. Kinoshita, K., Kidera, A. & Go, N. (1999). Diversity of functions of proteins with internal symmetry in spatial arrangement of secondary structural elements. *Protein Sci* **8**, 1210-7.
12. Murray, K. B., Gorse, D. & Thornton, J. M. (2002). Wavelet transforms for the characterization and detection of repeating motifs. *J Mol Biol* **316**, 341-63.
13. Taylor, W. R., Heringa, J., Baud, F. & Flores, T. P. (2002). A Fourier analysis of symmetry in protein structure. *Protein Eng* **15**, 79-89.
14. Murray, K. B., Taylor, W. R. & Thornton, J. M. (2004). Toward the detection and validation of repeats in protein structure. *Proteins* **57**, 365-80.
15. Shih, E. S. & Hwang, M. J. (2004). Alternative alignments from comparison of protein structures. *Proteins* **56**, 519-27.
16. Abraham, A. L., Pothier, J. & Rocha, E. P. (2009). Alternative to homo-oligomerisation: the creation of local symmetry in proteins by internal amplification. *J Mol Biol* **394**, 522-34.
17. Abraham, A. L., Rocha, E. P. & Pothier, J. (2008). Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics* **24**, 1536-7.

18. Chen, H., Huang, Y. & Xiao, Y. (2009). A simple method of identifying symmetric substructures of proteins. *Comput Biol Chem* **33**, 100-7.
19. Guerler, A., Wang, C. & Knapp, E. W. (2009). Symmetric structures in the universe of protein folds. *J Chem Inf Model* **49**, 2147-51.
20. Sam, V., Tai, C. H., Garnier, J., Gibrat, J. F., Lee, B. & Munson, P. J. (2006). ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification. *BMC Bioinformatics* **7**, 206.
21. Sam, V., Tai, C. H., Garnier, J., Gibrat, J. F., Lee, B. & Munson, P. J. (2008). Towards an automatic classification of protein structural domains based on structural similarity. *BMC Bioinformatics* **9**, 74.
22. Rose, G. D. (1979). Hierarchic organization of domains in globular proteins. *J Mol Biol* **134**, 447-70.
23. Janin, J. & Wodak, S. J. (1983). Structural domains in proteins and their role in the dynamics of protein function. *Prog Biophys Mol Biol* **42**, 21-78.
24. Holland, T. A., Veretnik, S., Shindyalov, I. N. & Bourne, P. E. (2006). Partitioning protein structures into domains: why is it so difficult? *J Mol Biol* **361**, 562-90.
25. Veretnik, S., Gu, J. & Wodak, S. J. (2009). Identifying structural domains in proteins. In *Structural Bioinformatics* Second edition edit. (Gu, J. & Bourne, P. E., eds.), pp. 485-513. John Wiley & Sons, Inc., Hoboken, N.J.
26. Holm, L. & Sander, C. (1998). Dictionary of recurrent domains in protein structures. *Proteins* **33**, 88-96.
27. Kim, C. & Lee, B. (2007). Accuracy of structure-based sequence alignment of automatic methods. *BMC Bioinformatics* **8**, 355.
28. Jung, J. & Lee, B. (2000). Protein structure alignment using environmental profiles. *Protein Eng* **13**, 535-43.
29. Holm, L. & Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics* **16**, 566-7.
30. Tai, C. H., Vincent, J. J., Kim, C. & Lee, B. (2009). SE: an algorithm for deriving sequence alignment from a pair of superimposed structures. *BMC Bioinformatics* **10 Suppl 1**, S4.
31. Kim, C., Tai, C. H. & Lee, B. (2009). Iterative refinement of structure-based sequence alignments by Seed Extension. *BMC Bioinformatics* **10**, 210.
32. Zhu, J. & Weng, Z. (2005). FAST: a novel protein structure alignment algorithm. *Proteins* **58**, 618-27.
33. Kawabata, T. (2003). MATRAS: A program for protein 3D structure comparison. *Nucleic Acids Res* **31**, 3367-9.
34. Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S. E. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Res* **32**, D189-92.
35. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. (1999). A census of protein repeats. *J Mol Biol* **293**, 151-60.
36. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-40.
37. Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A* **70**, 697-701.

38. Jones, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C. & Thornton, J. M. (1998). Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci* **7**, 233-42.
39. Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv Protein Chem* **34**, 167-339.
40. Tress, M., Tai, C. H., Wang, G., Ezkurdia, I., Lopez, G., Valencia, A., Lee, B. & Dunbrack, R. L., Jr. (2005). Domain definition and target classification for CASP6. *Proteins* **61 Suppl 7**, 8-18.
41. Veretnik, S., Bourne, P. E., Alexandrov, N. N. & Shindyalov, I. N. (2004). Toward consistent assignment of structural domains in proteins. *J Mol Biol* **339**, 647-78.
42. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093-108.
43. Taylor, T. J. (2006). Analysis of the structure and topology of real and model proteins using Delaunay tessellation Ph. D., George Mason University.
44. Holm, L. & Sander, C. (1994). Parser for protein folding units. *Proteins* **19**, 256-68.
45. Gibrat, J. F., Madej, T. & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Curr Opin Struct Biol* **6**, 377-85.
46. Tai, C. H., Lee, W. J., Vincent, J. J. & Lee, B. (2005). Evaluation of domain prediction in CASP6. *Proteins* **61 Suppl 7**, 183-92.
47. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science* **273**, 595-603.
48. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**, 320-2.
49. Yin, Y. & Fischer, D. (2006). On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. *BMC Evol Biol* **6**, 63.
50. Long, M., Betran, E., Thornton, K. & Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**, 865-75.
51. Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K. D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C. Y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W. & Johnston, M. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-91.
52. Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D. S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J. N., Lu, H., Menard, P., Munyana, C., Parsons, A. B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A. M., Shapiro, J., Sheikh, B., Suter, B., Wong, S.

L., Zhang, L. V., Zhu, H., Burd, C. G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F. P., Brown, G. W., Andrews, B., Bussey, H. & Boone, C. (2004). Global mapping of the yeast genetic interaction network. *Science* **303**, 808-13.

53. Kellis, M., Birren, B. W. & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617-24.